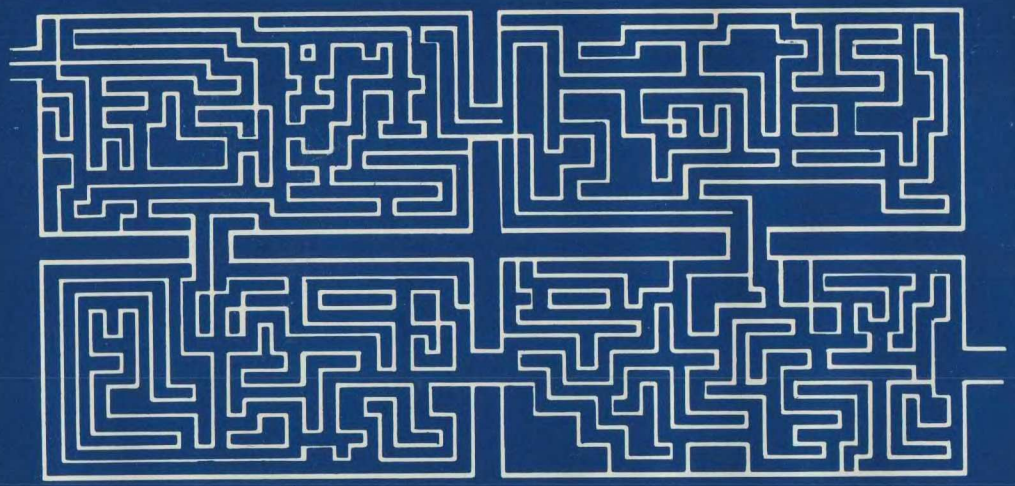
Evaluation Studies Review Annual

Volume 9



Edited by

Ross F. Conner

and ASSOCIATES

Typifying the multidisciplinary character of evaluation studies, this ninth volume in the Evaluation Studies Review series provides a representative sample of the most vital concerns, popular research topics, and contemporary issues in the field today. The editors have organized the material into a nine-part framework that coincides neatly with their view of the principal steps in the entire evaluation process. Philosophical issues; conceptual models; selecting approaches; validity considerations; analyzing, interpreting, disseminating, and using evaluation data—all these areas and more are explored in readable detail by distinguished contributors from all disciplines of the evaluation field.

Ideal as a beginning handbook for evaluation students on any level, or as a contemporary sourcebook for practicing professionals, **Evaluation Studies Review Annual,** Volume 9 is a comprehensive collection of state-of-the-art literature.

ISBN 0-8039-2386-4

ABOUT THE AUTHORS

ROSS F. CONNER is Associate Professor of Social Ecology at the University of California, Irvine. He is a member of the Council of the Evaluation Research Society and has served on several of the society's committees. He is the coauthor of *Sesame Street Revisited* (with T. Cook, H. Appleton, A. Shaffer, G. Tamkin, and S. Weber) and *Attorneys as Activists: An Evaluation of the American Bar Association's BASICS Program* (with R. Huff). Professor Conner's recent writing focuses on methodological and ethical aspects of evaluation and on the use of evaluation results in policymaking. Dr. Conner recently was named a W.K. Kellogg Foundation National Fellow.

DAVID G. ALTMAN recently completed his Ph.D. in social ecology at the University of California, Irvine. Currently he is a postdoctoral fellow with the Stanford Heart Disease Prevention Program. His interests are in evaluation research, health psychology, and community psychology.

CHRISTINE JACKSON is a doctoral candidate in social ecology at the University of California, Irvine. Her dissertation research involves an evaluation of one of the American Heart Association's community health education programs. Her interests are in health psychology, social intervention, and evaluation research.

SAGE PUBLICATIONS
The Publishers of Professional Social Science
Beverly Hills London New Delhi

EVALUATION STUDIES REVIEW ANNUAL
Volume 9

Evaluation Studies

Review Annual

Evaluation Studies

Review Annual

Volume 9 1984

Edited by

Ross F. Conner

David G. Altman

Christine Jackson

SAGE PUBLICATIONS Beverly Hills London New Delhi

For information address:

SAGE Publications, Inc.
275 South Beverly Drive
Beverly Hills, California 90212

SAGE Publications India Pvt. Ltd.
C-236 Defence Colony
New Delhi 110 024, India

SAGE Publications Ltd
28 Banner Street
London EC1Y 8QE, England

Printed in the United States of America

International Standard Book Number-8039-2386-4

International Standard Series Number 0364-7390

Library of Congress Catalog Card No. 76-15865

FIRST PRINTING

CONTENTS

ABOUT THE EDITORS

ROSS F. CONNER is Associate Professor of Social Ecology at the University of California, Irvine. He is a member of the Council of the Evaluation Research Society and has served on several of the society's committees. He is the coauthor of *Sesame Street Revisited* (with T. Cook, H. Appleton, A. Shaffer, G. Tamkin, and S. Weber) and *Attorneys as Activists: An Evaluation of the American Bar Association's BASICS Program* (with R. Huff). Professor Conner's recent writing focuses on methodological and ethical aspects of evaluation and on the use of evaluation results in policymaking. Dr. Conner recently was named a W.K. Kellogg Foundation National Fellow.

DAVID G. ALTMAN recently completed his Ph.D. in social ecology at the University of California, Irvine. Currently he is a postdoctoral fellow with the Stanford Heart Disease Prevention Program. His interests are in evaluation research, health psychology, and community psychology.

CHRISTINE JACKSON is a doctoral candidate in social ecology at the University of California, Irvine. Her dissertation research involves an evaluation of one of the American Heart Association's community health education programs. Her interests are in health psychology, social intervention, and evaluation research.

1984: A Brave New World for Evaluation?

Ross F. Conner, David G. Altman, and Christine Jackson

Evaluation, as an established field, is now in its late adolescent years. The bubbling, exciting, fast-developing childhood years of the late 1960s and early 1970s gave way in the mid to late 1970s to the less self-assured, serious, introspective early adolescent years. Now, in the early 1980s, evaluation is making the transition from late adolescence to adulthood. The coincidence of this transition with the presence of the "brave new world year" of 1984 (with apologies to Huxley and Orwell) provides a convenient excuse to reflect on the state of evaluation. What kind of a brave new world is in store for evaluation? What aspects of evaluation's development need more attention to facilitate the positive aspects of what lies ahead?

In assembling the contents of this volume, we developed some insights into these issues as we reviewed hundreds of recent evaluation-related articles, journals, books, reports, and conference presentations. Following a discussion of our approach to the *Annual,* we share these insights and suggest some areas of the evaluation enterprise that will require greater attention in the years to come if evaluation is to continue to develop successfully into middle age.

OUR APPROACH TO THE *ANNUAL*

Before we provide our overview of the past and present trends in evaluation research, we need to explain our approach to assembling this edition of the *Annual* so that the reader can make his or her own judgment about the adequacy of the data base from which our conclusions are drawn.

We began our review in the fall of 1983 when, as a result of our evaluation experience and understanding of the field, we developed a conceptual framework of the evaluation process. This framework, adapted as we continued our review of recent material, consists of these principle steps, which we believe characterize most evaluation projects: adopting an evaluation philosophy or ideology, learning about the context of the specific program under study, selecting a general evaluative approach, developing a model of the program and its operations, designing an evaluation plan, developing sampling and measurement strategies, analyzing the data, then disseminating and utilizing

the evaluation results. Some of these steps are often unconsciously taken (for example, adopting an evaluation philosophy or ideology), while others are so well known that they occasionally overshadow other steps in the process (for example, designing an evaluation plan). With the exception of the last step (that is, dissemination/utilization, which we believe occurs throughout an evaluation study), we view these steps as generally occurring in the sequence listed. While the relationship among steps is fixed, there is flexibility within each step. This flexibility results from the need to match the most important general features of a step to the particular demands of the evaluation setting.

With this framework as a general organizing guide, we began our review for the *Annual* in the fall of 1983. We assembled lists of journals, books, and reports that seemed likely to contain significant evaluation work. We were greatly assisted in our task by the members of our editorial board, who suggested both noteworthy articles and likely sources of unknown but good evaluation work from their particular fields. We encouraged our editorial board members to take a very catholic view of the field and to suggest works from untraditional journals and unusual sources. It is our feeling that one of the great strengths of evaluation is its multidisciplinary character, and we wanted to try to typify that in our selections. Unfortunately, we identified many more worthy articles than we could publish, due to strict page limitations from the publisher. Consequently, the papers included in this volume are not all of the best recent works in the field. Instead, the papers here are a selection of very good work that reflects a number of important and timely evaluation issues and concerns.

The main sections of the volume reflect the main steps in the evaluation framework we developed. Part I of the volume addresses philosophical and ideological issues surrounding evaluation research. The next parts of the volume focus on the steps most evaluators would take in developing a specific evaluation project. The initial step in many evaluations is an informational one, as the evaluator learns about the context of a program and about its main actors, clients, and activities (covered in Part II). Next, the evaluator selects a general approach or several approaches to his or her study (Part III). Then, in a step that is still rare but generally favored, the evaluator develops a conceptual or rhetorical model of the program, which guides him or her in observing the program as it actually is implemented and in selecting the most critical causal links for careful examination (Part IV).

The next steps in this generalized process are perhaps the most familiar ones. The details of the evaluation are set out, with special attention to validity considerations (Part V), and sampling and measurement decisions (Part VI). The evaluation plan then is implemented, producing data to be analyzed and interpreted (Part VII). The evaluator is then ready to disseminate his or her findings and focus on utilization issues (Part VIII). This last step, in our view, really occurs with each of the previous steps if dissemination and utilization are to be accomplished successfully. We reject the traditional view that

utilization is something that occurs only after the study is completed. Nonetheless, for clarity, we have set out papers focusing on dissemination and utilization in a separate section.

In addition to these sections covering generalized steps of the evaluation process, we have included a concluding section on professional issues, such as standards and ethics, and on future directions of the field (Part IX).

PAST TRENDS IN EVALUATION RESEARCH

As the journals, professional organizations, books, awards, and even the existence of this annual review series attest, evaluation research is now a well-established discipline. This outcome certainly was now clear, or perhaps not even envisioned or desired, when the field arose from various branches of more long-standing disciplines. The following summary history of evaluation's emergence gives some idea of the significant developments and problems that characterized the field. We should note that this history is slanted toward the national view, although evaluation work at the state and local levels generally followed developments at the national level. In addition, the events and trends that we present below are probably best viewed as several people's view of the field's development rather than as an official history. That task will have to fall to others less involved and more objective than we are.

While it is impossible to establish the exact date of its birth, evaluation research emerged in the mid-1960s as a number of developments converged. Foremost among these was the beginning of the great emphasis on the development of social programs under Presidents Kennedy and Johnson. At about the same time, two important books, by Campbell and Stanley (1966) and Suchman (1967), were published and widely read. Out of these developments as well as others, smaller in scale but nonetheless significant, emerged the field of evaluation. The main trends during evaluation's early years were large-scale studies, often national in scope, and a bias toward quantitative methodologies. The large studies resulted from the support that social intervention programs generally enjoyed in Washington at that time. The quantitative bias resulted from the novice evaluators' previous training and participation in traditional social science disciplines.

In the early 1970s, following the first euphoric years, evaluation research began to institutionalize itself. Informal groupings of those working in evaluation changed into formal professional organizations. The Evaluation Research Society was formed, as were the Evaluation Network and the Council for Applied Social Research (subsequently merged into the Evaluation Research Society). Partly as a result of the formation of these organizations, evaluators began to evaluate evaluation. Both evaluation researchers and those who were the intended users of evaluation research began to look closely at the conduct and outcomes of the many evaluations that had been completed. What they found did not always please them. Evaluations in some cases were too equiv-

ocal, too costly and too late to be useful. Some evaluators complained that we were using the wrong tools, in terms of designs, methods and measures, or that we simply did not have the tools we needed to produce good studies. Others complained that the problem was not us but the turbulent setting in which we had to work. The arguments at the professional meetings were lively and healthy, if not always conclusive.

In the mid to late 1970s, evaluation research had its scientific christening when two journals devoted solely to evaluation were founded: *Evaluation Quarterly* (subsequently retitled *Evaluation Review* and published more frequently to accommodate all worthy articles) and *Evaluation and Program Planning*. With this institutionalization of an evaluative mechanism for the field itself, evaluators began seriously to indulge in self-criticism. Certain issues had to be faced squarely, such as the irrelevance and misapplication of the control-group design in some studies or the absence of use of evaluation results in policymaking.

Evaluators were beginning to develop creative solutions to these and other issues as the 1980s began. Just at that point, however, the Reagan administration started its cutback in social programs and its layoffs of federal workers. Not only did the funds for evaluation work begin to shrink significantly, but people working in the relatively new evaluation units in federal agencies also began to lose their jobs, as the last hired became the first fired. These developments caused evaluation to shift from a growth industry, as some had characterized it, to a mildly embattled enterprise. Evaluation definitely has not died, but it no longer has the fervor of the earlier years, when interest in and funds for evaluation studies were high.

THE STATUS OF EVALUATION IN 1984

In the course of our review of material for this volume of the *Annual,* we had the opportunity to read many papers, reports, and books focusing on every aspect of the evaluation enterprise. While at times fatiguing, this exercise did have one very valuable aspect: It gave us the rare opportunity to learn what many evaluation researchers were doing and thinking. The papers in this volume provide a representative sampling of important concerns and popular research foci in the evaluation field at this time. (The one exception would be the area of meta-analysis, which, because it was covered so extensively in the previous *Annual,* (Light, 1983), we underemphasized in this volume.) In the sections below, we note some of the most significant trends and, with some trepidation due to the questionable validity of our crystal ball, provide observations on likely or desirable developments as evaluation marches into the brave new world beyond 1984.

The Quantitative-Qualitative Debate: A Truce

A long-standing dispute in the evaluation research literature has centered on the advisability of using quantitative or qualitative methods. Initially, the

quantitative methods held the superior position and qualitative approaches rarely were mentioned. To be heard, advocates of qualitative methods had to emphasize the differences and superiority of their favored approach. This resulted in an understandable but unfortunate debate that pitted quantitative approaches against qualitative approaches and cast the evaluator's initial design decision as an "either-or" situation.

In our review, we were struck by the degree to which evaluators seem genuinely to have gone beyond the quantitative versus qualitative distinction. The issue no longer seems to be which approach is better but, instead, is how we can capitalize on the complementarity of these approaches to design more sensitive studies (see Parts I and III in this volume for examples). We believe this is a healthy development for the field, which may lead to the development of a methodology for evaluation that is unique, rather than an amalgam of traditional social science approaches. Should this happen, it is conceivable that evaluation research may make a significant methodological contribution the social science fields from which it sprang.

Multiple Methods

There has been an important related development in the area of evaluation approaches, spurred perhaps by the end of the obsessive focus on the qualitative versus quantitative issue. A number of writers are advocating that methods be mixed in conducting an evaluation (See Parts I, II, and III).

The challenge is to mix the best parts of multiple methods to accomplish our evaluation tasks. Thus far there are more calls for the use of multiple methods than actual examples of how this can be accomplished successfully. Nonetheless, this important shift in thinking is a necessary precondition for the development of new models. Consequently, we anticipate that some very creative multiple-method models will begin to appear in the new few years.

New Approaches for Lean Times

As too many evaluators are well aware, funds for evaluation have decreased in recent years. One outcome has been less evaluation work, but another has been the advocacy of new evaluation approaches for leaner times (see Parts III and IX).The approaches we identified were not actually new; instead, they were adaptations of older methods to the evaluation setting. Notable among these adaptations are the use of social indicators and the use of archival measures to assess program coverage and effectivness. These approaches typically are less costly to implement than traditional evaluation approaches because the evaluator is using data already collected by others. Evaluators using these approaches frequently will need to be creative in selecting process and outcome measures because they will be working with data others have defined and collected. The challenge will be to assemble sets of measures that individually focus on different aspects of a phenomenon but that collectively elucidate the phenomenon in a multi-dimensional way. This task will require

evaluators to use tried and true, as well as unusual and unexpected, measures and indices.

Decentralization of Evaluation

In its earliest years, evaluation research involved many large-scale programs, usually administered at the federal level. Many of the well-known educational program evaluations, such as Head Start, exemplify this trend. As resources for social amelioration programs have decreased in recent years and as the available funds have been shifted to the state or local levels for administration, the focus for evaluation activities also has shifted. The number of large federal-level evaluations has dropped dramatically, and the number of state- and local-level evaluation activity has increased (see Parts II and IX). We expect this trend to continue, at least for the near future.

This trend has positive and negative implications for evaluation research. On the positive side, evaluators will be working close to those directly responsible for implementing, administering, and improving programs, thereby increasing the likelihood of relevant evaluation data and timely evaluation use. Unlike the large federal programs, where many people are involved in these different processes in many locations, small local-level programs typically involve fewer people and fewer intermediaries between the planning for a program and its implementation. The evaluator, then, can work more closely with these people and can develop and implement a study design that is known to be more responsive to the needs of the program personnel. (Some of the articles in Part VI discuss related issues.)

This trend toward the local level will not necessarily preclude national-level assessments of similar types of programs. Some current evaluation writers are advocating the use of data convergence to obtain cumulative knowledge of programs or evaluation practices. Some believe that particularly meaningful knowledge of this type will come from multiple small-scale studies that employ a variety of methods and measures.

On the negative side, evaluators will be working with much smaller evaluation budgets and so may not be able to implement some of the sophisticated designs and measures that were possible on the large-scale level (e.g., the New Jersey Negative Income Tax Experiment and its counterparts in other cities across the country). In addition, evaluators may not have the degree of high-level support for evaluation activities that characterized their work at the federal level. While there certainly were negative consequences to some aspects of federal mandates for evaluation, there were definite boons to the expression of strong evaluation support, such as the requirement from Congress that evaluation must be done on certain federal programs or the directives from various agency or department heads that evaluation activities would be an integral part of program activities. The absence of these high-level supports for evaluation will require that evaluators cultivate these kinds of supports at the state and local levels. (Articles on related topics are found in Parts II, VIII, and IX.)

New Conceptions of Validity

A central part of evaluation researchers' thinking has been Campbell and Stanley's (1966) distinction between external and internal validity, based on Campbell's work nearly 10 years earlier (1957). Internal validity, most of us learned, related to whether X in fact caused Y in a particular situation. External validity related to generalizability of the X-Y relationship to other settings, populations, and treatment-measurement variations. Campbell and Stanley provided us with a listing of possible confounding factors that could invalidate the conclusions of our studies and ways to ameliorate these factors.

Recently, evaluation writers have begun to question the validity of this distinction and to reformulate the idea of validity (see Parts V and VII). We think this is an exciting development for the field because it will force us to reexamine our approach to determining causation.

Too many of us, with Campbell and Stanley's catalogue of designs in hand, thought that selecting the right design would solve our invalidity problems. This assumption frequently was incorrect for the multifaceted evaluation setting and resulted in unsatisfactory and unsatisfying attempts to institute theoretically proper designs. Part of the surge in popularity of qualitative evaluation designs about a decade ago resulted from these frustrating experiences.

Now, however, evaluation writers are challenging us to think beyond validity issues to more basic concerns of how we can best determine and measure cause in an evaluation context. The focus of our efforts needs to be an explanation of *how* and *why* a process, program, or product is working, not simply on *whether* it is working. The outcomes from the discussion of these validity issues over the next few years could play a significant role in advancing evaluation's methods and techniques and, perhaps, social science's methods and techniques as well.

The Use of Evaluation

As we noted in our brief history of evaluation research, the absence of utilization of evaluation results was one troubling factor in causing a good deal of soul-searching in the field. That soul-searching is still going on, but its tone is of a much different quality. The definition of evaluation at that early time was too restricted to an instrumental view, where immediate and direct details of use were examined, usually those the evaluator had recommended. If these things did not occur, the judgment was that no use had resulted from the evaluation.

This limited view of evaluation began to change in the late 1970s and has changed even more in the past few years. Conceptual uses and symbolic uses also are recognized as examples of utilization, although these may not be in exactly the directions the evaluator would favor. Even instrumental uses appear more frequently than once thought, perhaps because evaluators have become better judges of the decision requirements of policymakers and decision makers. It is clear to us from our review that evaluations are indeed used,

and that the agonizing over nonuse has appropriately subsided (see Parts II, IV, VIII, and IX). The focus now is on how to facilitate and encourage more use of evaluation findings.

The Role of Conflict

In the early years of evaluation's development, conflict between evaluators and program personnel or decision makers was viewed as one of those unfortunate realities of doing evaluation in the "real world." There were disagreements over methods and measures, results and implications. Some evaluation writers recommended avoiding doing evaluations in settings where too much conflict existed. While this admonition is still applicable in extreme situations, it is noteworthy that conflict no longer is viewed as an unpleasant, uncontrollable aspect of the evaluation setting. Instead, evaluators seem to be accepting conflict as an integral part of working in public and private programs (see Parts II, III, VIII, and IX).

Conflict need not always be a disadvantageous factor in evaluation planning, implementation, and utilization. If an evaluator is flexible enough and rational enough, he or she can capitalize on disagreements and turn them into assets. For example, disagreements between program personnel about the goals and objectives of a program can be identified by the evaluation staff and fed back to the program personnel. Because these disagreements sometimes have never been clearly or concretely set out, they have not been resolved. The evaluator can serve a useful function for the program, as well as for the planning of the evaluation, by serving as the catalyst for goal and objective clarification. By doing this, the evaluator can enhance his or her credibility and perceived usefulness to the program and, as a consequence, be granted even more freedom in the conduct of a comprehensive evaluation.

Recognition of the Importance of the Evaluator's Sensibilities

There has been much attention in the past on the technical skills an evaluator must have; little attention was devoted to the other skills an evaluator must have to complement his or her technical abilities in the actual conduct of an evaluation project. Increasingly, evaluation writers have come to recognize that the development of these evaluation sensibilities is critical in training novice evaluators.

These sensibilities include an understanding of the importance and the operation of the context within which an evaluation occurs and the decision-making situations into which evaluation data are introduced (see Parts II and VIII). Other important sensibilities relate to the professional conduct of an evaluation project and to the ethical questions that may arise as an evaluation progresses (Part IX). The latter issues are receiving increasing attention in the field, but there is still room for improvement. It is clear that the increase in understanding of these factors has resulted in more successful and useful eval-

uations. The challenge for the evaluation field now is to develop ways to teach novice evaluators about these sensibilities, probably using experiential components in evaluation classes. In this way, students of evaluation will be able to learn the science as well as the art of evaluation.

Evaluator Training and Jobs

As in any field, training of new professionals and retraining of current members should be important issues for evaluation. Some activities related to both of these areas have occurred at recent joint meetings of the Evaluation Network and the Evaluation Research Society, namely the Evaluation Teaching Materials Exchange and the preconference workshops, several of which have focused on training in specialized topics for evaluation professionals. While these activities are useful ones, the field could benefit from more systematic thought and action on the issue of training.

We searched in vain for articles describing current evaluation training or retraining efforts. While we did locate and include several good papers related to evaluation jobs (Part IX), these papers only indirectly relate to the issue of training. There are several possible reasons for the absence of attention to training. Until recently, those wanting to be evaluators had little difficulty finding jobs, even if their detailed evaluation training was limited. Advanced social science or education training usually was sufficient to qualify someone for an evaluation position. This situation has changed, however, as competition for evaluation jobs has increased. Those seeking employment are now very interested in specialized training in evaluation to enhance their marketability.

Another reason for the absence of attention to training has been the disagreement among evaluators about the critical skills that should make up a good training package. As the evaluation field was developing, no one could be certain about which skills and knowledge were more or less critical. Now, however, the field is at a point where serious discussion and study about training and retraining evaluators could lead to general agreement about the set of core evaluation skills. Important topics of such studies would include the goals, content, and outcomes of training programs that currently exist; the skills and knowledge that evaluators believe are important for training; and the views of employers of evaluators about the skills they value. This information would provide a basis on which to redesign current training classes, workshops, and programs, as well as to plan responsive, relevant retraining programs. Without this information, the evaluation field is likely always to be behind, rather than in step with or ahead of, the job market's needs.

CONCLUDING THOUGHTS

In the first edition of this *Annual*, Glass (1976) noted that a discipline begins and grows if it is centered around a set of intellectually engaging ques-

tions. It is clear to us that the discipline of evaluation research easily meets this criterion. Since Glass produced his volume, the discipline has continued to generate important and engaging questions of two main types: those that address more and more basic aspects of old issues (for example, questions about validity) and those that address new issues previously unrecognized or ignored (for example, questions about decision making and the use of evaluation results). If the past can be used to predict the future, the discipline of evaluation should continue to prosper in the years beyond 1984.

The evaluation enterprise has a second aspect, however, that also must be considered in assessing the state of the field. Evaluation arose from the demands of social policymakers who were attempting to solve or at least ameliorate important social problems. Evaluation has grown because it has contributed meaningful, useful information to the policy process. Evaluators continue to make such contributions and to develop new approaches that will produce even more meaningful information. A special strength of the evaluation enterprise has been its ability to adapt to and even capitalize on, the complex and confusing but creative environment of social programs. This strength is still very much in evidence, as the papers in this volume demonstrate. As the evaluation enterprise moves beyond 1984, we can look forward to a brave, exciting world—an ever changing set of engaging questions anchored and tested in the important arena of social problem solving.

REFERENCES

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin,* **54**, 297–312.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Glass, G. V (Ed.). (1976). *Evaluation studies review annual* (Vol. 1). Beverly Hills, CA: Sage.

Light, R. J. (Ed.). (1983). *Evaluation studies review annual* (Vol 8). Beverly Hills, CA: Sage.

Suchman, E.A. (1967). *Evaluative research.* New York: Russell Sage Foundation.

I

THE PHILOSOPHY AND IDEOLOGY OF EVALUATION

The biggest questions and issues facing evaluation research are related to our operating assumptions, made both consciously and unconsciously. Too often our attention as evaluation researchers is on the microscopic questions, such as which program goals to evaluate, which measures to use, which data analysis strategies to employ. Less frequently we think about somewhat more macroscopic questions, such as which general evaluation approach is best suited to our purposes. An evaluator facing this question might ask himself or herself whether, for example, a goal-based or a goal-free approach would be better suited to a particular evaluation task. Rarer still is the evaluator who asks the ultimate macroscopic questions, such as, Can we really be scientific? How do our unconscious biases affect our conduct of evaluation? How is the way we conceptualize evaluation biased by our built-in preconceptions? and Is there an ideal paradigm?

These important but difficult questions are easy to ignore in the day-to-day pressures of the evaluation environment. We evaluation researchers, however, ignore them at our peril, since it is precisely the process of considering and attempting to answer such macroscopic questions that determines the future success or failure of the evaluation enterprise. Too much attention to microscopic issues and too little to macroscopic concerns can result in a dangerous situation for the profession of evaluation.

Several of evaluation's best-known figures address just such macroscopic issues in this section. Donald Campbell, one of the acknowledged fathers of the evaluation field, contributes the text of a speech delivered to the American Educational Research Association. Following a brief overview of the limitations of the logical positivist approach to doing social science and a short presentation of post-positivist standards for the conduct of social science research, Campbell discusses ten significant features of applied social science conducted for policy purposes. (In one part of this discussion, Campbell introduces an extension of the experimental/quasi-experimental categorization of research design for which he is well known: queasy experiments.) These ten features are characterized by a more modest view of the ability of scientists to make causal inferences using our fallible, fragile tools. Campbell concludes with several suggestions of alternative models and approaches that will foster better evaluation work.

Michael Scriven has addressed the most macroscopic evaluation questions and issues for many years. In his paper in this section, he continues this

healthy and important tradition. Scriven discusses four ideologies or fundamental biases that he believes have pervaded the evaluation enterprise: separatism, positivism, relativism, and a managerial focus. Scriven argues instead for a consumerist ideology and a multimodel approach to evaluation, one that takes multifield, multidisciplinary, and multilevel perspectives, to name just a few of the multiplicities that Scriven describes. The article ends with an evaluation checklist that presents 15 dimensions that Scriven believes must be considered in doing most any type of evaluation, whether of products, programs, people, or proposals.

The following paper by Ernest House also addresses the question of how we think about evaluation and the factors that underlie our thinking. House's intriguing contention is that our evaluation conceptions are unconsciously influenced by metaphors. Metaphors for House are vital intellectual tools which are central to our perception and understanding of the world around us. In the case of evaluation, the same metaphorical thinking applies. As an illustration, House analyzes the metaphors underlying the popular evaluation text by Rossi, Freeman, and Wright (1979), *Evaluation: A Systematic Approach.* Their metaphors include a view of social service delivery as industrial production ("program elements are defined in terms of time, cost, procedures, or a product"), of social programs as machines (social programs can be "fine-tuned"), and of social delivery systems as conduits ("the unreliability of measuring instruments may dilute the difference in outcomes"). House demonstrates how different conceptions of evaluation result from these different metaphorical views. House's novel view of what determines our evaluation approaches raises the possiblity that very different perceptions of evaluation, with associated unusual approaches, are possible, at least for those who can conceive of a new metaphor.

The final paper in this section also deals with questions of fit between the conception of evaluation and the execution of evaluation. Palumbo and Nachmias explore several aspects of the "identity crisis" that confronts evaluation now that the dominant evaluation paradigm (the goal-based model) has been supplemented by a variety of alternative paradigms. They discuss such issues as the ideal role of evaluation in decision making and the congruence (or, more frequently, incongruence) between evaluation methodologies and actual organizational behavior. Palumbo and Nachmias are not sanguine about the development of an ideal evaluation paradigm, but they argue that it is nonetheless worthwhile to work toward it.

The four papers in this section share several common threads. First, all the papers have an anti-positivist bias and call for a change in our thinking. Scriven clearly argues for a change not only away from any lingering positivist bias in evaluation but also away from the management bias in past evaluation studies. Campbell and Palumbo and Nachmias also urge evaluation researchers to look more closely at the evaluation situation before deciding on approaches and methodologies. Although more subtle, House also challenges us to change our standard way of thinking about evaluation.

The ideas in these papers go beyond the old, simple distinction between quantitative and qualitative evaluation approaches to even more basic conceptions. Rather than argue about different models and different methods, these evaluation theorists focus instead on different philosophies and ideologies. By probing to the core of the evaluation enterprise, these evaluation thinkers are conducting the most basic kind of evaluation of evaluation and, in so doing, are opening up the possibility that we can begin genuinely to understand the points of agreement and disagreement. Once we have this understanding, we are in a position to make real progress in advancing the science of evaluation.

1

Can We Be Scientific in Applied Social Science?

Donald T. Campbell

Can we be scientific in applied social science? My ability to take a middle-of-the-road, sensible position in a militant polemic way makes you know that my awswer will be both *yes* and *no*. Certainly it is much harder to be scientific where financially enormous policy decisions hang upon our fragile social science tools. Let me give you one preliminary *yes* and two preliminary *no's*.

A feeble *yes:* We can be somewhat more scientific than we are now or have been (in educational program evaluation, for example). Changes feasible within the current financial, administrative, and political climate could make us able to be more scientific. An equally feeble *no:* If we present our resulting improved truth claims as though they were definitive achievements comparable to those in the physical sciences, and thus deserving to override ordinary wisdom when they disagree, we can be socially destructive. We can be engaged in a political misuse of the authority of science that has not been fully earned in our own field. Another *no:* Using quantitative social science measures for administrative control and budgetary decision making (as in the accountability movement) can be destructive of the institutions and processes over which control is intended, and destructive as well of whatever prior validity the social science measures employed once had.

I want to come to these conclusions, or get close to them, by briefly reviewing recent developments in the philosophy of science, sociology of science, and sociology of knowledge, including the argument within our own program evaluation community as to whether we should employ the methods of physics or the methods of the humanities. In light of a fragmentary, modern, post-positivist theory of science, I will then discuss the special problems of policy-relevant

From Donald T. Campbell, "Can We Be Scientific in Applied Social Science?" original manuscript.

Author's Note: This is an edited version of a transcript of the "awards address" delivered at the annual meeting of the American Educational Research Association, March 22, 1982, in New York. (One or two of the topics on my outline that were not actually delivered have been included here.) Preparation of the talk, transcript, and edited version have been supported by NSF Grant BNS 7925577 and by my university professorship at Lehigh University. A modified version of this paper will appear in the *Educational Researcher.*

social science research, including problems resulting from the politicization of our own mistaken view as to what an applied social science should look like, which we offered in the heyday of the Great Society Program of the 1964–1968 period, under the regime of one of our two presidents named Lyndon Johnson (that is, "Lyndon Johnson the Good"). I am thinking of the Office of Economic Opportunity, Program Planning and Budgeting Systems, and program evaluation.

POST-POSITIVIST THEORY OF SCIENCE

Twenty years ago logical positivism dominated the philosophy of science and, through concepts like *operational definition,* dominated our thoughts about research methods. Today the tide has completely turned among the theorists of science in philosophy, sociology, and elsewhere. Logical positivism is almost universally rejected. This rejection, in which I have participated, has left our theory of science in disarray. Under some interpretations it has undermined our determination to be scientific and our faith that validity and truth are rational and reasonable goals. What we should have learned instead was that logical positivism was a gross misreading of the method of the already successful sciences. Logical positivism was wrong in rejecting causal processes imputed to unobserved variables. Logical positivism failed to recognize that even at its best, experimental research is equivocal and ambiguous in its relation both to the real physical process involved and to scientific theory; and that attention to this equivocality calls for use of multiple methods, none of them definitional, triangulating on causal processes that are imperfectly exemplified in our experimental treatments and measurement processes. Properly interpreted, the dethronement of logical positivism should have led to an *increase* in methodological concern rather than its abandonment. Positivism's worst gift to the social sciences was definitional operationism, and this still persists in applied social science, as in the accountability movement in which goal statements and achievement claims are rigidly defined in terms of singular, quantitative indicators. (In practice, the use of such indicators for practical decision making reduces or emasculates the validity of the measures involved [Campbell, 1979a, pp. 84–86]).

Campbell and Stanley's *Experimental and Quasiexperimental Designs for Research* (written in 1961 and 1962, first published in 1963), was lucky to be already post-positivist. (At least in a whiggish rewriting of history, I can claim that. In Cook and Campbell [1979, pp. 10–14] the assessment is more mixed.) First of all, we cited N. R. Hanson (1958), who was the first in the Hanson-Kuhn-Feyerabend tradition to emphasize the theory-ladenness of the factual observations of science. It cited Popper (1959) with approval (although it didn't cite my favorite slogan of his: "We don't know, we can only guess"). We emphasized the equivocality of both the treatment implementations and the observations. We gave a section head and two paragraphs to evolutional epistemology (in Campbell [1959] version, if not the 1974a, and thus did not

include my now standard, "Cousins to the amoeba, how could we know for certain?"). Most importantly post-positivist was the concept of "plausible rival hypotheses," putting so much scientific weight on that squishy concept of *plausibility*.

I would like to point out five post-positivist points with which I agree and with which I think you also should agree. I am borrowing from Hanson (1958), Polanyi (1958), Popper (1959), Toulmin (1961), Kuhn (1970), Feyerabend (1975 and before) and other wild characters including Quine (1951, 1969).

1. Judgmental, discretionary components are unavoidable in science. They appear in the choice of experimental design, the choice of a specific apparatus, the wording of the particular questions in our questionnaires, in the interpretation of results, and in the choice between competing theories. These subjective discretionary links cannot be avoided. Logical positivisim wanted to remove all discretion. This effort to achieve foundationalist explicitness took two forms: completely explicit observational foundations (meter readings, sense data, and so on) and logical deductive manipulation of these sense data. Logical positivism failed at both levels.

Campbell and Stanley (1966, p. 35) joined in this rejection of logical positivism when they said that "true" experiments at their very best only *probe* theories; they do not *prove* them. But the rejection was most important in our emphasis upon the role of *plausibility*. We took the position that there could be lots of threats to validity that were logically uncontrolled but that one should not worry about unless they were plausible. The general spirit was that any interpretation of a body of data or research procedure should be regarded as innocent until judged guilty for plausible reasons, as determined through the scientific method of mutual criticism.

I've often wondered why there were no hostile logical-positivist reviews of Campbell and Stanley, accusing it of undermining scientific standards. We failed to get one as far as I know. It is with mixed pride that I note we are now regularly being used as an exemplar of logical positivism, and of the mistaken effort to import into the social sciences the inappropriate methods of the physical sciences. (While I am grateful for every citation, I think this is a misreading, as will be argued below.)

2. The paradigm theme. We are inevitably encapsulated in some paradigm of presuppositions, inexplicit or explicit. Historically, we can look back and see how provincially we were imbedded. We cannot do with presuppositions. We cannot pull each presupposition out individually and prove them one at a time. In every expansion of scientific knowledge we have to expand the number of things we assume are true and that have to go unproven. In the evolutional-epistemology version of this, with the recipe of variation, selection, and retention, there is emphasis upon the presuppositions about the nature of the world that are built into our retinas, the nerve wiring of our brains, our language, and our own research tradition. From evolutionary epistemology comes the crucial question of balance between variation and reten-

tion. These are incompatible, and knowledge becomes impossible if either totally dominates.

In accepting paradigm-embeddedness again we are rejecting the *foundationalism* that was so central to logical positivism. There are no untouchable axioms: All are criticizable and revisable. Nor are there any foundational observations or facts. There are indeed at any historic period of time in any successful science a vast array of trusted facts, but none is immune from revision. For the atomistic (sense data, observations, or axioms) foundationalism of the positivists, we must substitute a holistic, squishy, quasi-foundationalism, a composite foundationalism that I call the 99 to 1 trust-doubt ratio. This is like the holistic network imagery of Quine (1951), but I'll give it to you in my version.

For the cumulative evolutionary process of knowing, our only available tactic is to trust most of our current beliefs while we use that distributed fulcrum to revise a few of them. The ratio of the trusted to the doubted has got to be in the order of 99% trust to 1% doubt. In biological evolution, 99% of the genes are trusted while mutation and recombination vary 1% of them. However wrong-headed the initial beginnings, nature is stuck with this great mass of presuppositions on how to design an animal. Similarly, in a science such as physics, the great revolutions have been achieved by trusting 99% of the cumulated facts and using that basis to revise 1% of its beliefs and their theoretical integration. This produces a kind of gradualism at the level of facts (wherein lies my only real disagreement with Kuhn).

Don Moyer (1979) has studied the belief changes following the 1919 eclipse observations, where English physicists and astronomers moved from 5% adoption of Einsteinian general relativity to 99% adoption in a five-year period. He documents the ways this revolution was based upon profound trust of previous physics, which provided the factual leverage for overthrowing the dominant Newtonian theory. It was palace revolution in conceptual organization and theory, in which most of the facts (all being theory-laden "facts") were retained.

Before going on to the other three points, I would like to use these first two points (paradigm dependence, and discretionary human judgments) to discuss the qualitative versus quantitative agenda which is so important right now in educational research and program evaluation. Should we be using the methods of the humanities or the methods of the physical sciences? I would like to argue that if we had not misread the record of the physical sciences we would recognize that these methods are very similar.

Let us start out with that old tradition, at one time called philology and now called hermeneutics, which asks such questions as, What did Homer mean by this particular phrasing? or, What did Saint Paul mean in this particular verse? In philology and hermeneutics, one had generations of scholars quarreling about these issues, but remaining within the same social communication net, a quarreling collective committed to getting the truth. Now, part of

this hermeneutic tradition is this presuppositional and contextural dependence that I have called the 1 to 99 doubt-trust ratio, a composite fallibilist foundationalism generating and criticizing plausible rival hypotheses as to alternative interpretations, including the hypothesis that some copyist had made a clerical error that was subsequently transmitted by other loyal copiers, and so on. This self-critical community of interpreters, by looking at a wider range of manuscripts from this same time, and thus extending the grounds of judgment, often eventually arrived at consensual decisions as to the best interpretations of a particular manuscript.

Or, look at the method of the historians as taught and exemplified by Collingwood (Levin, 1970), who was a historical relativist with a historical paradigm theme. His method was explicitly the method of a detective in a detective story. The method is epitomized by trying to rule out plausible rival hypotheses.

When we get down to our own practical work, a plausible-rival-hypothesis approach is absolutely essential, and must for the most part be implemented by common-sense, humanistic, qualitative approaches. In program evaluation the details of program implementation history, the site-specific wisdom, and the gossip about where the bodies are buried are all essential to interpreting the *quantitative* data (Campbell, 1974b, 1975, 1979a; Cook & Reichardt, 1979).

3. Historicism. At any given time, even in the best of science (even in physics), we are in a historical context and our experiments and our theoretical arguments are historically imbedded. They have a historical provincialism; they are reactions to what has gone before; they are dated and uninterpretable outside of that context. The contrasts with the past are, in some kind of a problem-solving way, almost necessarily exaggerated. So we have a dialectic of contrast, in which exaggerated, oversimplified corrections for what has gone before are an essential part of the process, and the past that has gone before is essential for understanding the new terms and new experiments that are introduced. In an effort to speak in the extreme forms of post-positivist jargon, I have called this the "dialectic historical indexicality of scientific terms" (Campbell, 1982). Gergen (1982) presents the historicist argument for social psychology.

4. Relativism. This treasure of post-positivism encompasses epistemological, historical, cultural, and paradigm relativism. In the evolutionary epistemology tradition (Campbell, 1974a) my slogan is "blind variation and selective retention." This is an emphasis on exploring in the dark, with the fumbling of a blind person being a better model for epistemology than clairvoyant vision. All of this commits me to a profound epistemological relativism.

Now, while I am a thoroughgoing epistemological relativist, I reject an ontological relativity, or, since Quine (1969) has used that term in a different sense, an ontological nihilism. Evolutionary epistemology has in it an unproven assumption of a real world external to the organism, with which the organism is in dialectical interaction. I have been spending a lot of time recently reading

and meeting with (Campbell, 1981) exciting young sociologists of science such as Barry Barnes (1976), David Bloor (1976), and Michael Mulkay (1978), Karen Knorr-Cetina (1981), Bruno Latour and Steven Woolgar (1979), and Harry Collins (1981). Also relevant is the book Robert Merton and Thomas Kuhn have resurrected, Ludwig Fleck's 1935 *Genesis and Development of Scientific Fact* (1979). Harry Collins calls this the "relativist" program in the sociology of science. Latour and Woolgar and Knorr-Cetina call it the "social constructivist" program. David Bloor and Barry Barnes (Bloor, 1976; Barnes & Bloor, 1982), call it the "strong programme," meaning that in doing sociological, historical studies of science (asking the question, What were the causes for their changing their scientific beliefs?) it is illegitimate to use our current confidence in the truth of the belief as an explanation for why, back then, they came to believe it.

This agnosticism I find methodologically correct. After all, those past scientists were not clairvoyant, and many of the changes we now regard as in the mainstream of scientific development we do not now regard as "true". But these new sociologists of science carry this agnosticism too far. They refuse to speculate in an ontologically-realist way about what kinds of social processes, what kinds of systems of interaction among scientists and between scientists and society, could produce *improved* beliefs. They refuse to undertake what I call an epistemologically relevant internalist sociology of science (Campbell, 1979b, 1981). I am continuing to work on such a sociology of science (Campbell, 1984).

5. Sociologism and psychologism. Science is a social process, scientists are thoroughly human beings: greedily ambitious, competitive, unscrupulous, self-interested, clique-partisan, biased by tradition and cultural memberships, given to mutual backscratching, and the like. James Watson's *The Double Helix* (1968; but see Olby, 1974, for Crick's perspective) is one of the most used texts in the sociology of science relevant to this.

Out of this, I want us to keep the goal of *truth,* and to attempt to understand and foster a social system of science (differing greatly from our recent dominant theory of applied social science for policy purposes) in which it becomes sociologically plausible that the processes would lead to beliefs of increasing validity. The scientific method itself is a social system product. Science is itself a social system, it is "tribal" in that sense (Campbell, 1979b), but strangely, its norms preach against that very tribalism: against deference to authority, against deference to majority rule. A key part of this sociology of successful science is a mutal criticism that keeps those who are criticizing each other still remaining in the same group, rather than splitting off into their own insulated cults. Competitive replication, threat of replication, a reward system that encourages competitive innovation but punishes dishonesty in the resulting competition (Merton, 1973) are all parts of it (Campbell, 1984).

From this sociological point of view, combined with an evolutionary-epistemology point of view, it follows that large numbers of independent decision

makers are essential for objectivity in science. It follows too that we must maintain scientists' collective interests in the trust given the system of science by the larger public (Merton & Geiryn, 1982). We must maintain the individual scientist's interest in reputation, recognition, and fame, without allowing these interests to undermine the self-interest in science's collective validity. We scientists cannot avoid being dependent on the trust of fellow scientists. We must avoid creating a motivational system that generates truth claims or belief assertions that we distrust. We need a scientific method (as a social invention and social process) that will counteract the ill effects that a cynical and nihilistic interpretation of point 4 (relativism) and point 5 (sociologism and psychologism) can produce.

This epistemologically relevant internalist sociology of science will not deny the scientist's paradigmatic provincialism, self-seeking competitiveness, and human fallibility, but will rather propose a social system designed to curb side effects that produce invalid beliefs. Inevitably our model of science will show science as a fragile and vulnerable social institution, one that is capable of flourishing only now and then, only here and there, on the face of the earth. A validity-producing social system of science is nothing we should take for granted.

APPLICATION TO APPLIED SOCIAL SCIENCE

If we move such a post-positivist theory of science into the problem of the validity of applied social science, we find that we need all of the social system features of pure science (e.g. physics, laboratory psychology, and biology). From this perspective, when we move into the arena of policymaking, there are some regular features of applied social science for policy purposes that come to our attention.

First is clearly the *greater equivocality of causal inference for research done in policy settings.* There are many, many more plausible rival hypotheses. There is much less control. Looking back at the "artificiality" of physical science laboratories (their soundproof walls, atmospheric controls, insulation against electromagnetic and magnetic fields, achievement of vacumms, and all of the other accoutrements of "experimental isolation"), we can see that all of this laboratory apparatus is designed to control or to rule out *plausible rival hypotheses,* or at least to render them "implausible," thus achieving an *artificial* situation in which causal inference can be done more competently.

When biologists left the insulated laboratory where apparatus and walls are the essence of the scientific method, to move out into the agricultural experimental station where the winds blew and the rains rained, they invented another type of artificiality to render implausible large classes of plausible rival hypotheses. This was the *randomized* experiment. We should note that slightly before that, educational researchers such as E. L. Thorndike and his students, moving from the insulted psychology laboratory out into the classrooms, independently invented randomized assignment to experimental treat-

ments and latin-square designs, again as artificialities that operated somewhat like experimental isolation in generating controls, in reducing the plausibility of rival hypotheses such as selection, selection-treatment interaction, practice effects, and the like. While we educational psychologists did not do it with Fisher's mathematical elegance, we were *first* with these great tools of artificiality. McCall's (1923) *How to Experiment in Education* summarizes this early achievement.

Today, as so many of us react to the frustrations of social science research with the hope that humanistic methods will turn out to be more appropriate than physical science ones—an exploration that I too favor (Campbell, 1974b, 1975, 1984), our troubles are often blamed on a prior, mistaken, subservient borrowing of physical science methods. Indeed, Campbell and Stanley (1966) are often accused of this fallacy. Close analysis will, I believe, show that this is unfair. Thorndike and McCall were *not* borrowing random assignment and the "rotation experiment" (latin-square) from the physical sciences, nor from R. A. Fisher and his agricultural experiment stations. Instead, they were reacting to the mutual criticisms of their own educational-psychology research community, and inventing research designs that would help rule out the recurrent very plausible rival hypotheses generated by their fellow critics.

So too, Campbell and Stanley's list of threats to validity is an accumulation of our field's criticisms of each other's research. The list of quasi-experimental designs is a cumulative listing of our discipline's inventions of ways of ruling out some of the very plausible rival hypotheses. We can thank Campbell and Stanley for being conscientious collectors of the achievements of this tradition of collective self-criticism. (That's what they were: collators, bookkeepers, reviewers of the literature.) Their collection of designs is not at all drawn from physical science. Of course, from the quasi-experimental perspective, just as from that of physical science methodology, it is obvious that moving out into the real world increases the number of plausible rival hypotheses. Experiments move to quasi-experiments, and on into queasy experiments, all too easily.

A second difference between applied social science and laboratory research is that the still greater likelihood of *extraneous, nondescriptive interests and biases* entering through the inevitable discretionary judgmental components that exist in all science at the levels of data collection, instrument design and selection, data interpretation, and choice of theory. As we move into the policy arena there is much less social-system-of-science control over such discretionary judgment favoring descriptive validity, and there are much much stronger nondescriptive motives to consciously or unconsciously use that discretionary judgment, to, so to speak, break the glass of the galvanometer and get in there and push the needle one way or the other so that it provides the meter-reading wanted for nondescriptive reasons (Campbell, 1982, 1979a, pp. 84–86).

The next few points about moving the theory of science into the applied social science arena stem in considerable part from the seriously mistaken model of applied social science that we social science methodologists offered

to ourselves and to government in the 1960s, in the period of the Great Society, in the era of the Office of Economic Opportunity, and Program Planning Budgeting and Systems. Many of these I have gone over on previous occasions (Campbell, 1974b; 1979a).

My third point is the *mistaken belief that quantitative measures replace qualitative knowing.* Instead, qualitative knowing is absolutely essential as a prerequisite foundation for quantification in any science. Without competence at the qualitative level, one's computer printout is misleading or meaningless. We failed in our thinking about program evaluation methods to emphasize the need for a qualitative context that could be depended upon. One example is frequent separation of data collection, data analysis and program implementation that was once characteristic of Washington's funding of programs, in which one firm would collect the pretest, another firm would collect the posttest, and a third firm would analyze the data. This easily led to a gullible credulity about the numbers on the computer tape, with the analyst in total innocence about what was actually going on in the program implementation and testing situations.

To rule out plausible rival hypotheses we need situation-specific wisdom. The lack of this knowledge (whether it be called ethnography, or program history, or gossip) makes us incompetent estimators of program impacts, turning out conclusions that are not only wrong, but are often wrong in socially destructive ways.

Fourth, the evaluation model we offered mistakenly bought into the logical positivist's *definitional operationism,* specifying as program goals fallible measures open to bureaucratic manipulation (Campbell, 1969a; pp. 414–417, 1979a, pp. 84–86).

Fifth, *a one decision/one research* ideal was a central feature of our original program evaluation model. (This is diametrically opposed to the social system of the successful physical and biological sciences.) Each program evaluation was to be done to support a specific administrative decision. One researcher-evaluator was to have a monopoly on the resulting truth claims. This one study was to be the basis for the decision. With this often went a disregard of prior wisdom and prior science in making the decisions about the future of the program (Lindblom & Cohen, 1979). The program evaluation was conceptually tied to refunding, to be the sole or an important base for expanding or contracting the program.

Such a policy violates common sense as well as the sociology of knowledge. Had we sat down and thought, What will it do to all of those discretionary points in data collection if next year's funding is going to ride on them? Where are the discretionary points and how can they be distorted?, we would have recognized that this program evaluation model belied our common experience, the sociology of bureaucracy (Blau, 1955, 1956; Ginsberg, 1984) and of our knowledge as psychologists as to the multiple motives the individuals implementing programs have, including the motive of being able to feed one's children next year. ("Where will another job come from if this program is

discontinued?" or, "If we report to our client our unpleasant results, where will next year's contract come from?" and so on.) These considerations add into the recurring conflict we all have observed between the evaluation staff and the program delivery staff. Program evaluation became destructive of program delivery morale.

A sixth mistake in the model that we in the 1964–1968 period recommended to government was the emphasis on *external evaluation* of programs rather than evaluation by the delivery team itself. This again is the complete opposite of the social customs of the physical sciences, in which passionate believers in new theories design the research and carry it out. The objectivity of physical science does not come from turning over the running of experiments to people who could not care less about the outcome, nor from having a separate staff to read the meters. It comes from a social process that can be called competitive cross-validation (Campbell, 1984), and from the fact that there are many independent decision makers capable of rerunning an experiment, at least in a theoretically essential form. The resulting dependability of reports (such as it is, and I judge it usually to be high in the physical sciences) comes from a social process rather than from dependence upon the honesty and competence of any single experimenter. Somehow in the social system of science a systematic norm of distrust (Merton's [1973] "organized skepticism") combined with ambitiousness leads people to monitor each other for improved validity. Organized distrust produces thrustworthy reports. In contrast, in program evaluation, the monopoly of a single evaluation for each program, with but one decision maker to use it, and the dogma of external evaluation, all combined to make impossible this crucial aspect of the social system of the successful sciences.

Another type of mistake involved *immediate evaluation,* evaluation long before programs were debugged, long before those who were implementing a program believed there was anything worth imitating.

A totally unnecessary feature was recommending a *single national once-and-for-all evaluation* that would settle the issue forever.

Point nine: There was *gross overvaluing of, and financial investment in, external validity,* in the sense of representative samples at the nationwide level. In contrast, the physical sciences are so provincial that they have established major discoveries like the hydrolysis of water (in which electrical anodes and cathodes generate bubbles of oxygen and hydrogen) by a single water sample taken from the Soho neighborhood of London in 1903 (see Campbell [1969b] for a more extended and complex discussion), never cross-validating the discovery on a "respresentative sample" of all of the water of the world.

The so-called Northwestern School—whose center of strength is still at Northwestern with Bob Boruch, Tom Cook, and their colleagues, and within which I still include Lee Sechrest, Paul Wortman, myself, and most of our Northwestern Ph.D.s—has been criticized for overemphasis of internal validity at the expense of external validity. This accusation must be, in a historical

sense at least, wrong. Who, after all, introduced the great emphasis on, itemized all of the threats to, and assembled the controls for external validity (Campbell, 1957; Campbell & Stanley, 1963, Cook & Campbell, 1979)? Of course, we are interested in external validity, but we see no point in having a representative national sample of a repeated regression artifact, or of some other internally invalid research design.

Tenth, is *the neglect of the fact that scientific truths are a collective product of a community of scientists at any given time.* Such a community is self-critical, gets into the guts and looks under the cover and tries to decide what was going on in specific experiments. There was a neglect of this insulating layer of human judgments that are well informed and mutually disciplined. We somehow assumed in our OEO-PPB&S model that a single computer output could speak directly to the administrator. Now, however, as post-positivist fallibilist critical realist, we want our realism to include the real and fallible processes of data collection and conclusion drawing. We can see vision as the product of imperfect lenses, imperfect nervous systems, and oversimplified presumption systems, which lead to generally valid perceptions but also to optical illusions (Campbell, 1983).

This physicalization, this materialization of the process of knowing, is a very important part of the historical development of epistemology. Extended to science we should have seen from the very beginning that social data collection and social experimentation were social system intrusions into the ongoing processes, and that putting policy-decision pressure on them would distort every crushable, squishy, little discretionary link. We were guilty of a doctrine of "immaculate perception" (as it has been called in epistemology), guilty of assuming a noninteractive acausal observational process in which all of our questionnaries and arrangements could describe without disturbing what they were describing, and in which the people being described as well as the describers would be unmotivated to bias the meter readings.

BETTER STRATEGIES FOR APPLIED SOCIAL SCIENCE

Our post-positivist theory of science with its social system of science emphasis is far from complete. Nor have we yet applied it adequately as an alternative ideology for applied social science, ready as advice to Washington whenever the spirit of the experimenting society, that existed under the regime of the good President Johnson, returns. To be so ready, we must start arguing now about the pros and cons of alternative models. To help initiate this I offer the following.

1. I'll call the first alternative the contagious cross-validation model of program evaluation. A generous and concerned government provides funds for developing local programs addressed to chronic sores of society. This local program funding includes funds for whatever evaluation the program designers want, including funds for academic consultants. Lots of local programs result.

When any one of them, after a year or so of debugging, feels they have something hot, a program worth others borrowing, we will worry about program evaluation in a serious sense. Our slogan would be, "Evaluate only proud programs!" (Think of the contrast with our present ideology, in which Washington planners in Congress and the executive branch design a new program, command immediate nationwide implementation, with no debugging, plus an immediate nationwide evaluation.)

When the high morale program and program results were disseminated, there would no doubt emerge a group of willing adopters. (Note that before we had our program evaluation ideology, such borrowing was usually on the basis of persuasive program *plans,* and took place prior to even the first full tryout, as Addie and Murray Levine [1970] have documented so well in one classic instance.) At this stage, our federal funding would support adoptions that include locally designed cross-validating evaluations, including funds for appropriate comparison groups not receiving the treatment. (We might at this or the next stage have large-scale "external" evaluations, as long as these did not preclude interpretable comparisons at each site not depending upon full national implementation.)

After five years we might have 100 locally interpretable experiments. We would also have a community of applied social scientists familiar with them all, that had cross-examined each others' data, suggested and done reanalyses, performed bias-sensitive meta-analyses, and so on. Many of these scholars would be tenured university or public school faculty, whose job security would not depend upon the outcome. From the consensus of this mutually monitoring research community we would advise government and potential adoptors.

I leave it an open question whether or not the full-scale dissemination of a clearly successful program would be done without local cross-validation by adopters. Fully facing the problems of external validity, and the social historicity (Gergen, 1982) as to what will work when, would require this. I do believe we could make it feasible for many programs, and provided classroom teachers, for example, with realistic means of evaluating the competence of their own practice, albeit usually without synchronous parallel comparison groups except for exploratory innovations.

By moving the primary evaluation to the dissemination stage, we are evaluating the transferable, borrowable aspects of the program. In the initial zeal of program developers, exceptional success is frequently due to heroic 80-hour weeks on the part of key staff, and these are not aspects of the transferable program. We need to know about effectiveness for the program's routinized form. While the problem of generalizing in applied science is substantially different than in theoretical science, one essential of the "knowledge" produced is still reusability on different occasions and times.

The contagious cross-validation model is much closer to the model of the physical sciences, as noted in the previous section under point 6. Let us

remember that applied social science has more, rather than less, need for mutual criticism, argumentative reanalysis, and cross-validation than does physical science. This is just because we lack the possiblity of experimental isolation, just because our data have to be generated through the cooperation of persons with strong stakes in the outcome, and just because science (either physical or social) is done in an arena in which the rival interests in what the outcome is are so powerful that objective description can become a minor motive.

Let me give a concrete illustration that is banal and simple. I was an observer from a nearby but safe distance from the Chicago school system for many years. Here they were spending millions of dollars on testing programs that used national norms for an annual humiliation of half of the grammar schools in the city. That testing program was destructive in its net effect. The annual humiliation did nothing to improve the schools, told them nothing about what they could do to make education better, and put tremendous corruption pressure on test administrations. (Rumored practices were to classify as many children as possible as abnormal ineligibles, and to manipulate the time schedules to optimize performance). Thus the annual humiliation was destructive both of the validity of measurement and of the morale of the teachers.

While there were continual plans and expenditures for individual student data retrieval, neither the school system nor we designers of quasi-experiments ever provided a teacher with the ability to tell whether the text chosen for the current year was better than last year's.[2] We could have also provided individualized data retrieval disguising the scores so that no one knew what they were in terms of national norms, providing a comparison base for teachers based solely upon the previous performance of their own pupils. No national-norm humiliation need have been involved—merely an ability to tell whether one was doing better than last year. Such de-normed retrieval capability would also have provided adventuresome teachers the capacity to try out alternatives in teaching style. It's a great failure that we never got around to doing this. We program evaluation methodologists never provided the perspective nor the conceptual tools, nor lobbied the school system for this usage and against the other.

2. *Getting competitive replication into national policy pilot studies.* The contagious cross-validation model is appropriate only where the program under study can be implemented autonomously by a local unit (be it school, classroom, city, retail store, or factory). Where the program being piloted has to be eventually implemented nationally, different sources of competitive cross-validation must be sought. I am thinking of such heroic studies as the New Jersey Negative Income Tax Experiment (Watts & Rees, 1977; Kershaw & Fair, 1976; Pechman & Timpane, 1975; Rossi & Lyall, 1976) and the several subsequent still larger experiments with guaranteed annual incomes in rural North Carolina, Gary, Seattle, and Denver. Belonging here too are the Hous-

ing Allowance Experiments (see Lowry [1982] for the Supply experiments, Abt Associates and the Urban Institute for the Demand experiments) and the big health insurance experiments. We need such enormous studies, but should run them in the future with deliberate efforts to build in some degree of independent replication and mutual monitoring. Here are several ways this might be done.

A. Rather than awarding a single contract, each should be *split into two or more independent experiments,* so that all of the hundreds of discretionary decisions as to how to present the experimental treatment and design the questionnaires and interviews would be made and implemented by at least two independent research teams. Such heteromethod replication (Campbell, 1969b; Cook & Campbell, 1979) is needed for interpretive validity. It would also provide a small group of informed scientists for competitive cross-examination.

B. There should be *adversarial stakeholder* participation in the design of each pilot experiment or program evaluation, and again in the interpretation of results (Krause & Howard, 1976; Bryk, 1983). We should be consulting with the legislative and administrative opponents of the program as well as the advocates, generating measures of feared undesirable outcomes as well as promised benefits.

C. There should be *competitive reanalysis* of data from the big studies. The Office of Economic Opportunity set a great precedent to which we have inadequately responded. The Institute for Research on Poverty, University of Wisconsin, has available for reanalysis the data tapes for the New Jersey Negative Income Tax Experiment, and proper scientific disagreements are emerging, for example, as to how they handled the attrition problem (Boeckmann, 1981). The have the data from the first big Head Start evaluation, a data set with a fine record for productive second-guessing (Smith & Bissell, 1970; Barnow, 1973; Magidson, 1977; Bentler & Woodward, 1978). I hope they have the big Performance Contracting study (Gramlich & Koshel, 1975) with the rebuttals from the performance contractors. Major classics in this area come from my Northwestern colleagues (Cook et al., 1975; Boruch, 1978; Boruch et al., 1981; Trochim, 1982).

The original Coleman report (1966) on educational desegregation has been thoroughly reanalyzed, so that now we could assemble a half-dozen volumes the size of Mosteller and Moynihan's (1972); and from a modern post-positivist theory of science, we can recognize that only now do we have a competent applied social science community ready to use the Coleman report in conjunction with all related research, prior and subsequent, to guide governmental policy. The original image of one research (one data collection, one analysis, by one analyst team) to guide one governmental decision, was based upon a fallacious theory even for pure science, and still more wrong for applied social science.

While these secondary analyses are of great value, and should become obligatory for all expensive data collections, we should remember that they cannot fully correct for the hundreds of idiosyncratic discretionary judgments involved in the initial data collection.

D. *Legitimating dissenting-opinion research reports* from members of the research team. The Freedom of Information Act of the late 1960s was one of the great social inventions increasing the possibility of a valid, policy-relevant, applied social science. While Rights of Subjects legislation (another great innovation) has been used to greatly curtail its practical implementation (needlessly so—see Campbell, Boruch, Schwartz, & Steinberg, 1977; Boruch & Cecil, 1979, 1982) the legitimating value is still there. It should make possible competitive reanalyses. Indeed, the Seattle Teachers' Union had used it in demanding the data tapes from The Office of Economic Opportunity's (OEO) Performance Contracting Study before the final report was ready, and OEO had agreed to this in an out-of-court settlement. (This never lead to a rival analysis, in part at least because OEO's official analysis when it came out supported the interests of the teachers' union). In my unpublished but widely circulated "Methods for the Experimenting Society" (1971a), drawing upon the unpublished and minimally circulated Gordon and Campbell (1971), we argue that the voting booth rather than the rat lab should be the methodological model in policy research, and that the right to reanalyze data employed in governmental decision making is fundamentally related to the right to demand a recount in an election.

Another background for my argument is the great value that whistle-blowing has had for the validity of physical and biological research results when these have been done under conditions of extreme policy relevance. (I am thinking of research on the dangers of chemicals to manufacturing workers and food consumers, the dangers to and effects on humans and sheep of irradiation from nuclear experiments and power generators.) While such whistle-blowing occurs, it is still experienced as a guilt-producing team disloyalty, both by the whistle blower and coworkers, who may react with ostracism. It would improve the scientific and political validity of applied physics, chemistry, and biology if whistle-blowing were legitimated by reconceptualizing it as the right and duty to generate dissenting-opinion research reports, and if all laboratory staff were provided official access to all data for this purpose. Insofar as our research results are inherently more ambiguous, even more do we need this in applied social science.

I am making a radical suggestion, but one that we in the American Educational Research Association, the Evaluation Network, and the Evaluation Research Society, could right now be put into our guidelines on research ethics (Stufflebeam, 1981; Rossi, 1982). Moreover, we as individuals could start it now with our own research assistants. Imagine if you gave every research assistant (including the neurotic ones with negative Oedipal resolu-

tions whom you never should have hired in the first place) the right of access to all of the data and the right to generate minority reports. I have no doubt that this would increase the validity of the official report (as well as provide some of the needed competitive reevaluation). We research directors would write up our reports differently knowing that our righteous and sore-headed assistants were potentially free to dig up the items on the interview that we neglected in our final report, to dig up and publicize the disappointing analyses we failed to find room for in our final report, or to reanalyze the data with a different perspective. Our profession should start designing a model contract specifying such rights that could be given each employee when hired.

3. Writing up our evaluation research reports for our fellow evaluation researchers in and out of the universities, is my third suggested reform of our original OEO-PPB&S model of applied social science. I state it thus because we so often in those early days chided ourselves for letting our academic standards and interests get in the way of writing program evaluations geared to fluent administrative decision making. (I need help in assembling good examples of this literature.) While I am not attempting to condone irrelevant "pure" research smuggled in under the applied budget, I am insisting on having available (along with the data available for reanalysis) a full academic analysis for cross-examination by our applied social science colleagues.

Let me stress this through an aside to those of my students (face-to-face and by the printed word) who feel that the Campbell and Stanley superego has ill prepared them for life in the real world of program evaluation. Let your employer or the administrator whose neck in on the block write up the "executive summary." Be sympathetic to the social role and predicament of program administrators and developers. Do not be a "sadistician" (as one of our psychoanalysts might diagnose it), forcing them to live up to your own most punitive standards of scientific rigor (note Devereaux's *From Anxiety to Method in the Behavioral Sciences* [1976]). You protect your own superego by signing your name to the 200-page appendix addressed to your fellow scientists. We too should be like the physical scientists who advise government from the consensus of a well-informed, mutually monitoring scientific community focused on the problem area. These appendices, proper government funding of conferences, and reanalyses in terms of the plausible rival hypotheses we generate, will provide an applied social science base that is more optimal, politically and scientifically.

The complete sociology of applied-science validity, which I wish I had, would take into account environmental impacts on commitments to validity which applied science careers involve. I will use this future agenda, and my earned status as an academic garrulous grandfather, to permit inserting here (rather than properly reorganizing this paper) some further advice on maintaining the Campbell and Stanley superego in program evaluation careers. It will help if one recognizes that our initial OEO-PPB&S rhetoric got fused

with a legislative and administrative rhetoric in a way that we should avoid being mousetrapped by.

Still today, governmental funds are needed to provide relief to the ill, aged, and underfed. Let us call this *problem-specific revenue sharing.* But it became politically necessary for such relief funds to be disguised as "new programs" that would cure the problems they were designed to alleviate. Including in the legislation the requirement that the "program" be "scientifically evaluated" became in many, many instances just a part of the escalated rhetoric, a routine part of assuring conscientious, responsible custodianship of governmental funds, on a par with requiring proper bookkeeping and auditing. In such cases the genuinely worthy goal was achieved when funds were spent locally on the problem. *Local fund-spending on the need was the real "program."* (Paying too much attention to pork-barrel motives supporting the same goal can distract from attending sympathetically to the local relief aspects, and the rhetorical requirements for providing for these needs.) Most such so-called programs involved no alternate disseminable program package. At best, funding and staff are added in ungeneralizable ways to preexisting agencies.

For these programs I recommend avoiding laying one's scientific superego on the line. Save up those negotiating energies and costs in interpersonal goodwill that comparable untreated comparison groups, meaningful pretests, and interpretable before-after comparisons involve to apply to that rare occasion when a potentially valuable innovation is being tried out, or that still rarer occasion when unique circumstances permit an impact assessment of current practice. For the "only-rhetorically programs," do evaluations that are low-cost in both rapport and money. Collect the opinions of well-placed observers as to what would have happened without the "program," and as to what aspects of it failed and what succeeded. Put in the final report appendix useful "input" descriptions. Include discussion on suggestions as to how promising disseminable aspects of local practice, or practitioners' suggested innovations as yet untried, might be implemented in the future in ways that might probe their usefulness.

For such non-programs, evade (if you can) producing any quantitative estimate of impact. If you cannot, a least in the long appendix surround them with full discussion of how the setting makes them equivocal. If a cost-benefit analysis is required, try to get this subcontracted to an economist or operations researcher whose training has not troubled his conscience with all of the plausible threats to the validity of the "benefit" estimates available. Avoiding quantified guessing in highly equivocal evaluation settings is a matter of political conscience also. Evaluation reports should enter into political decision-making processes as one component to multiparty argument and negotiation. Due to the general prestige of quantified science, not yet earned in our area, quantitative guesses and computer output carry more weight than they should in competition with the qualitative judgments of well-placed observers.

"Street wisdom," or theoretical understanding of the encompassing social system and the political realities (sympathetic to actors and roles, avoiding

hostile cynicism) are important components of our "methodology for the experimenting society" (Campbell, 1971a). It will help to remember what Rossi (1969) has taught us. The legislative and administrative setting is always one in which many needs are competing for funds. The most important needs may indeed have priority for funding. But importance means that these are stubborn, unsolvable, chronic problems on which normal societal problem solving has failed. The competition for funds almost guarantees that the tentative solutions for these urgent chronic problems will be underfunded.

If often seems that programs are designed and implemented just so as to preclude interpretable comparisons. This may indeed be so, and so because the designers and administrators have been aware (perhaps unconsiously) that the program could only be a drop in the bucket, and had no chance of living up to the claims for panacea that were politically necessary for getting even that drop (the "overadvocacy trap," Campbell, 1969a; 1971a; 1971b; Shaver & Staines, 1971). We program evaluators, expanding our methodological responsibilities beyond the narrow issues of experimental design (while not at all abandoning these concerns) to include a sociology of applied-scientific validity, must be sympathetic to this predicament. We must avoid reacting with the hostile disdain of wounded idealism and methodological righteousness. We must avoid this not only for the health of the social system in which we participate, but also for our own mental health. Our economic and career predicament may give us no alternative but to keep our job. The reaction of unsympathetic hostile disgust can trap us in self-contempt for prostituting our scientific skills and ideals. I believe we can avoid this by aspiring to a sympathetic understanding of our program director's and our own social-system predicament, and by working as best we can within that system to produce validly interpretable evaluations whenever feasible and when there is a potentially disseminable program alternative worthy of such efforts.

4. Avoiding "ad hominem" and "ad institutionem" research. A final radical shift I would like us to consider is the recommendation that we stop using our fragile tools of experimental design and measurement for purposes of managerial control and "accountability." (I thus reverse the implicit recommendations of my early [1956] view that leadership effectiveness is a causal hypothesis to be demonstrated optimally by quasi-experimental methods.) Financial costs are one reason; these tools are too expensive to be used for personnel selection purposes (for selecting the better teacher, principal, superintendent), nor is quasi-experimental comparability likely to be available to make such data interpretable as an effect of skill, effort, and merit. Nor can we really solve our organizational problems by promoting effective persons out of their current locations of effectiveness. There are not enough dedicated geniuses. Overall, we must improve organizations by discovering optimal use of the energies and abilities of current staffs, rather than by hiring those of proven effectiveness away from their current jobs.

But my main reason for recommending that we exclude the research goals of evaluating institutions, social organizations, and persons, is my conviction

that this use, beyond all others, corrupts the validity of the measures, and may also corrupt the very social processes the measures are designed to monitor (Campbell, 1979a, pp. 84–86; Blau, 1955; Ginsberg, 1984). We are thoroughly dependent upon the staffs we evaluate for the qualitative background required by discretionary judgments, as well as for generating much of the data. The social control, organizational management, and personnel evaluation purposes maximize the nondescriptive motives, the motive to influence the decision rather than (or in addition to) provide a valid description. It would be my thesis and hope that these distortion pressures are at minimum when what is being evaluated is a *program,* an alternative that present staffs could adopt without losing their jobs. *Let us evaluate alternative programs, not persons or social units.*

This principle of post-positivist applied social science obviously also supports again the abandonment of the single evaluation, single-decision model, and the decoupling of evaluation from refunding decisions, or a radical reversal of the present coupling. I return to my near-but-safe-distance observation of the old Chicago: It was my sincere judgment that there would have been a substantial saving of program and evaluation funds had the evaluation-funding linage read, "In the event of no-effect or undesirable-effect outcomes, the same staffs should continue to work on the same problem with an alternative program, and with a 10% budget increase above inflation." (We would, of course, have needed econometric tuning of that percentage to avoid pressures toward faking failure.)

CONCLUSION

The problem is turned over to you unfinished and inadequately formalized. But I hope that I have convinced you that we need sociology of scientific validity, and an applied social science specialty within it, as a part of the methodology we bring to our tasks. I hope that you share my conviction that this can be done in a way that still makes valid applied social science possible (or, at very least, that we can produce beliefs of enough improved validity and subtlety to make continuing in our profession worthwhile). If you are convinced of both need and possibility, I call upon you vigorous youngsters to take up the task of creating an adequate social theory of validity-increasing applied social science. But if you are convinced of the impossibility, then it is your moral duty to publicly denounce the pseudo-science in which we inadvertently find ourselves engaged. Let us at very least create around the problem a mutually monitoring, disputatious community of scholars who listen carefully to each other's arguments and rebuttals.

NOTES

1. Suppes's "Facts and Fantasies of Education" (1973, pp. 14ff.) comes closest. Listing us under "second order fantasies," he chides us, sympathetically, for offering no reasons for our "wholly enthusiastic support of experiments," no "abstract principles for which ... principles of experimentation are derived," no "collection of empirical evidence bearing on the theory of experimentation," no "defense of the reasons for randomizing in experiments," and as needing "derivation from first principles in at least one example." While Cook and Campbell (1979) may have gone part way to meeting these and Suppes's other objections, probably Charles Reichardt's (1983) as yet unpublished paper best fills the conceptual gap he and others have noted.

2. Chicago, for all its reputation for corruption, still allowed teachers a list of texts they could choose among, so they could have experimented with textbooks. Textbook evaluation is a good place for a science of program evaluation to cut its teeth. A text obviously differs depending on who is using it, but still it is a relatively specifiable and disseminable program package.

REFERENCES

Barnes, B. (1976). *Scientific knowledge and sociology theory.* London: Routledge & Kegan Paul.

Barnes, Barry, & Bloor, David. (1982). Relativism, rationality, and the sociology of knowledge. In M. Hollis and S. Lukes (Eds.), *Rationality and relativism.* Oxford: Blackwell.

Barnow, B. S. (1973). The effects of Head Start and socioeconomic status on cognitive development of disadvantaged children. Unpublished doctoral dissertation, University of Wisconsin, Department of Economics.

Bentler, P. M., & Woodward, J. A. (1978). A Head Start reevaluation: Positive effects are not yet demonstrable. *Evaluation Quarterly,* **2,** 493–510.

Blau, P. (1955). *The dynamics of bureaucracy.* Chicago: University of Chicago Press.

Blau, P. (1956). *Bureaucracy in modern society.* New York: Random House.

Bloor, D. (1976). *Knowledge and social imagery.* London: Routledge & Kegan Paul.

Boeckmann, Margaret E. (1981). Rethinking the results of a negative income tax experiment. In R. F. Boruch et al. (Eds.), *Reanalyzing program evaluations* (pp. 341–363). San Francisco: Jossey-Bass.

Boruch, Robert F. (1978). *Secondary analysis (New Directions for Program Evaluation,* No. 4). San Francisco: Jossey-Bass.

Boruch, R. F., & Cecil, J. S. (1979). *Assuring privacy and confidentiality in social research.* Philadelphia: University of Pennsylvania Press.

Boruch, R. F., & Cecil, J. S. (Eds.). (1982). *Solutions to ethical and legal problems in social research.* New York: Academic Press.

Boruch, R. F., Wortman P. M., Cordray, D. S., & Associates. (1981). *Reanalyzing program evaluations.* San Francisco: Jossey-Bass.

Bryk, Anthony S., (Ed.). (1983). *Stakeholder-based evaluation (New Directions in Program Evaluation,* No. 17). San Francisco: Jossey-Bass.

Campbell, D. T. (1956). *Leadership and its effects upon the group.* Ohio Studies in Personnel, Bureau of Business Research Monograph No. 83. Columbus: Ohio State University.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin,* **54**(4), 297–312.

Campbell, Donald T. (1959). Methodological suggestions from a comparative psychology of knowledge processes. *Inquiry,* **2,** 152–182.

Campbell, Donald T. (1969a). Reforms as experiments. *American Psychologist,* **24,** 409–429.

Campbell, Donald T. (1969b). Prospective: Artifact and control. In R. Rosenthal & R. Rosnow (Ed.), *Artifact in behavior research* (pp. 351–382). New York: Academic Press.

Campbell, Donald T. (1971a). "Methods for the experimenting society." Presented at the meeting of the American Psychological Association, Washington, DC, September 5.

Campbell, Donald T. (1971b). Comments on the comments by Shaver and Staines. *Urban Affairs Quarterly,* **7**(2), 187–192.

Campbell, Donald T. (1974a). Evolutionary espistemology. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper.* LaSalle, IL: Open Court.

Campbell, Donald T. (1974b). Qualitative knowing in action research. Kurt Lewin Award address, Society for the Psychological Study of Social Issues, New Orleans, September.

Campbell, Donald T. (1975). "Degrees of freedom" and the case study. *Comparative Political Studies,* **8**(2), 178–193.

Campbell, Donald T. (1977). Descriptive epistemology: Psychological, sociological and evolutionary. William James Lectures, Harvard University. (Unpublished, duplicated copies available)

Campbell, Donald T. (1979a). Assessing the impact of planned social change, *Evaluation and Program Planning,* **2**, 67–90.

Campbell, Donald T. (1979b). A tribal model of the social system vehicle carrying scientific knowledge. *Knowledge,* **2,** 181–201.

Campbell, Donald T. (1981). ERISS Conference, *4S Newsletter* (Society for the Social Studies of Science), **6**(3), 24–25.

Campbell, Donald T. (1982). Experiments as arguments. *Knowledge,* **3**(3), 327–337.

Campbell, Donald T. (1983). Varieties of neurological embodiments of knowledge in an evolutionary epistemology. Prepared for inclusion in A. Shimony, D. Nails, & R. S. Cohen (Eds.), *Naturalistic epistemology: A symposium of two decades.* In preparation.

Campbell, Donald T. (1984). Science's social system of validity-enhancing collective belief change and the problems of the social sciences. 1984 Prepared for inclusion in D. W. Fiske & R. A. Shweder (Eds.), *Pluralisms and subjectivities in social science.* In preparation.

Campbell, D. T., Boruch, R. F., Schwartz, R. D., & Steinberg, J. (1977). Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. *Evaluation Quarterly,* **1**(2), 269–299.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* **65,** 81–105.

Campbell, D. T., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research* Boston: Houghton Mifflin. (Originally published 1963).

Coleman, J. S. et al. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Department of Health, Education and Welfare, Office of Education.

Collins, H. M. (1981). Stages in the empirical programme of relativism. *Social Studies of Science,* **11,** 3–10

Cook, T. D., Appleton, H., Conner, R., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *Sesame Street revisited: A case study in evaluation research.* New York: Russell Sage Foundation.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings.* Boston: Houghton Mifflin.

Cook, T. D., and Reichardt, C. S. *Qualitative and quantitative methods in evaluation research.* Beverly Hills, CA: Sage Publications.

Devereaux, George. (1967). *From anxiety to method in the behavioral sciences.* The Hague: Mouton.

Feyerabend, Paul K. (1975). *Against method.* London: NLB Press.

Fleck, Ludwik. (1979). *Genesis and development of a scientific fact.* Chicago: University of Chicago Press.

Gergen, Kenneth J. (1982). *Toward transformation in social knowledge.* New York: Springer-Verlag.

Ginsberg, Pauline E. (1984). The dysfunctional side effects of quantitative indicator production: Illustrations from mental health care. *Evaluation and Program Planning,* **7**(in press).

Gordon, A. C., Campbell, D. T., el al. (1971). Recommended accounting procedures for the evaluation of improvements in the delivery of state social services. Northwestern University Center for Urban Affairs. (Duplicated paper)

Gramlich, Edward M., & Koshel, Partricia P. (1975). *Educational performance contracting.* Washington, DC: The Brookings Institution.

Hanson, N. R. *Patterns of discovery.* Cambridge, MA: Harvard University Press.

Kershaw, D., & Fair, J. (1976). *The New Jersey income-maintenance experiment* (Vol. 1). New York: Academic Press.

Knorr-Cetina, K. D. (1981). *The manufacture of knowledge: An essay on the constructivist and contextual nature of science.* Oxford: Pergamon.

Krause, Merton S., & Howard, Kenneth I. (1976). Program evaluation in the public interest. *Community Mental Health Journal,* **5,** 291–300.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Latour, Bruno, & Woolgar, Steve. (1979). *Laboratory life: The social construction of scientific facts.* Beverly Hills, CA: Sage Publiciations.

Levine, Adeline, & Levine, Murray. (1970). Introduction to the new edition. In Randolph S. Bourne (Ed.), *The Gary Schools.* Cambridge: MA: Massachusetts Institute of Technology Press.

Levine, Joseph M. (1978). The autonomy of history: R. G. Collingwood and Agatha Christie. *Clio,* **7,** 253–264.

Lindblom, C. E., & Cohen, D. K. (1979). *Usable Knowledge.* New Haven, CT: Yale University Press.

Lowry, Ira S. (1982). *Experimenting with housing allowances: Executive summary* (Report No. R-2880-HUD). Santa Monica, CA: Rand Corporation.

Magidson, J. (1977). Toward a causal model approach for adjusting for pre-existing differences in the nonequivalent control group situation. *Evaluation Quarterly,* **1**(3), 399–420.

McCall, W. A. (1923). *How to experiment in education.* New York: Macmillian.

Merton, Robert K. (1973). *The sociology of science* (Norman W. Storer, Ed.). Chicago: University of Chicago Press.

Merton, Robert K., & Geiryn, Thomas I. (1982). Institutionalized altruism: The case of the professions. In R. K. Merton, (Ed.), *Social research and the practicing professsions* (pp. 109–134). Cambridge, MA: Abt Books.

Mosteller, Frederick, & Moynihan, Daniel P. (Eds.). (1972). *On equality of educational opportunity.* New York: Vintage Books.

Moyer, Donald Franklin. (1979). Revolution in science: The 1919 eclipse test of general relativity. In Behram Kursunoglu, Arnold Perlmutter, & Linda F. Scott (Eds.), *On the path of Albert Einstein.* New York: Plenum Press.

Mulkay, M. (1979). *Science and the sociology of knowledge.* London: Allen & Unwin.

Olby, Robert. (1974). *The path to the double-helix.* Seattle: University of Washington Press.

Pechman, Joseph A., & Timpane, P. Michael (Eds.). (1975). *Work incentives and income guarantees: The New Jersey negative income tax experiment.* Washington, DC: Brookings Institution.

Polanyi, M. (1958). *Personal knowledge: Toward a post-critical philosophy.* London: Routledge & Kegan Paul.

Popper, Karl. (1959). *The logic of scientific discovery.* New York: Basic Books. (Originally published in German, 1934)

Quine, Willard Van Orman. (1951, January). Two dogmas of empiricism. *Philosophical Review.* (Reprinted in Willard Van Orman Quine [Ed.], *From a logical point of view.* New York: Harper & Row, 1963.)

Quine, Willard Van Orman (1969). *Ontological relativity.* New York: Columbia University Press.

Reichardt, Charles S. (1983). Assessing cause. University of Denver, Department of Psychology. (Duplicated manuscript)

Rossi, Peter H. (1969). Practice, method, and theory in evaluating social-action programs. In J. L. Sundquist (Ed.). *On fighting proverty* (pp. 217–235). New York: Basic Books.

Rossi, Peter H. (Ed.). (1982). *Standards for evaluation practice.* (New Directions in Program Evaluation, No. 15). San Francisco: Jossey-Bass.

Rossi, Peter H., & Lyall, Katharine C. (1976). *Reforming public welfare: A critique of the negative income tax experiment.* New York: Russell Sage Foundation.

Shapin, Steven. (1982). History of science and its sociological reconstructions. *History of Science,* **20,** 157–211.

Shaver, Phillip, & Staines, Graham. (1971). Problems facing Campbell's "experimenting society." *Urban Affairs Quarterly,* **7**(2), 173–186.

Smith, M. S., & Bissell, J. S. (1970). Report analysis: The impact of Head Start. *Harvard Educational Review,* **40,** 51–104.

Stufflebeam, D. L., Chair, Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials.* New York: McGraw-Hill.

Suppes, Patrick. (1973). Facts and fantasies of education. In M. C. Wittrock (Ed.), *Changing education: Alternatives from educational research.* Englewood Cliffs, NJ: Prentice-Hall.

Toulmin, Stephen E. (1961). *Foresight and understanding: An inquiry into the aims of science.* Bloomington: Indiana University Press.

Toulmin, Stephen E. (1972). *Human understanding.* Princeton, NJ: Princeton University Press.

Trochim, W. (1982). Methodologically based discrepancies in compensatory education evaluations. *Evaluation Review,* **6**(4).

Watson, James. *The double helix.* New York: Signet.

Watts, H. W., & Rees, A. (Eds.). (1977). *The New Jersey income-maintenance* (Vols. 2–3). New York: Academic Press.

Zuckerman, Harriet. (1977). Deviant behavior and social control in science. In Edward Sagarin (Ed.), *Deviance and social change* (pp. 87–138). Beverly Hills, CA: Sage Publications.

2

Evaluation Ideologies

Michael Scriven

New disciplines are often wracked by ideological disputes. In this respect, evaluation is no different from some of the other new entries in the disciplinary sweepstakes — in recent decades these include sociobiology, computer science, feminist theory, non-formal logic, serious parapsychology, ethnic and policy studies, ecobiology, molecular biology, structural linguistics, computerized mathematics, physiological and cognitive psychology, psychohistory, and others. There is nothing new about this, as some reflection on the history of evolutionary theory and astronomy will remind us. But it is hard to achieve perspective on any revolution of which we are part. The proliferation of evaluation models is a sign of the ferment of the field and the seriousness of the methodological problems which evaluation encounters. In this sense, it is a hopeful sign. But it makes a balanced overview very hard to achieve; one might as well try to describe the "typical animal" or the "ideal animal" in a zoo.

Evaluation is a peculiarly self-referent subject. In this respect, it is like the sociology of science; that is, the sociology of science includes the sociology of the sociology of science and, hence, is self-referent. Similarly, systematic objective evaluation — the kind with which the discipline is concerned — is not restricted to the evaluation of microscopes. If it were, it would not include itself. But evaluation applies to the process and products of all serious human endeavor and hence to evaluation. The application of evaluation to itself is sometimes called

From Michael Scriven, "Evaluation Ideologies," pp. 229–260 in *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, edited by G. F. Madaus et al.

meta evaluation, and it has generated the standards for educational program evaluation that are summarized and discussed elsewhere in this book.

Just as it is especially disappointing that the sociology of science — a subject older than this century and dedicated to a self-referent activity — was almost blind to the sexist bias in science, no doubt because that bias pervaded sociology of science as well as other branches of science, so it is depressing to notice the extent to which certain prejudices continue to shape the practice of evaluation. I have no doubt that many more apply than I shall mention here — Ernest House has warned us about some others in *Evaluating With Validity* (Sage, 1980) — but the ones discussed here may constitute a useful start for creating the kind of anxiety and self-scrutiny that will uncover the rest. Later in the paper, I critique standard evaluation processes in the light of these biases, and I also talk about methods and models which avoid them.

These ideologies or fundamental biases that have pervaded much of evaluation include:

1. The Separatist Ideology. "I am an evaluator, you are a subject, she is an object" — i.e., the denial or rejection of self-reference, less kindly described as a kind of criticism. This is most clearly seen in the failure of evaluators to turn their attention to the procedures by which they are themselves evaluated as — and which they use to evaluate others — members of the scientific community. The most scandalous of these procedures include peer review — for research funding or personnel decisions — by uncalibrated, unvalidated, and un-followed-up review panels. It was easy to get away with this as long as evaluation was treated as meaning first of all the evaluation of students (when the word *evaluation* occurs in the title of a book published before 1960, it almost invariably refers to the practices of student performance assessment), and then program evaluation. Program evaluation is not self-referent, since evaluating a program does not itself constitute a programmatic activity. This may have been one of the reasons for the almost phobic intensity of the focus on program evaluation, though undoubtedly another reason was that the funding lay in that direction. In any case, we see here an unhealthy example of parasitism; the constricted notion of what evaluation was all about fed on the improper practices in everyday scholarly operations, from the allocation of funds to the selection of personnel. I postulate as the psychological dynamics behind this kind of error, which would be hard to explain unless there was a deep motivation for it, the existence of something which I will call *valuephobia*, a pervasive fear of being evaluated, which I take to be a part of the general human condition — with rare exceptions — and to apply to scientists very generally, evaluators amongst them. We have frequently seen examples of "going native," the phenomenon of field evaluators posted at program sites who are unable to withstand the social tensions of that role and succumb to the pressure of need-affiliation, joining the staff in point of view and commitment. Often one

finds that within a year, staff evaluators begin to develop significant blindnesses to obvious weaknesses in the program which they are supposed to be evaluating — weaknesses that they would never have overlooked when they first came in. Going native may be an empathic response to valuephobia of the staff under one's evaluative eye, or it may be motivated by the anti-evaluative backlash from that staff.

Thus, the phenomenon of the unscientific scientist, psychologically comprehensible in terms of epidemic valuephobia, represents a simple distortion of scientific inquiry — separatism — which misrepresents it as requiring a permanent role separation between the observer and the observed. In fact, though objectivity is hardest to achieve in self-reference, it is an ideal towards which we must strive, and which we do commonly recognize as part of the obligation of professionalism. Moreover, though claims to achieve it should be viewed with suspicion, there are many ways to approach it. So the first ideology that affects evaluation, driven by valuephobia, is the ideology of the separation of subject and object in an inappropriate way.

2. The Positivist Ideology. The various phases in the development of evaluation proceeded against a most important backdrop of a great ideological battle in the philosophy of science, indeed in philosophy as a whole. This was the battle between the positivists and their opponents, originally the idealists and later many others. Right though the positivists were to attempt a drastic reduction in the cant and circumstance of much then-current philosophy, they over-corrected heavily, and we are still a long way from recovering our equilibrium along with a sense of the possibility of objectivity in ethics and other domains of value inquiry such as evaluation.

Since it is obvious from a cursory review of the contents of scientific works that they are frequently highly evaluative and that the evaluations in them are frequently and carefully rendered highly objective by analysis and documentation (I particularly have in mind evaluations of experimental designs, scientific instruments, the contributions of other scientists, and alternative explanations of the data), it is somewhat bizarre that science of the twentieth century represented itself as value-free. Again, one must consider the possibility that this was an ideology generated to reduce valuephobic anxieties. Surely it is necessary to reach for psychological explanations of such glaring discrepancies as that between the assertion that no evaluative judgments can be made with scientific objectivity and the ease with which evaluative judgments about the performance of students were produced by the very instructors who had just banned them from the domain of objectivity. Thus, both in their pedagogical practice and their professional publications, scientists acted as evaluators who were prepared to back up their evaluations as objective and appropriate, yet who denied the possibility of any such process within the field of their expertise. Since the field of expertise of an educational psychologist includes the practice of grading educational efforts, those academics were guilty of the most direct inconsistency.

Thus, while the separatist ideology or bias rejects the self-referent nature of science or evaluation, the positivist ideology rejects the evaluative nature of science. Both involve inconsistencies between professed philosophy and professional practice, and both have constricted the growth of evaluation severely, since it violates both taboos. One has only to observe the vehemence with which many scientists attack the idea of student evaluation of their teaching on *a priori* grounds without the faintest consideration of whether there is scientific evidence for its validity (see the January 1983 correspondent columns of the *Chronicle of Higher Education*) to see the separatist ideology at work; and the rejection slips which accompanied submissions of articles about evaluation to social science journals prior to the mid-1960s amply demonstrate the power of the positivist ideology, the value-free component of which was often and misleadingly called "empiricism." The wolfdog of evaluation is acceptable as a method of controlling the peasants, but it must not be allowed into the castle — that is the message which each of these ideologies represents, in its own way.

3. The Managerial Ideology. When program evaluation began to emerge, who commissioned it? Program instigators and managers, legislators and program directors. And whose programs were being evaluated? Programs initiated by the same legislators and managers. It is hardly surprising that a bias emerged from this situation. In the baldest economic terms, the situation could often be represented in the following way: someone looking for work as an evaluator (e.g., bidding on an evaluation contract) knew that they could not in the long run survive from the income from one contract. It followed that it was in their long-term self interest to be doing work that would be attractive to the agency letting the contract. Since that agency was typically also the agency responsible for the program, it also followed the evaluators understood that favorable reports were more likely to be viewed as good news than unfavorable ones. Absent extreme precautions, such as radical separation of the evaluation office from the program offices and direct reporting/promotion, etc. of the evaluators by the chief-of-staff, on a highly professional basis, there was a strong predisposition towards favorable evaluations. It is extremely noticeable that when the General Accounting Office or the Congressional Budget Office or the Audit Agency or the Inspector General's Office — all of which are well-insulated evaluation shops — do evaluations of federal programs, the results are very much more critical than those done by allegedly independent contractors, when the contract is let by the agency itself. Even these "internal-external" evaluation shops — the General Accounting Office for example — are not immune to the bias of ultimate shared self interest, since all are agents of a government that wants to look good; but there is a great difference in degree. When we move further down the spectrum, to the usual situation in a school district where the Title I evaluator may be on the staff of the Title I project manager, the pressures toward a favorable report become extreme. Everyone

knows of cases where the project manager simply removes the critical paragraphs from the evaluator's report and sends it on upstairs as a co-authored evaluation.

The managerial ideology went far beyond a simple conflict-of-interest bias, though that reaches so far that perhaps only the appointment of lifetime evaluators, following the standard legislative model of the appointment of superior court justices, could be taken seriously as a countermeasure that showed the society to be fully aware of the problem. The managerial ideology generated a major conceptual scheme, which pervasively contaminates almost all contemporary program evaluations. This is the achievement or success model for evaluation, translated to the view that program evaluation consists of identifying the goals of programs and determining whether they have been met. Relevant though that is to the concerns of the manager, it is of no interest at all to the consumer. The road to hell is paved with good intentions, and the road to environmental desolation is paved with successful programs of pest eradication. The distinction between intended effects and side effects is of no possible concern to the consumer, who is benefitted or damaged by them alike, and consumer-oriented evaluation is, on the whole, considerably more important than manager-oriented evaluation. Although goals and objectives are considerably overrated as aids to good management, resulting in the absurdities of detailed daily lesson plans which may inhibit good teaching more than they facilitate it, there is at least some argument for them in a planning context. There is no argument for them in the evaluation context, *except* for providing managerial feedback and for providing meta-managers with some index of the success of their subordinates in projecting reasonable goals.

Once again, we can find here the cavalier disregard of one's own behavior so characteristic of the separatist syndrome. The very program manager who thinks that goal-free evaluation is either absurd or obscene or illegal, walks straight into the local automobile dealership and proceeds to evaluate the products there without the slightest inclination to request a statement of objectives from the General Motors design team that labored long and hard to produce them. Nor will any reference to such goals be found in *Consumer Reports,* widely read by scientists who loudly proclaim the impossibility of objective empirical evaluation and by managers who proclaim the impossibility of goal-free evaluation.

Consumer Reports is an irrefutable counter-example to the paradigm of goal-achievement program evaluation. The coterie of program managers and their consultants work up many rationalizations to keep program evaluation separate from product evaluation ("people aren't products," etc.), lest the obvious incongruity between the goal-based paradigm they espouse and the needs-based paradigm they employ in their own affairs should become too apparent. It is a phenomenon of some significance that for 15 years all books about the "new discipline of program evaluation" were entitled *evaluation,* talked about evaluation, and turned out to only deal with program evaluation. Not only did they thereby ignore product evaluation, the one kind of evaluation for which we had

many decades of thoroughly reliable development; but they also ignored personnel evaluation, an extraordinary achievement since no serious program evaluation can be done without looking at the treatment of personnel in the program, i.e., at personnel evaluation. Now the treatment of personnel involves considerations of justice — that is, ethics — as well as some other quite sophisticated methodological issues, and it comes perilously close to home since it involves the evaluation of people — and even program managers are people.

So we find valuephobia once more leading to extraordinary global and logistical maneuvers designed — unconsciously, no doubt — to screen off the ethics and the personnel evaluation as if somehow they could be avoided in the course of program evaluation. If they were brought in, of course, then we would have to face the possibility that managers had to be evaluated, that the goals of programs were just as evaluable as their impacts, and that even ethics itself had to be faced as a legitimate part of serious comprehensive program evaluation. In particular, affirmative action issues could not be treated as merely part of the legal background of program evaluation. They would have to be dealt with as serious issues with respect to which correct answers have to be discovered — or else most programs could not be given a clean bill of health.

The managerial ideology dovetailed very nicely with the positivist ideology, because treating a program as equivalent to its success in achieving its goals was a wonderful way of avoiding having to make any value judgments. It merely passed the value judgment buck along to the program managers, accepting their determination of goals as the presupposition of the investigation. "You tell us what counts as a good outcome, and we (scientists) will tell you whether you got it" was the posture, and it was a very attractive one for the valuephobe. The manager, in turn, could often pass the buck back to a legislature, and they — if they so desired — could always blame the public. Goal-achievement evaluation was thus a smokescreen under which it was possible for adherents of value-free dogma to come out of the woodwork and start working on some rather well-financed evaluation contracts. They were not, they said, violating the taboo on making scientific value judgments; they were just investigating the success of a means to a given end. They were also, thereby, committed to connivance-without-cavil in some pretty unattractive programs, including the efforts of the CIA in Central and South America as well as Southeast Asia. When the radical left of the sixties turned up these activities, it concluded that such behavior showed that science was not in fact value-free. All it showed was that scientists were not value-free, a conclusion which no one had ever denied. Although badly bitten over the politics of these exposés, establishment social scientists rightly regarded them as irrelevant to the fundamental logical propriety of the value-free position. For that position maintained only that scientific evidence could not substantiate evaluative judgments, and it never involved the claim that science could not be used for good or ill, by scientists or others. I have mentioned above, and argued in greater detail else-

where, that the fundamental logical position — that science cannot substantiate value judgments — was completely wrong, and indeed obviously wrong; it is for this error that the social scientists must be condemned, and it was this positivist error that led to the managerial error. For only if one believed oneself incapable of disciplined and scientific investigation of value claims could one so readily adopt, without careful scrutiny, the shoddy value premises of the counterinsurgency programs.

Substantial branches of the federal government are in fact concerned with product evaluation — perhaps the Federal Drug Administration is the most conspicuous example. The very methodology that they employed was one which placed an absolutely minimal emphasis upon the achievement of the goals or objectives of the manufacturers or vendors of the product; there was never much doubt that if something came through the doors of the FDA labeled "post-anesthetic analgesic," it would reduce post-operative pain. The problem was always focussed on the side effects. Now one can hardly evaluate side effects by asking whether they represent the achievement of the intended effects, to which they are by definition irrelevant. So what does one use in evaluating side effects? One uses the needs of the patient — or client, or consumer, or user, or student. Thus, in order to evaluate side effects, which one cannot avoid doing if one is to do responsible program or product evaluation, one must have some kind of needs assessment in hand. But if one has some kind of needs assessment in hand, then one can use it to evaluate all effects, whether intended or not. Indeed, it is exactly the appropriate device for doing so. Consequently, one can completely by-pass the reference to goals. Programs, like products should be evaluated by matching their effects against the needs of those whom they affect. And that is what the doctrine of "goal-free evaluation" recommends.

What happens in the managerial ideology is of course that one presupposes the goals of the program were based upon an infallible and eternally valid needs assessment, so that one can use the goals as a surrogate for needs. Unfortunately, that leaves the side effects out of consideration; and it is of course ludicrous to assume either that managers (or those who employ them) always do needs assessment, or always do valid needs assessments, or that any such needs assessments, even if done and valid, will *still* be valid years later when the time has come to evaluate the program. Needs change, not only because we come to recognize new ones, but because programs come and go, population demographics change, the state of the economy varies, and the extent to which needs have been already met varies. Hence up-to-date needs assessment — or something equivalent to this, such as the functional analysis that is often a surrogate for needs assessment in the case of product evaluation — is an essential part of any serious evaluation.

The managerial ideology has another extremely unfortunate error built into it. Not only does it ignore the consumer's point of view, disregard side effects and the justice of the delivery process, but it also pays little attention to a special concern of the taxpayer. One often hears managers arguing that their programs should only be

evaluated on the basis of whether the program goals were achieved, "because that is all that they undertook to do." The evaluation point of view is not concerned solely with — and frequently not at all concerned with — the narrow legal obligations of managers, but also with their ethical obligations, and — transcending the managers altogether — the true merit or worth or value of the program itself. Now *that* raises such questions as whether the same results could have been achieved for less money via another approach, or even for considerably less money using this approach, despite the fact that the contract was completed within the allowed budget.

It is of great significance that the whole question of serious cost analysis was virtually unknown to academic circles until quite recently and that even now it is not part of the standard training of social scientists within the applied fields. Those of us in evaluation who have pushed hard for cost analysis as an equal partner in the team of evaluation methodologies, recall vividly that the notion of cost effectiveness originated not in the academy, but with the Army Corps of Engineers. And cost analysis is by no means conceptually clear to this day; the standard references contradict each other even on the definition of cost (Scriven, 1983).

The effective use of the money available on the project for which it was allocated is one dimension of cost effectiveness; another dimension involves opportunity costs, that is, the comparison of this particular way of expending the resources with other ways that would have achieved similar or better results. This second dimension in cost analysis raises the awkward spectres of a series of "ghosts at the banquet," the ghosts of all the alternative possibilities that were not realized. Should the evaluator have to evaluate not just the program under evaluation, but all the alternatives to it? The cost of such evaluations would be unrealistically great. But if no evaluation is done of the critical competitors — the most important alternatives — then one can never say that the expenditure on the present project was justified. And that conclusion, that the project represented the best or even a justifiable expenditure, is precisely the type of conclusion that many clients for evaluations request, or need even if they do not request it. In particular, it is part of the evaluation imperative to address that question unless there are specific reasons for avoiding it, since it is the question that directly concerns society as a whole rather than special interests of the funding agency and the managers and staff of the program.

So, it is clear that the managerial bias furthered an ideology that omitted a number of important dimensions of the most important kind of evaluation — the systematic and objective determination of worth or value. It is also clear that there are procedures available to reverse this bias and move towards needs-based rather than goal-based evaluation, to what we might call consumer-oriented rather than manager-oriented evaluation. These methodologies include a full range of techniques of cost analysis, including techniques for the analysis of opportunity costs and non-money costs; the provision of opportunities for those who are evaluated to respond to the drafts of the evaluation before it is given to the client officially; and the procedures of goal-free evaluation.

The latter approach not only represents a counterveiling methodology, but a useful methodological simplification, because the practical task of identifying the true "goals of the program" is often completely beyond reasonable solution. One may dig into the historical transcripts — the General Accounting Office goes back to the discussions in committee hearings prior to the formation of legislation — but one then faces the fact that the working goals of the program change with the experience of program delivery. Should one then use the goals of the senior staff members; of the firing-line staff; of the responsible individuals in the funding agencies; or all of the above at the beginning of the evaluation, or during the evaluation, etc., etc.? The problems of converting these goals, expressed informally or rhetorically, into behavioral objectives; of avoiding or resolving inconsistencies in them; of handling the prioritizing of them; of dealing with clear cases of mistaken empirical assumptions in them; and so on, still remain to be solved. Goals are often best seen as inspirational devices — they make poor foundations for analysis.

It is also important to note that for the evaluators to be aware of the goals of the program is for them to be given a strong perceptual bias in a particular direction, which, in conjunction with whatever positive or negative effect they possess for the program, unleashes the possibility of a distorted perception of the results. It is entirely typical for evaluators to look mainly in the direction of the intended results, because they know that the client is particularly interested in that direction; they know that not doing a thorough job in that direction will count against them for future contracts or employment, and they know that they typically will be completely off the hook as far as the client is concerned if they report only on results in that general area. The possibility of this kind of "lazy evaluation" thus opens up, and it is all too often enough to keep one busy without a serious search for side effects. When the field staff do not know the goals of the program, except in the most obvious and general sense, and are only allowed to talk to the program's clients rather than program staff, then they are much more likely to pick up other effects. For one thing, they are on their mettle with no clues; for another, they begin to identify with the recipients and that is a much more appropriate identification — if one has to be made — than with the program staff, not only methodologically (since it generates a new set of biases that can offset the managerial ones), but also ethically. After all, the program staff existed only to serve the recipients, not the other way around. It is therefore extremely unfortunate if evaluators spend most of their time talking to program staff and relatively less of their time talking to program clientele. Social linkages created by these contacts are another source of bias in addition to the perceptual bias in knowing the goals.

There is no need for program evaluation to be done on a wholly goal-free or wholly goal-based commitment. A mixture of the two — with some staff aware of goals and others, isolated from the first group, not aware of them — often works very well. A mode reversal is also possible, with the staff beginning their work in ignorance of the goals and proceeding as far as the preliminary report in writing;

then being informed about the goals, and proceeding through such further work as may appear necessary at that point. So one can often eat one's cake and have it, if one does it in the right order. Goal-free evaluation roughly corresponds to double-blind design in the medical field, and for those same reasons is to be deemed advantageous where possible. It is not, in general, more expensive, though it will certainly be so in some cases, and it will be less expensive in other cases — especially since the cost of disruption of staff and services (so often not counted into the cost of evaluation) is largely eliminated.

Given that evaluation is an essential part of quality control, one learns something extremely important from the discovery that the very term *evaluation* is such anathema in many quarters — for example, in large parts of the federal government system — that people go to great lengths to use other language such as *assessment* or *policy analysis* to cover precisely an evaluation process. It is clear that valuephobia, given the educational background and professional commitment of most people working in the human services area, is far more powerful than their commitment to quality. While it may well be true that evaluation is often performed extremely badly, that it may be a damaging activity for worthwhile programs and involve a risk of unfair treatment for worthy people, that hardly justifies the extraordinary defensive maneuvering that goes on in order to avoid it or its impact. The interest in quality control that the Japanese have shown with the institution of Quality Circles has been widely remarked, but a much deeper and more serious deficiency underlies the fact that Quality Circles, invented here, were disregarded until Japan took them up. Valuephobia runs deep.

Another example: there is no such thing as professionalism without a commitment to evaluation of whatever it is that one supervises or produces — and to self-evaluation as well. Yet few professional schools have even the most superficial curriculum commitment to evaluation training of any kind, let alone of professionals.

At the very least, one would expect to find some willingness among managers to treat investment in evaluation on a straight investment basis; since it is clear that it makes claims to pay off in much the same way as any kind of management consulting pays off, or indeed in the way in which computerization pays off, managers who were seriously oriented towards quality consideration would certainly run up some experimental evidence as to the extent to which evaluation by certain evaluators, done in certain ways, etc., pays off or does not pay off. While most program evaluation may be too biased and superficial to be worth following up, it is patently obvious that good product evaluation and good personnel evaluation can pay off very many times over. There are also a number of clear cases where large-scale program evaluation has paid off by factors between 10 and 100. (The doctrine that evaluation should more than pay for itself, on the average, is a meta evaluation criterion of merit and has been referred to as the doctrine of cost-free evaluation.) Thus, the managerial bias is carried to the extreme of a very self-serving indulgence in valuephobia.

4. The Relativist Ideology. Whereas the positivists were committed to the view that there was some kind of definite external world about which we learned through our senses and through experiments, more recent philosophy of science has tended to move away from this "realistic" or "external world" commitment towards the view that everyone has his or her own reality, all equally legitimate. And evaluation has been very much influenced by this movement in the philosophy of science. Throughout this book, in articles by the most distinguished workers in evaluation, one finds not only a shying away from the notion of objective determination of worth — as in Cronbach's aversion to summative evaluation — but also a shying away from even the notion of objectively correct descriptions of programs. Multiple perspectives, yes; multiple realities, no. While it is in my view perfectly appropriate to respond to the obvious need for multiple perspectives and multiple levels of description by abandoning any naive assumption about the existence of a single correct description of objects in the external world (including programs), it is equally mistaken to overreact in the direction of solipsism or relativism. The relativist ideology or bias is, in my view, a case of such overreaction; it is often to be recognized by the emphasis placed by its supporters upon the impossibility of establishing "the truth," or "the existence of a correct view of the world," and so on. If it were really the case that there is no objective superiority of some descriptions above others, then there could be no discipline of physics any more than evaluation. The concept of relativism is self-refuting; if everything is relative, then the assertion that everything is relative cannot itself be known to be true. So, although we may reject the existence of a single correct description, we should not abandon the idea that there is an objective reality, though it may be a very rich one that cannot be *exhaustively* described. It may even be one which can only be described in a non-misleading way by giving descriptions which are relativised to each audience; we may concede all this, and yet insist that in many cases there is such a thing as a correct — though not a unique — description (given a certain audience and level) by contrast with a number of incorrect descriptions. Indeed, these descriptions may involve descriptions of the merit, worth, or value of parts or aspects of the entity being investigated.

It has been argued above that the very core of science, as of other disciplines, is committed to the objectivity of evaluation — in fact, if one could not distinguish good from bad scientific explanations, one could not be said to be a scientist at all — and there is thus no shame or indeed any further commitment involved in treating evaluation as an objective discipline. The fact that ethical issues must also be handled raises the question of the status of objectivity in that subfield of evaluation; but whatever decision one comes to there, one cannot weaken the resolve with which one must address the search for the best and the better and the ideal when evaluating all aspects of a program other than the ethical. Programs are simply very complicated institutions, but they are no more complicated than theories or even experimental designs, which we have no hesitation in evaluating

by strictly scientific criteria. It is a modest enough — and surely a scientific — suggestion that we should evaluate programs in terms of their latent rather than their alleged function.

Thus, I see the re-emergence of relativism as the latest and most serious bias in evaluation methodology, because it comes from the evaluators themselves. It is quite easy to show that those who support it officially actually disregard it in their common practice. Just as managers act as goal-free evaluators of consumer goods, so relativists act as objectivists in their grading of their students or of the interpretations by their colleagues of certain experimental results. This inconsistency between practice and philosophy is a sure sign of the immaturity of this field at the present moment. There are many other such signs, and in the ensuing paragraphs we will call attention to a few standard evaluation practices that violate some of the most obvious criteria for systematic evaluation — and yet have not been universally condemned by professional associations of evaluators and often are not even seen as particularly relevant to the narrowly conceived business of program evaluation. In the course of discussing these examples, albeit very briefly, we will also take the opportunity to introduce one or two conceptual distinctions that clarify practices and malpractices as well as referring to the four fallacious ideologies that we have outlined above.

The Social Science Model. This set of four fallacious ideologies often seems to congeal into something that could be called the traditional social science model of evaluation. Since we are here proposing a set of alternative positions or ideologies, which we will elaborate in modest detail below, it can be argued that we are proposing an alternative and more appropriate model for the social sciences. Thus, if this argument is correct, evaluation should lead us to a considerable sophistication of the rather primitive philosophy of science that has been associated with the social sciences, and one might sum this up by saying that evaluation turns out to be a better model for the social sciences than they have proved to be for it. Taking this view seriously, one looks more carefully at the publications in the traditional social science journals and sees many ways in which these could be increased in their value, to science and to society, if a range of further questions were to be addressed about them, both at the design level and the meta level. So there is a second goal for this paper, the commitment to substantial reform of the ideology and hence the practice of the social sciences and not just of evaluation.

The examples that follow come from educational experience, not just because we are all familiar with such cases, but because it may be that the largest payoff from improvement in evaluation can be achieved if reforms in educational evaluation take place — by contrast with reforms in the administration of criminal justice or other human services. The examples chosen scarcely exhaust the area; we could have focussed solely on the kind of evaluation that underlies the current mania about computers, e.g., the absence of serious needs assessment behind the push for teaching BASIC as "computer literacy." But we focus on older sins.

The Evaluation of Student Work. In this most common of all educational experiences, we find example after example of methodological misconceptions and misdirections, which clearly show how well segregated our intellectual efforts were from our pedagogical practices. It is only as the discipline of evaluation has grown to some degree of autonomy and as external social pressures have forced us to re-examine the evaluation of students that we have come to raise our eyebrows over practices which many of our most intelligent and best-trained social scientists had set up and nurtured for decades.

We will not here rehearse the whole sorry story of the abuses of norm-referenced testing and the gradually improving mix with criterion-referenced testing that is emerging. As the fights over minimum competency achievement tests for graduation or promotion, over the definition of test bias, over the concept of instructional validity, and about other issues are reaching a more mature level of discussion, assisted by the courts and public opinion as well as the scholars, we are seeing the development of evaluation by contrast with mere testing. We will here simply comment on a basic logical point that has not been treated with appropriate respect in the literature on measurement, but which becomes crucial as we attempt to develop the logic of evaluation in any consistent and comprehensive way. The basic logical relations in evaluation seem to be four in number: grading, ranking, scoring, and apportioning. The following definitions are partly stipulative, but involve very little straightening out, being mainly a reflection of the implicit logic of the common terms. *Grading* is the allocation of objects to a set of classes that are ordered by merit or worth; the number of classes usually being small compared to the number of entities graded, and the description of each class being given in terms that refer to some external standards of merit or worth, i.e., not simply to relative position. *Ranking* is the allocation of individuals to some position in an ordering, usually one where the number of positions is equal to or almost equal to the number of individuals; the order being by merit or worth. *Scoring* is the most elaborate standard mensurable approach associated with evaluation; it involves the ascription of a quantitative measure of merit or worth to each individual in the group being evaluated. And *apportioning* is the process of allocating a finite valuable resource in varying amounts to each individual as a means of expressing an assessment of merit or worth. Certain obvious connections and lack of connection can be quickly stated. Ranking does not imply grading nor vice versa; scoring will entail a ranking but not a grading (in general); neither grading nor ranking will entail an apportioning, although apportioning can be defined in terms of a very complicated set of gradings and rankings of parts of whatever is being evaluated, whenever such parts can be identified. Both criterion-referenced and norm referenced tests require cutting scores in order to define a grading; normed tests always, and criterion referenced tests sometimes, define a ranking. The body of basic training in tests and measurement is weak on these distinctions, because of the valuephobic exclusion of explicit discussion of merit. As a result, elementary mistakes are to be found in almost every text and in many published tests, where

confusions between these types of evaluation are rampant. A typical example occurs when the translation of the ratings on a five point rating scale is given as excellent; very good; average; below average; very poor. The first two of these refer to grading; the next two refer to a norm-referenced or ranking approach, and the last reverts back to a grading approach. The scale is logically unsound since the average performance of the group being rated may be very good, or poor, or anywhere else, so there are often two correct responses. The "anchors" given presuppose a more or less normal distribution *and* a coincidence of the upper reaches of the distribution with excellent performance (and correspondingly with the lower reaches), both of them are extremely implausible assumptions in most contexts of student evaluation.

The concept of grading on the curve, another symptom of valuephobia, exhibits the same distortion of the difference between grading and ranking. With typical managerial bias, it assumes that the difficulty level of the test has been set at precisely the right point so that the top ten percent (or 15 percent) which are automatically given an *A* will in fact deserve to be regarded as having performed not merely superior work (which is tautologically correct) but excellent work, and similarly for the other grades. If it is argued that psychologists ascribe no more significance to the *A* than top decile performance, then we must focus on the bottom end of the class and inquire why it should be assumed that there must always be ten percent who fail. Obviously, such an assumption is completely false in many circumstances, and, if false at that end of the distribution, the converse must be in question at the other end. And in the middle.

Of course, built into the very conception of scoring that leads to the normal distribution used in grading on the curve is precisely that identification of merit with a point in the scoring system, the commitment to an independent assessment of worth or value, that is supposedly rejected by going to grading on the curve. If one is prepared to commit oneself to the view that any point, however earned on whatever question in the test, is of equal value — the assumption without which one cannot justify scoring at all as a basis even for ranking — then one might as well commit oneself to the rather more modest assumption that one can identify a truly excellent or hopeless performance not just by its salience.

Another example of logical confusion occurs in funding decisions, where the review panel is instructed to rank or grade programs, whereas apportioning is the question at issue. (Using the wrong instructions may, however, make managerial manipulation of the results much easier.)

Teacher Evaluation. The evaluation of research has always been thought to be relatively straightforward by comparison with the evaluation of teaching; close examination of the implicit assumptions in the way research evaluation has been done has led to increasing disquiet with this in recent years, and a great deal more needs to be done towards developing reasonably objective standards for the

evaluation. But the evaluation of teaching and teachers is much more of a scandal. A great deal has been written about this recently, and we will simply make two points here. First, it has rarely been remarked that there is a complete difference between an evaluation of merit and an evaluation of worth in teaching, and that these two considerations have quite different relevance to different kinds of personnel decisions. The evaluation of worth (to the institution) is an evaluation which brings in questions of the salaries in the marketplace, of the extent to which the subject matter is popular or essential to mission, of payoffs from fame (in the media sense) of the instructor and so on. None of these is involved in the evaluation of professional merit, a property of the individual and his or her performance against the standards of merit in that profession. Thus, a teacher at the college level may have the greatest merit and be of so little worth to the institution that it does not make sense to grant tenure, simply because the subject matter in which this instructor specializes no longer draws any students at all; the reverse may also be true of the great showman or grantsman who attracts income and/or students but does so without a foundation of true professional merit. Roughly speaking, initial and tenure appointments should be made on the basis of worth as well as merit, but promotions and awards should be made solely on the basis of merit.

A second interesting point that can be made about the evaluation of teachers concerns the fact that the universal procedure employed in the evaluation of primary and secondary school teachers is invalid for every possible reason. That procedure consists of visiting a very few classes, often with advance notice and using checklists or subjective judgment to determine whether appropriate practices are occurring during the visit. The sample size is too small to be of any use, even if the sample is random; the sample is not random, since the measurement process may affect the treatment; the judge is not free of significant social biases from non-classroom relationships with the teacher; the checklists are invalid; and finally the judge is completely invalidated as a detector of learning gains, which must be regarded as at least a major part of what teaching is all about. The continuance of this practice in the light of these obvious invalidities is a reflection upon the state-of-the-art of (or interest in) evaluation amongst professional administrators and teachers. It should, of course, be noted that neither unions nor management would benefit from switching to an alternative approach since neither is rewarded for the replacement of bad teachers by good ones, and indeed would be heavily punished by the emotions, costs, and struggles that would be involved in a changeover. Only the children and the taxpayer are cheated and their representatives are not yet sufficiently sophisticated to speak up about the impropriety of this process.

Apart from this generally dismal situation, there is an extremely interesting and more sophisticated point involved. Supposing that we *had* established a very reliable list of indicators of good teaching, and that we *were* able to observe teachers at work without affecting the way they teach, in a *large enough* sample

of lessons. It seems that then our problems would all be solved. (In fact, we do have one such indicator, not the dozens which are widely touted; it appears that sample of lessons. It seems that then our problems would all be solved. (In fact, we do have one such indicator, not the dozens which are widely touted; it appears that "interactive time on task" is a good indicator of amount of learning.) We now come to see one of the more radical differences between formative and summative evaluation. For purposes of summative evaluation — that is, in this context, the making of personnel decisions — we cannot use statistical indicators of merit that refer to only one or some aspects of the performance. This claim of course directly contradicts the standard operating procedure in the evaluation of teachers. We cannot use such an indicator any more than we can use skin color, even when we are in possession of job-related, valid generalizations about skin color, e.g., that the crime rate is higher among blacks, and the oppression rate higher among whites. We cannot use such generalizations in the evaluation of individual cases, because, in the first place, they apply only to randomly chosen samples from the population to which they refer, and the individual in a personnel evaluation situation is by no means a random sample — we know much too much about such individuals for them to be "representative" or "typical" or "random" samples of that population. In the second place, if we do *not* know more than this about the individual in a personnel decision case, then we can and should go out and get some more evidence, evidence directly related to track record performance in this or the most similar work situation we can identify in their case history. This is scientific common sense. The ethical imperative, in addition, requires that we not use membership in a very general class as the basis for judgment about the individual; we have various terms for the associated error, for example, "guilt by association," or "stereotyping." Since there are always feasible and superior alternatives to these generalizations in personnel work, there is no justification for using them. In the case of summative teacher evaluation, the clearly superior alternative is the use of direct evidence of learning, plus appropriate standards obtained from suitable comparisons with other teachers of similar children. (Even *holistic* ratings by judges present *most of the time*, who lack the chance to acquire *social bias*, will be superior; which is to say, student evaluations of teaching.)

The various absurdly primitive attempts to use pupil performance as an indicator of teacher merit have produced an understandable backlash against this kind of approach; but when the comparisons are made with other teachers of children in the same school, where allocation to classroom is almost entirely random — or to children in similar schools serving essentially similar populations — then the difference in final achievement on a sound common test must be due to the differences in teaching ability. Minor differences are of no interest since the matching is not perfect and circumstantial variables will have some minor effects (e.g., classroom architecture, the presence of a single highly disruptive student, etc.) However, if multiple measurements of student gains are made (e.g., in an elementary school, three successive measurements across three successive terms) there is not going to be much doubt that teachers who are always two standard

deviations off the mean are either genuine super-teachers or genuine failures. The courts having upheld this kind of evidence as grounds for dismissal; we should now be using it. (Where it is not available, student evaluations are the best alternative.)

Of course, even though the courts have upheld the use of comparative gain score evidence alone, it is not all that we should be gathering. We also need evidence of the quality of the content taught and not covered in the test. This is readily obtainable by inspection of materials (especially student products) by a curriculum specialist or even by a principal with experience in this area. We also need evidence about the ethicality and professionality of the teaching process. (Where student ratings are used instead of gain scores, the evaluation of all content becomes crucial.) The ethicality of the teaching process is not a matter of whether one uses negative reinforcement rather than positive reinforcement — often inappropriately regarded as cruel and unusual punishment by supervisors and principals. It is rather a matter of whether there is flagrant disregard of due process and considerations of justice, e.g., by the use of sexist or racist remarks or practices; by unfair grading practices; and by inappropriate test construction. This will best be picked up by a review of the test materials and anonymous student responses. Finally, although it is not absolutely essential, it is highly desirable to use evidence of professionality, usually best based upon a dossier submitted by the instructor. Professionalism requires self development, so evidence of advanced courses in both subject matter and method would be relevant. It requires self-evaluation; so it requires evidence that testing of one's teaching success, including (usually) the use of student evaluations, has been obtained. Both of these considerations require a steady process of experimentation, with new materials and approaches. Even a program of critical reading of new and promising literature or current research literature would be relevant to these considerations and could be documented in such a dossier.

The preceding will generate a highly satisfactory model for summative evaluation. But does it not, in one version, involve a violation of the very principles which it was set up to support? In using student evaluations, especially as our only indicator of learning, are we not using an indicator that has only a statistical correlation with merit in teaching? This is true, but this is one of the cases where a statistical indicator may be justifiable. To see why, consider an even more extreme example. Test scores by students on well-constructed scholastic achievement tests are used in order to select the entering class for colleges and graduate schools. But it is well known that such tests are not infallible indicators of what we may take to be the criterion variable — success at those colleges. If they are "merely statistical indicators," then surely we are not entitled to use them since they violate the principle of judging the individual on the basis of his or her own work rather than on the performance of people who are related by some statistical generalization to the individual being evaluated? The reader will no doubt notice two crucial differences about this case. In the first place, we *are* using the individual's own

work, a comprehensive and relevant work sample, in fact. In the second place, we do not have a feasible and better alternative available, (cf., also the validity of an end-of-term course exam).

People sometimes propose that the use of the high school teachers' evaluations of the college-bound student — based, as they are, upon very extensive observation — would be superior to the use of test scores. Investigation shows that this is not usually the case, essentially because of the problem of inter-judge unreliability. In short, it is not a *systematic* alternative because there is no feasible *system* of having the same set of judges look at all candidates, so the test — which is administered in the same form to all candidates — wins on the swings of reliability what it loses on the roundabouts of inadequate work sampling. And so it is with student evaluations of teachers. Especially when the questionnaire is appropriately constructed and administered, a high score has a good positive correlation with the learning outcome. Of course we could always directly measure the learning outcomes — that is not the problem; the problem is identifying the extent to which the gains are due to teaching merit (as opposed to the textbooks, peer interaction, and intellectual or familial background), and deciding on the cutting scores that will separate good teaching performance from bad. Absent the comparative situation described earlier, our only alternative is the use of student evaluations. Now these evaluations are holistic evaluations of the particular work of the particular individual, not evaluations of part or one aspect of what the teacher does (cf., brief visits or time-on-task measures); they are probably related to learning, *and* they include allowances for other causes and for what could have been done, by contrast with what *was* done. The method is imperfect of course, but based on considerable exposure to other teachers, in the consumer's role. In short, they provide us with the comparative dimensions that we lack if we just collect gain scores. (It does not follow, by the way, that we should use a comparative *question:* "Rate this teacher against others you have had. . . ." That will get you a ranking — but few personnel decisions can be based on a ranking, certainly not a promotion or tenure decision. That's grading on the curve. You must ask for a grading: "Rate this teacher A–F, where A means excellent . . . F means extremely bad." The student's experience with other teachers will create the range of the *feasible;* the top of that range is the locus of excellence.)

While time-on-task measurements *are* empirically related to the performance of the individual, as is skin color, the relationship is of a weaker kind, one that does not survive an increased specification of the individual's characteristics. Student evaluations are holistic of both individual and performance and, though by no means perfect, are — as far as we know and as we'd expect — superior to ratings by any other general category of judges (e.g., principals or supervisors or process experts) though we certainly need more sampling of the matrix of subject matter by age, by school environment, etc., to support this claim more substantially. Hence, we should be using them in the high school and college situation, where there are

usually no comparative norms available. When comparisons *are* possible, as with multi-section freshman courses in college, it is preferable though sometimes politically impossible to set up random allocation, common tests and blind grading, and revert to the use of comparative norms.

The preceding discussion will make clear the way in which ethical considerations interact with scientific ones in personnel evaluation. It should also make clear the important distinction between holistic and what can be called analytic evaluation — one might use the terms macro evaluation and micro evaluation instead. The holistic evaluation is an evaluation of the total relevant performance, whereas the analytic evaluation evaluates some component or dimension of that performance. The evaluation of components is in some ways more useful for formative evaluation than the evaluation of dimensions, because it is likely to be easier to manipulate components than dimensions. But either may provide an adequate basis for assembling *or justifying* an overall evaluation. Counterintuitively, however, it transpires that we have clear evidence showing holistic evaluation is sometimes considerably more valid — as well as far more economical — than syntheses of micro evaluations. The problem with the analytic approach to overall evaluations is that the assembly of component scores or grades involves a weighting and combining arrangement of typically unknown validity. (See *The Evaluation of Composition Instruction*, Davis, Scriven, & Thomas, 1981). The evaluation of teaching also illustrates clearly the differences among evaluation, explanation, and remediation, so often confused in program evaluation, where the client frequently *demands* that the evaluator submit remedial recommendations as well as an overall evaluation. Attractive though that is to the client, and important though it is to do it when possible, there is often an urgent necessity to choose between sound summative evaluation and relatively unreliable and more expensive formative evaluation. It is fairly easy to evaluate teachers on the basis of their success, where one can get appropriate comparison groups set up; but it is not a consequence of the validity of this evaluation that one can give any advice whatsoever to the teachers who perform less well as to how to improve their performance. The reason for this is not only that the best approach to summative evaluation is often holistic; it is also that we lack the grounded theory to provide the appropriate explanations, since all efforts to find components of a winning style (apart from interactive time-on-task which is only marginally describable as "style") have so far failed. Absent a diagnosis of the causes of failure, whence comes a prescription?

Although the traditional approach to remediation is through explanation, the occasional success of "folk-medicine" demonstrates the *possibility* of finding remedies whose success is not inferred from a general explanatory theory, but discovered directly. And so it is with teaching; we might find that a certain kind of in-service training package is highly successful, although it does not proceed from an analysis of the causes of failure. It is thus triply wrong for a client to demand

micro explanations as part of an evaluation as a route to remediation or justification. They will not necessarily lead to remediation; there are other ways to get to remediation to provide evidence for the validity of the evaluation. The latter is provided on a holistic basis, e.g., by correlational data relating evaluations by this method (or these judges) with the subsequent performance of the criterion variable. Of course, remedial suggestions are often obvious or easily uncovered from an analytic summative evaluation; but not always and the analytic approach is often not the best one.

I have already mentioned that if one approaches the evaluation of something by evaluating components or dimensions of it, which are then assembled into an overall evaluation, serious problems of validation arise about the formula used for assembly. I have discussed elsewhere the use of some traditional approaches, e.g., weighted-sum with overrides, and we have of course the well known model of cost-benefit analysis, in which we reduce costs and benefits to a single dimension and thereby convert evaluation into measurement. Much more needs to be done about the synthesis step in program evaluation; the present trends, partly because of the difficulty of this step and partly because of the influence of the relativist ideology, is towards mere "exhibiting" of performance on the multiple dimensions involved. This is simply passing the buck to the non-professional, and represents far less than the appropriate response by a professional evaluator.

Review

What is emerging from our discussion of these common evaluation practices? Two points. On the one hand, we are seeing gross errors of practice emerge under critical study, and it is not hard to see how these reflect — directly or indirectly — the ideologies or biases we have discussed. By far, the greatest influence of those ideologies is indirect in that they have discouraged recognition of the essential self-reference and evaluative nature of science; discouraged emphasis on the client's perspective; and discouraged any sustained commitment to the existence of correct versus incorrect conclusions.

The Consumerist Ideology. For many people, committed to the relativist ideology, it follows from the fact that one is attacking some ideologies that one must be supporting another. This is in error as a general conclusion, but it would be fair to say that the sum total of all the criticisms so far does add up to a point-of-view that needs to be made explicit at this point. I'll use a label for it that has been contaminated with largely irrelevant opprobium, but still retains enough common meaning and a connotation of an ethically appropriate position; I'll say that we have been presenting a *consumerist ideology*. Consumerism is like unionism; both came into existence to represent a movement which, even from the

beginning, involved some wrong activities, while representing a long overdue balancing of power and involving an essentially moral concern with people who had been left out of the reckoning. By and large, consumerism has done well by us, from the first day that Ralph Nader provided an over-simplified and in many ways unjustified analysis of the General Motors Corvair automobile, although it has brought with it some overkill pseudo-safety and pseudo-consumer protection legislation. The essential point of the consumerist ideology in evaluation is that all parties affected by something that is being evaluated should be taken into account and given at least their appropriate moral weighting — and in many cases, an appropriate opportunity for explicit participation and/or response to the evaluation process or outcome.

We can proceed quite briefly with a few more examples of bad practice still tolerated because of acceptance of the fallacious ideologies, and then conclude with a brief description of a model of evaluation methodology that can be said to unpack the consumerist ideology, just as the goal-based evaluation model unpacks the managerial ideology.

The Evaluation of Educational Institutions by Accreditation. Just as there is a completely standard model for primary or secondary teacher evaluation, so there is one for the evaluation of primary, secondary and professional schools. This model, accreditation, has a number of distinctive features, some virtues, and a number of serious weaknesses that cannot be dismissed as due to constraints on resources available for accreditation.

The distinctive features of accreditation, nearly all present in all applications of this approach, are:

1. The use of a handbook of standards, involved in several other components, beginning with
2. A self study by the institution, resulting in a report on how well they are achieving what they see as their mission; which is read by
3. A team of external assessors, usually volunteer members of the same general professional enterprise, who not only read the self-study, but also make
4. A site visit, usually for one to three days, which involves direct inspection of facilities, interviews with staff, clients, and students, plus review of prior reports, and which results in
5. A report on the institution, which usually makes various recommendations for change and for/or against accreditation (possibly with various conditions); this report is subject to
6. A review by some august panel, at which the right to appeal against the

recommendations is sometimes granted to the institution being evaluated and at which some censoring of the recommendations sometimes occurs; after which

7. A final report and decision is issued.

Some of the desirable features here include: some use of external evaluators, self-scrutiny as a method of preparing the ground for the external suggestions and for providing a linkage group with the external assessors, a review process which gives some chance to address injustices, and a rather modest cost. Within this general framework, good evaluation could indeed be done. But it is rare to see it done.

We'll pick up only a few of the problems, more to illustrate than to provide a thorough analysis. We can conveniently group the problems under the same heading as the components.

1. The handbook of standards is usually a mishmash ranging from the trivial to the really important, and there is usually no weighting suggested. (Sometimes there isn't even a handbook of standards.) Consequently, the bits and pieces can be assembled in more or less any way that the panel feels like assembling them, without any focus on the justification of the implicit weighting of such a synthesis. It is common for the handbook of standards to begin with some piece of rhetoric about how institutions should only be judged against their own goals, but yet we will find buried in the handbook a number of categorical standards that must be met by all institutions. This inconsistency reflects a failure to resolve the ideological tension between managerial and consumerist approaches. Managers do not want to be blamed for not doing what they did not undertake to do; on the other hand, consumers do not like to be treated badly and don't much care whether the maltreatment was unintentional or not. Ethics obviously requires that the rights of consumers be protected at least in certain respects, so that minimum standards of justice should be met by all educational institutions. It might also be argued that public institutions have some obligations to provide a service that is reasonably well-tailored to public needs, and that even private institutions — who may select more or less whomever they wish to enroll — must nevertheless provide services that are related to the needs of those whom they do enroll. (Note that the absolute standards one does encounter in these typical standards checklists are usually considerably less ethics-related than the ones just mentioned, indeed are often highly debatable; e.g., the requirement of vast libraries for graduate programs.)

2. The self-study is frequently devoted towards a review of goals in the light of mission, and of achievements in the light of goals. This tends to involve the usual managerial biases, because of the failure to give due weight to the consumer; in particular, there is poor attention to the need to search for side effects, there is little concern with comparisons or cost-effectiveness, and usually little concern with the ethics of the process. (This of course varies considerably across the huge range of

accredited institutions, but of the many that I have seen from the medical and legal area as well as from many college and high school reports, the above seems to be a fair generalization.) Another type of weakness emerges at this point; there is rarely a professional evaluator on the internal self-study review team, and consequently many of the usual traps are fallen into, including careless ascriptions of casual efficacy to programs, misinterpretations of data about learning gains, and alleged success of graduates and so on. It is impossible to expect that there will not be some adjustment of goals to achievement — and this may sometimes be healthy — but it does provide an opportunity to duck behind goal-relativism, which is allegedly the standard by which the accrediting association will make the final judgment. Thus the managerial bias is supported by the relativist one.

3. The team of external assessors is usually picked from volunteers, and, consequently, professional evaluators and the busiest administrative analysts and consultants are more or less automatically excluded. Professional evaluators are by no means automatically an advantage on these panels; it would be absurd for a professional evaluator to assume that they are. The only imperative is that they should sometimes be present and that careful meta-studies should be done to see if this does lead to any improvement. The idea that one can dismiss the supposed experts entirely seems naive, given the low quality of the usual reports. It must be expected that professional evaluators will have to be paid for this activity, so the price goes up; that price could be offset by reducing the size of the panel, since the indirect costs per diem and travel are quite substantial. We should find out whether some professionalism would offset some loss of numbers. There could also be systematic studies with funds from foundations, to see whether the addition of the best management consultants and evaluators will yield cost-saving suggestions that would compensate for increasing the fees to cover their costs. There would then be problems about equity as far as the still-unpaid members of the panel are concerned and serious problems about total cost. However, the quality of the evaluation reports, judged against professional evaluation standards, is so spotty that the entire process should be subject to serious scrutiny; it hardly constitutes an acceptable way in which to evaluate most of our important educational institutions.

Professionals and other busy people are not the only ones left off by the process of volunteering and subsequent selection, usually by central staff personnel. There is a strong tendency to leave radicals and other "extremists" off the panel. No doubt there are accreditation units here and there — I know of one — where this is not true; but it is certainly the general pattern, and it is a typical sign of managerial bias. If we were searching for truth, we would realize that radical perspectives often uncover the truth and can demonstrate it to the satisfaction of all panelists. And we would realize that establishment-selected judges are likely to be blind to some of the more deep-seated biases of the institution; one can see how serious this is by tracking back through old accreditation reports given during pre-feminist days. Not a sign can be seen of sensitivity to radical sexist exploitation

and inappropriate passing over of women for positions which they should have received; but there were plenty of feminists around in those days, if anyone had been looking for them.

This managerial bias is of course one that will favor the institution by not uncovering the skeletons in its closet; and it is not accidental that the whole accreditation process is run by a system of fees levied on the very institutions that are accredited and which provide the personnel for the accreditation. The system is thus in a fairly straightforward way incestuous; the question is whether one can conclude that it is corrupt. To the extent it is not, we must thank the innate professional competence and commitment and integrity of the panelists, which does not entirely evaporate under the background pressure towards pro-management, pro-establishment reports. However, to jump a few steps, it is important to notice that the report by the site team will sometimes be radically censored by the review board, which has of course not been to the site, in the direction of excising many or all of its most serious criticisms or conditions. This is an unattractive situation, and one which is not widely recognized. It suggests inappropriate bias, and when we look at the procedure whereby the review boards themselves are selected, we find in many cases an even more unattractive situation. For the review panels — for example, the governing board of the regional accreditation associations in the case of schools and colleges — are often entirely self-selected and often consist almost entirely of active or retired administrators.

4. The site visit is also not designed to capture the input of the most severe critics. Such obvious devices as setting up a suggestion box on the campus during the site visit, providing an answering machine to record comments by those who wish to call them in anonymously, or careful selection of the most severe critics of the institution from among those who are interviewed are practices that one rarely if ever encounters. Failure to adopt these practices simply shows a failure to distinguish between the need for a balanced overall final view and the need for input from the whole spectrum of consumers; both are imperative, the former does not exclude the latter, and the two are quite distinct.

So, from the use of inappropriate standards, such as the requirement of large research libraries for graduate programs instead of *access* to such libraries *or* to online databases, to the failure to enforce serious standards for the self-study (to the point where the great post-secondary institutions go through this stage without most of their faculty ever hearing that it is going on), we are dealing with grossly unprofessional evaluation. Nervousness about the incestuousness of the process is not lessened when one sees the defensive nature of the accreditation agencies' reactions to the proposal that federal or state governments should have some input to accreditation. Undesirable though this may be in various ways, a hybrid system would at least provide minimal insurance against the more outrageous examples of

"National Tobacco Research Institute" whitewashes. The extremely lax enforcement of professional standards by the medical and legal professions is a well known scandal and, although there are some professions — the psychologists are a pretty good example — which rise above this kind of managerial/separatist bias, it must be realized that the society and its legitimate government have extremely strong rights to be represented in a process which deals with the key services provided to its relatively unprotected citizenry. When we do get an occasional glimpse at the actual standards of competence in a profession — as when we see the results of competency exams on teachers, or the analysis of drug prescriptions written in a certain region — we have every right to suspect that the self-regulation process is not being done any better than one would expect, given the biases built into it. Accreditation is an excellent example of what one might with only slight cynicism call a pseudo-evaluative process, set up to give the appearance of self-regulation without having to suffer the inconvenience.

If one had to sum the whole matter up, one might call attention to the fact that in virtually no system of accreditation is there a truly serious focus on judging the institution by the performance of its graduates, which one might well argue is the only true standard. Not to look at the performance seriously, not even to do phone interviews of a random sample of graduates, not even to talk to a few employers and/or employment agencies who deal with graduates from this and others institutions; *this* is absurd.

It is scarcely surprising that in large areas of accreditation, the track record of enforcement is a farce. Among all state accreditation boards reviewing teacher preparation programs, for example, it is essentially unknown for any credential to be removed. Nor is it surprising that at one point the state of California was threatening to close down all unaccredited law schools, although some of these had a much higher success rate in getting their graduates past the bar exam than many prominent law schools in the state. And passing the bar exam is presumably one of the most important things that a law school is supposed to do for you — as far as I know, it is the only one for which we obtain a measurement. Crude measurements are not as good as refined measurements, but they beat the hell out of the judgements of those with vested interests.

Another example of crude measurement that turns out to be quite revealing is one that can be applied to the evaluation of proposals and the allocation of funds for research in the sciences, as well to the accreditation process, and it is such an obvious suggestion that the failure to implement it must be taken as a serious sign of the operation of the separatist ideology in the service of elitism. This modest proposal concerns checking the reliability of team ratings. When a review panel of peers judges that a particular proposal should be funded and another rejected, just as when a review panel judges that a particular institution should be accredited and another disaccredited (or warned, or not accredited), it seems reasonable for those affected to raise the question whether another panel drawn from the same pool of

professionals would have made the same recommendation. This is of course the question of inter-judge (in this case inter-panel of judges) reliability, and until very recently no such test had ever been made (although it is the simplest and most obvious recommendation that a freshman student of one of the social sciences would make about a judgmental process of any kind that was officially regarded as subject to scientific investigation). Only separatism insulates the scientist (or other professional) from this scrutiny; and in the couple of cases where a study of inter-panel consistency has been performed, the results have not been encouraging. The North Central Association sent in two teams to have a look at the school — Colorado Springs High School — and the results demontrated not so much a lack of agreement but some important disagreements coupled with the possibilty that most of the agreements were due to shared bias. A small National Science Foundation study of the results when more than one panel, drawn from the same pool of professionals, was assigned the task of rating proposals, showed striking and substantial differences. When these relatively crude measures are the only measures we have, the only appropriate conclusion from these results must be an extremely skeptical view of the validity of the accreditation approach to program evaluation.

Ideologies and Models

Ideologies are intermediate between philosophies and models, just as models are intermediate between ideologies and methodologies. Thus more than one ideology may support a particular model; just as the relativist ideology supports Elliot Eisner's connoisseurship model, so the empiricist ideology as well as the managerial and relativist ones support goal-based evaluation models. Some subtler relations can be plausibly inferred. Recently, for example, we have seen Cronbach's group coming out strongly in favor of formative evaluation as the only legitimate kind of evaluation, by contrast with summative. In this respect, their position matches that of some staff members of the American Federation of Teachers, who are willing to support the idea of evaluation of teachers for improvement, but not the idea of quality review. Apart from logical problems with the artificial nature of this separation, it is certainly an emphasis attractive to both the positivist and the relativist ideology, because each is much more willing to tolerate the idea of improvement — with its connotations of goals and local values as the criteria — than categorical assertions about merit and worth. Few people are valuephobic about the suggestion they are less than perfect, need some improvement; but to be told they are incompetent or even far worse than others, is less palatable.

In remediating (formative evaluation), as in ranking or grading, the fundamental task is that of determining the direction of improvement of superiority, and the mere avoidance of the "cutting scores" problem that is required before you can

establish grades does not avoid the logical task of establishing, i.e., justifying and evaluative assertion. Thus I see the preference for formative over summative as — from one perspective — an attempt to limit the amount of evaluative logic that one has to get into, but it does not eliminate the first and crucial step, the step that refutes both relativism and empiricism.

Relatedly, the recent tremendous emphasis on implementation and implementability as meta-evaluative criteria for the merit of evaluations can be seen as another attempt to duck the head-on confrontation with the necessity for demonstrating the *validity* of categorical value judgments, especially those involved in grading. The validity of value judgments, whether they are gradings or rankings, is what the empiricist and relativist deny; but it is a problem that must be faced, and it cannot be converted into the problem of whether the program achieves the goals of its instigators or whether an evaluation is implemented by its clients. Goal-achievement and evaluation-implementation are perfectly compatible with a categorical denial of all merit in the program or evaluation; their absence is perfectly compatible with a categorical assertion of flawless merit. In short, these proposed substitutes are not even universal correlates of the concept they seek to replace, let alone definitional components. (Perspectivism accommodates the need for multiple accounts of reality as perspectives from which we build up a true picture, not as a set of true pictures of different and inconsistent realities. The ethicist believes that objective moral evaluations are possible.)

So far we have talked very favorably about the consumerist ideology. Other strands in the position advocated here must also be recognized as implicitly supported by our criticism of the alternatives to them. These include the perspectivist and ethicist strands that stand opposed relativism and empiricism, the holistic orientation that is the alternative to reductionism (the other half of positivism), and the self-referent ideology that contrasts with separatism. We should add a word about what may seem to be the most obvious of all models for a consumerist ideologue, namely *Consumer Reports* product evaluations. While these serve as a good enough model to demonstrate failures in most of the alternatives more widely accepted in program evaluation, especially educational program evaluation, it must not be thought that the present author regards them as flawless. I have elsewhere said something about factual and logical errors and separatist bias in *Consumer Reports* ("Product Evaluation" in N. Smith, ed., *New Models of Program Evaluation*, Sage, 1981). Although *Consumer Reports* is not as good as it was and it has now accumulated even more years across which the separatist/managerial crime of refusal to discuss its methodologies and errors in an explicit and non-defensive way has been exacerbated many times, and although there are now other consumer magazines which do considerably better work than *Consumer Reports* in particular fields, *Consumer Reports* is still a very good model for most types of product evaluation.

The Multimodel

Evaluation is a very peculiar breed of cat. The considerable charm of each of a dozen radically different models for it, well represented in this book, can only be explained by the fact that it is a chimerical, Janus-faced and volatile being. Even at the level of aphorism, one is constantly attracted by radical variations in such claims as "evaluation is one-third education and one-third art — including the arts of composition, graphics, and politics" or "evaluation should be driven one-third by the professional obligation to improvement, one-third by the society's need for quality, and one-third by the need to economize." The "Ninety-Five Theses" of the Cronbach group carry this further. Analogies with other subjects keep springing into life: architecture is one that seems particularly appealing, with its powerful combination of aesthetic component with the engineering necessities, and with the economics and needs assessment that must be taken into account before a structure can be successful. A dozen others have been advocated as paradigms, from anthropology to operations research.

But during these last few years, it is not accidental that two rather similar approaches to clarification of the practice of evaluation have emerged and gained a certain amount of support. They both represent an attempt at distilling solid principles from the models, but they also represent a kind of model in their own right. These two approaches are the *Evaluation Standards* approach, and the *Evaluation Checklist* approach, to which we will turn in a moment. It is not accidental that both are consumer-oriented; we all know the kinds of checklists that we get out of consumer magazines and which facilitate our evaluation of alternatives for purchase, and we all know the way in which professional standards are used as checklists when supposedly questionable behavior by professionals is under scrutiny. More than this practical and value-orientation is involved here, however. I think that the checklist approach — if I may use the term to cover both instantiations of what I see as essentially a similar point of view — represents a kind of model in its own right. It is not like one of the relatively simple and relatively monolithic models with which we normally associate the term. But the emergence says somethimg about the subject of evaluation, something about its complexity and its relation to other subjects; I shall call it the Multimodel, an ungainly minotaur among models. (The complex CIPP model is an important intermediate case.)

The Multimodel is multiple in a number of ways. In the first place, it commits evaluation to being *multi-field* — that is, applicable to products, proposals, personnel, plans and potentials, not just programs. Then it is *multi-disciplinary* (rather than inter-disciplinary); this means that solid economic analysis, solid ethical analysis, solid ethnographics and statistical analysis, and several other types of analysis are often required in doing a particular evaluation, and not just some standard blend of small parts of these. (Consequently, teams and consultants

are often better than any soloist.) The investigations along each of these and other dimensions, some of which are devoted to entirely different disciplines, constitute a set of dimensions for an evaluation, which must eventually be integrated, since the overall type of conclusion for an evaluation (a grading, a ranking, and apportioning) is often pre-determined by the client's needs and resources. In many respects, the *multi-dimensionality* is the most crucial logical element in evaluation, because specific evaluative conclusions are only attainable through the synthesis of a number of dimensions; some involving needs assessments or other sources of value; others referring to various types of performance.

Another aspect of the multiple nature of evaluation concerns what can be called its need for *multiple perspectives* on something, even in the final report. It is often absolutely essential that different points of view on the same program or product be taken into account before any attempt at synthesis is begun, and some must be preserved to the end. The necessity here is sometimes an ethical one as well as a scientific one.

Relatedly, evaluation is a *multi-level* enterprise. When one gets a call over the phone to ask if one could possibly evaluate a certain program in an unrealistically short time-frame, it is entirely appropriate to respond that one most certainly can, indeed that one can evaluate it there and then, over the phone and without charge. One does have, after all, a considerable background of common sense and evidence about related programs which make it possible to produce an evaluation at this superficial level. We do not associate such evaluations with professionality or with high validity, but that may be a little too severe depending upon the extent of the evaluator's professional background, the similarity of the present example to other well-documented cases and the nature of the evaluative conclusion that is being requested. But if we move down from that superficial level, it is clear that there is a wide range of levels of validity/cost/credibility among which a choice must be made in order to remain within the resources of time and budget. Given certain demands for credibility, comprehensiveness, validity, and so on, there may not be a solution within the constraints of professionality, time, and budget. But more commonly there are many, and it is this that must lead one to recognize the importance of the notion of multiple levels (of analysis, evidential support, documentation) in coming to understand the nature of evaluation. One could go on; *multiple methodologies, multiple functions, multiple impacts, multiple reporting formats* — evaluation is a multiplicity of multiples.

To conclude, then, let me simply list the dimensions that must be taken into account in doing most evaluations, whether of product or program, personnel or proposals. There are certainly special features of the evaluation of — for example — teachers that do not jump out from this listing. But even the four-part checklist that I have suggested above for the evaluation of teachers can be seen to be buried in the following checklist, and indeed it can be enriched in a worthwhile way by paying more attention to some of the steps in this longer effort.

Checklists can function in different ways — there are checklists that list desiderata, and there are checklists that list necessitata. This checklist comes from the latter end of the spectrum, and it is relatively rarely that one can afford to dispense with at least a quick professional check on each of the checkpoints mentioned here. Checklists are also sometimes of a one-pass nature, and sometimes of a multiple pass, or iterative nature. Again, this is of the latter kind; one can't answer all the questions that come up under each of the early headings in adequate detail until one has studied some of the later dimensions; and, having studied them, one must come back and rewrite an earlier treatment, which will in turn force one to refine the later analysis that depends on the former. In designing and in critiquing evaluations, as well as in carrying one out, one is never quite done with this checklist.

The simple terms that I use for the title of each dimension need much unpacking, and they are there just as labels to remind the reader of a string of associated questions. More details will be found in the current edition of *Evaluation Thesaurus*, but I think enough is implied by the mere titles and the word or two that I attach to some of them to convey a sense of the case for the Multimodel. The traditional social science approach deals at most with half of these checkpoints and deals with those, in most cases, extremely superficially, as far as evaluation needs are concerned.

The Key Evaluation Checklist

1. *Description.* An infinity of descriptions is possible, of which a sub-infinity would be false, another sub-infinity irrelevant, another overlong, another overshort, and so on. Whereas relativisim infers from the fact that a large number would be perfectly satisfactory to the conclusion that there are no absolute standards here, perspectivism draws the more modest conclusion that there are a number of right answers, several of which need to be added together to give an answer that is both true and comprehensive, a fact which in no way alters the falsehood or irrelevance or redundancy of many other compound descriptions and hence the difference between right and wrong. The description with which we begin the iterative cycles through the checklist is the client's description; but what we finish up with must be the evaluator's description, and it must be based, if possible, on discussions with consumers, staff, audiences, and other stakeholders.

2. *Client.* Who is commissioning the evaluation, and in what role are they acting? (Distinguish from inventors, consumers, initiators, and so on.)

3. *Background and Context.* Of the evaluation and of whatever is being

evaluated: the hopes and fears. (This checkpoint will be set aside in the early stages of an evaluation that is to have a goal-free phase.)

4. *Resources (or strengths assessment).* For the evaluation and for whatever is being evaluated.

5. *Consumer.* Distinguish the targeted population from the impacted population (in a goal-based approach), and the directly impacted from the indirectly impacted.

6. *Values.* The needs assessment, the ideals review, the relevant professional standards, expert survey, functional or conceptual analysis, and so on. The source of values for the evaluation. To be sharply distinguished from a wants assessment ("market research") unless *no relevant needs* exist.

7. *Process.* Here we have to consider the legal, political, aesthetic, and scientific standards, some of which will have emerged from the values review, and apply them to the intrinsic nature of whatever is being evaluated.

8. *Outcomes.* Here the traditional social, scientific, engineering, medical, etc., methodologies come into their own, except that we must treat discovering unintended outcomes as of equal importance with the search along the intended dimensions of impact.

9. *Generalizability, Exportability, Saleability.* Across sites, staff, clients, and consumers.

10. *Costs.* Money and non-money, direct and indirect.

11. *Comparisons.* The selection of the "critical competitors" is often the most important act of the evaluator, since the winner may be one the client had not considered (but which is perfectly feasible).

12. *Significance.* A synthesis of all the above.

13. *Remediation.* There may or may not be some of these recommendations — they do not follow automatically from the conclusions of all evaluations.

14. *Report.* As complicated as the description, with concern for timing, media, format, and presenters, to a degree quite unlike the preparation for publication of scientific results in a scientific journal.

15. *Meta evaluation.* The reminder that evaluation is self-referent — the

requirement that one cycle the evaluation itself — its design and final form — through the above checklist.

Conclusion

Evaluation practice is still the victim of fallacious ideologies, because we have not applied the essential insight that evaluation is a self-referent discipline. The plethora of evaluation models provides a fascinating perspective on the complexity of this new subject, perhaps the keystone in the arch of disciplined intellectual endeavor. We can only build that arch strong enough to support the huge load of educational and social enterprises that it must bear if we come to understand its architecture and thus the function of its keystone considerably better, and in so doing, come to understand better everything else that we know.

References

Centra, J.A. *Determining Faculty Effectiveness*. San Francisco: Jossey-Bass, 1979.

Davis, B.G., Scriven, M., and Thomas, S. *The Evaluation of Composition Instruction*. Inverness, California: Edge-Press, 1981.

House, Ernest. *Evaluating with Validity*. Beverly Hills, California: Sage, 1980.

Scriven, M. "Product Evaluation." In N. Smith (ed.) *New Models of Program Evaluation*. Beverly Hills, California: Sage, 1981.

Scriven, M. *Evaluation Thesaurus* (3rd ed.). Inverness, California: Edge-Press, 1981.

Scriven, M. "Summative Teacher Evaluation." In J. Millman (ed.) *Handbook of Teacher Evaluation*. Beverly Hills, California: Sage, 1981.

3

How We Think About Evaluation

Ernest R. House

Much of our everyday thinking is metaphorical in nature. That is, we experience one thing in terms of another, according to such theorists as Lakoff and Johnson (1980). They present the following metaphor about argument as an example:

Arguments Are Wars

- Your claims are *indefensible*
- He *attacked every weak point* in my argument
- His criticisms were *right on target*
- I *demolished* his argument (Lakoff and Johnson, 1980, p. 4).

Underlying these separate metaphoric statements is a deep-seated metaphor: *Arguments Are Wars*. This generative metaphor is the basis for a number of expressions, and these expressions constitute a systematic, recognizable pattern. Based primarily upon such evidence,

I would like to thank Lee Cronbach, Robert Ennis, Don Hogben, Mark Johnson, Sandra Mathison, James Pearson, and Paul Silver for providing helpful comments.

From Ernest R. House, "How We Think About Evaluation," pp. 5–25 in *Philosophy of Evaluation* (New Directions for Program Evaluation, no. 19).

some linguists and philosophers contend that such extended metaphors, which occur in our ordinary thinking, are not haphazard or idiosyncratic: All of us employ them in a systematic fashion to structure the way we think about the world. Thus, these metaphoric concepts are extended, conventional, and intersubjective—much like language itself. Moreover, in structuring our thinking about argument in terms of concepts about war, we do more than just express ourselves colorfully. We actually win or lose arguments, attack and defend positions, and gain or lose ground. We live and experience arguments in these terms. The metaphor—*Arguments Are Wars*—shapes our actual behavior.

Until recently, the employment of metaphor was thought to be merely ornamental. Metaphor was used to make an expression more poetic or to emphasize a point rhetorically. However, novel experiences usually are structured in terms of more familiar ones, abstract concepts in terms of more concrete ones, and cultural notions in terms of physical ones. Metaphor is essential to our most complicated thought processes and a vital intellectual tool that we use to understand the world. For example, argument as war reflects aspects of our concept of *argument*. The metaphor highlights how participants in an argument relate to each other, how they treat one another, and how the argument might progress. However, argument as dance would indicate quite a different set of relationships between participants—that is, opponents would be partners. Therefore, *Arguments Are Dances* is not a common metaphor in our culture.

Complex concepts also can be structured by more than one metaphor. For example, the concept of argument is shaped not only by *Arguments Are Wars* but also by other metaphors:

Arguments Are Buildings

- The argument is *shaky*
- We need to *construct* a *strong* argument
- The argument *collapsed*
- Is that the *foundation* of your argument?
  (Lakoff and Johnson, 1980, p. 46).

Arguments as buildings indicates other aspects of our concept of argument that we consider to be important. *Arguments Are Buildings* highlights how arguments are put together, based, and constructed—quite different aspects that those conveyed by *Arguments Are Wars*. We might refer to how arguments proceed in waves, are calm or stormy, and appear on the surface as opposed to what is beneath the surface—that is, *Arguments Are Oceans*. But we do not.

The images of wars and buildings are quite different. But neither are they incompatible with one another. In emphasizing two distinct aspects of our notion of arguments, the two metaphors do not present a single, consistent image but they are coherent. This fundamental coherence is demonstrated by the fact that we mix *Arguments Are Wars* and *Arguments Are Buildings* in our thinking:

- When I *attacked* his argument, it *collapsed*
- The *foundation* of his argument is the *weak point*
- We need to *construct* an argument that is *defensible*
- Your *defense* is a *shaky* one

As the last statements indicates, even a strange mix of metaphors makes sense to us, since these two aspects of argument are used and associated with one another so commonly. Coherent metaphors often fit together by being subcategories of a major category and sharing a common entailment. For example:

Love Is a Journey

- It's been a long, bumpy road
- We're just spinning our wheels
- We've gotten off the track
- Our marriage is on the rocks
  (Lakoff and Johnson, 1980, p. 44).

Although all of these statements concern journeys, they are based on different kinds of journeys: a car trip, a train trip, and a sea voyage. The concrete images in each sentence define a more general category and, in that sense, are coherent rather than consistent. They fit together but do not compose a single image.

Quite a number of other metaphors also shape our conception or argument, usually in terms of familiar, concrete, and physical experiences like wars and buildings. Abstract, complex concepts are usually shaped by a number of metaphors that are coherent because the ideas themselves are too complex to be conveyed by one single, consistent image. Whether argument commonly is seen as a war or a dance is culturally determined, and the user of the concept ordinarily is not aware of the underlying metaphor that shapes his or her experience of the actual phenomenon. The user believes that arguments naturally happen that way. Thus, arguments follow certain social patterns because of the common conception that the participants have (Turner, 1974). These fundamental metaphoric concepts are essential to our under-

standing of the world because they form coherent systems of thought that we use extensively in everyday life (Lakoff and Johnson, 1980).

Metaphors Underlying Social Policy

Schön (1979) contends that social problem-setting is mediated by the stories people tell about troublesome situations. The framing of the social problem depends on the metaphors underlying the stories, and how the problems are framed is critical to the solutions that emerge. For example, a pervasive description of the social services is that they are "fragmented," and the implicit solution to this problem is that they be "coordinated." But services seen as "fragmented" could also be seen more benignly as "autonomous." Therefore, the underlying metaphor gives shape and direction to the problem solution.

Schön maintains that we are guided in our thinking about social policy by pervasive, tacit images that he calls *generative metaphors*, in which one frame of reference is carried over to another situation. These metaphors generally are used because the user is immersed in the experience of the phenomenon. Thus, these guiding images are necessary to his or her thinking. For example, urban renewal can be viewed in different ways. The slum can be seen as a once healthy community that has become diseased. A social planner with such an image envisions wholesale redesign and reconstruction as the cure to urban blight. However, the slum can also be viewed as a viable, low-income community, which offers its residents important social benefits. The second view obviously implies strikingly different prescriptions for improving the community.

The predominant image of the slum in the 1950s was that of a blighted community. However, in the 1960s the slum as a natural community arose as a countermetaphor that vied for public and expert attention in social planning. Each image features certain themes—taken from a reality that is ambiguous and indeterminate—that define the phenomenon of the slum (Schön, 1979). In the first vision, terms like *blight, health, renewal, cycle of decay*, and *integrated plan* are highlighted in descriptions of social planning. In the second vision, *home, patterns of interaction, informal networks*, and *dislocation* represent key ideas about what should be done with slums. Each overall image presents a view of social reality by selecting, naming, and relating elements within the chosen framework. According to Schön, *naming* and *framing* are the key processes in such conceptualization. By selecting certain elements and coherently organizing them, those processes explain what is wrong in a

particular situation and suggest a transformation. Data are converted to recommendations.

Naming and framing proceed by generative metaphor. The researcher sees the slum as a blight or as a natural community. In seeing A as B, the evaluation implicit in B is carried over to A. The first metaphor is that of disease and cure. The second is that of natural community versus artificial community. The transferred evaluations are based on images deeply ingrained within our culture, and once we define a complex situation as either health and disease or nature and artifice we know in which direction to move. Seeing A as B greatly facilitates our ability to diagnose and prescribe. On the other hand, it may lead us to overlook other important features in the situation that the metaphor does not capture. Since generative metaphors usually are implied rather than expressed openly, important features may pass undetected. Schön argues that we should be more aware of our generative metaphors, and that this is best done by analyzing the problem-setting stories we tell. The "deep" metaphor accounts for why some elements are included in the story while others are not, some assumptions are taken to be true in spite of disconfirming evidence, and some recommendations seem obvious. It is the metaphor of the slum as diseased—or as a natural community—that gives shape to the study and direction of a social planner's actions.

Industrial Production as a Metaphor for Social Programs

Evaluation concepts are often derived from fundamental, generative, and deep-seated metaphors that remain hidden. These metaphors guide one's thinking in certain directions. In this sense, evaluative thinking is no different from the metaphoric thinking in other areas. To illustrate this point, I turn to an examination of the ideas presented in Rossi and others' book, *Evaluation: A Systematic Approach* (1979). This book is one of the most widely used textbooks in the teaching of evaluation, and the authors' work is exemplary of thinking in the field of evaluation—and pervasively metaphoric.

The most fundamental metaphor that the authors use is that of the delivery of social services as industrial production. In their conceptualization, social services are utilities or commodities that are required by the public, and it is the duty of a social program to supply these services. The notion that services are produced by social programs and that they are to be delivered to a clientele manifests the production metaphor. For example, related ideas taken from the book include:

Social Service Delivery Is Industrial Production

- Program elements are defined in terms of *time, costs, procedures,* or a *product*
- A delivery system consists of organizational arrangements that provide program services
- These services are delivered to a target population
- Program development is equivalent to designing the system
- There are production runs
- Services can be calculated in terms of service units delivered
- One should monitor the delivery of these services
- There are operational indicators of success
- A *monitoring evaluation* is an assessment of whether the program conforms to the design and reaches the target.

Even more specifically, social programs as conceived in the preceding examples not only as industrial production in general but as a particular kind of industrial production—that is, an assembly line. At other times within the book, social programs are viewed as machines:

Social Programs Are Machines

- A program consists of elements
- Program elements are discrete intervention activities
- Programs may be broad, complex, but also have component parts
- They are implemented
- They operate according to a design
- They produce benefits, effects, and outcomes
- They can be replicated and replaced
- They can be tested
- They can be fine-tuned
- Accountability means conformity to program specifications
- A major failure is unstandardized treatment
- Variables can be manipulated to achieve results

Rossi and others employ yet a third specific metaphor of industrial production—that of a pipeline or conduit:

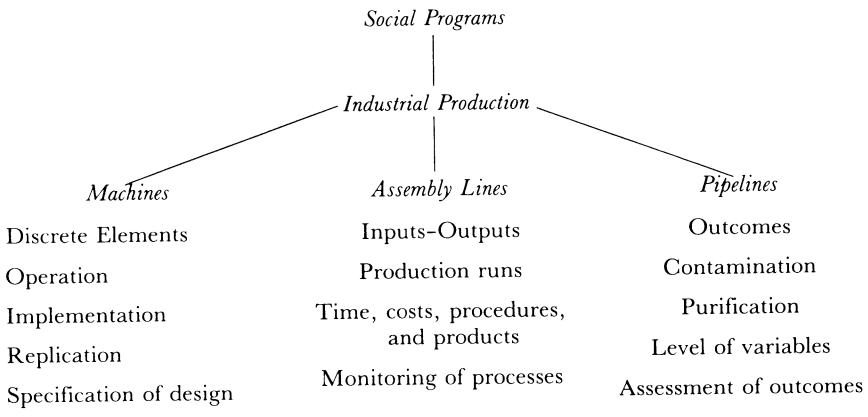
Social Delivery Systems Are Conduits

- A delivery system is a combination of pathways that allow access to services
- A major failure in systems is dilution of the treatment to an insufficient amount
- Outcomes always represent changes in the level of measurable variables
- Contaminants may either enhance or mask true changes
- Assessing net intervention effects requires purification of outcomes by purging contaminating elements
- The point of assessing the magnitude of effects is to rule out causal links between inputs and outcomes
- The unreliability of measuring instruments may dilute the difference in outcomes

Social programs as machines, assembly lines, and conduits all fit the overall metaphor of social programs as industrial production. But each metaphor emphasizes a slightly different aspect of the nature of social programs. That is, in thinking about social programs, one may emphasize the way social programs are put together and operate to produce benefits. Or the inputs and outputs, the raw materials, and labor that go into programs may be emphasized—or the way benefits or services are delivered to the program recipients. Therefore, social programs can be conceived as involving all of these aspects, and the various separate metaphors are used to emphasize different ones.

Different conceptions of what evaluation entails follow from these different metaphors of social programs: conformity to program design, monitoring of production processes, and measuring of purified outcomes. The evaluation of the program corresponds to the perceived nature of social programs. Sometimes the emphasis is on design specifications and the parts of the program; sometimes it is on the inputs and outputs, and other times the emphasis is placed on the outcomes—the latter metaphor being that of a pipeline with certain substances that issue from it and the corresponding evaluation resembling a chemical analysis from which the evaluator seeks to ascertain the results, purified of possible contamination. Of course, the overall metaphor is that of industrial production but there is no single, consistent image for all of the metaphors. Taken together, the three images present a coherent picture of social programs as industrial production (see Figure 1). The internal coherence among these metaphors is demonstrated in the

Figure 1. Metaphoric Conception of Social Programs



Source: The figure is based on Rossi and others, 1979.

mixed metaphors that make sense within this conceptual structure and used throughout the book. For example, delivery systems are said to deliver programs or program elements or treatments. Programs may produce benefits or outcomes or outputs. These terms are used interchangeably.

The internal coherence of these metaphors is derived from their shared entailments. That is, the better the discrete elements of the program fit together, the more efficiently the time, costs, and procedures are converted into products, and the more outcomes the program delivers. Hence, the design of the programs, the inputs of the program, and the delivery of outcomes are linked together, though by no means synonymous with one another. There is a sequentiality that underlies all three: a sense in which a social program must be created, made, or produced, and in which it must reach the people for whom it is intended. The concept of industrial production is not the only way in which this process can be conceived and made coherent, but it is one way of doing so. Of course, such an overall metaphor entails certain types of evaluations.

The ubiquitous metaphoric nature of these concepts is illustrated further by a detailed examination of the concrete images. For example, the assembly line is a fundamental image in our culture, and it is not surprising that Rossi and others apply this notion to social programs. Raw materials come in one end of the assembly line, labor is performed in stages, and products come out the other end. Underlying the assembly line concept are deeper metaphors that define both labor and time as material resources. A material resource is a kind of substance that

can be used in a manufacturing process, quantified precisely, assigned a monetary value per unit of quantity, serve a purposeful end, and used up progressively as it serves its purpose. If time and labor are material resources, they also can be quantified, assigned a value per unit, serve a purposeful end, and be used up (Lakoff and Johnson, 1980). In addition, in our society labor is seen as an activity—and an activity is defined as a substance. Hence, labor can be treated as a substance and a material resource; likewise, time commonly is viewed as a substance—defined in units. Conceiving of labor and time as substances and material resources permits them to be measured, used up, assigned monetary value, and used for various ends. Thus, in conceiving of social programs as assembly lines, Rossi and others can state "*Program elements* may be defined in terms of *time, costs, procedures*, and *products*" (p. 137). Doing a cost-benefit analysis of how time and labor are used in social programs is a logical next step and an important part of the authors' ultimate thinking.

In such a metaphoric framework, efficiency quite naturally looms large as a criterion for successful social programs. Social programs are expected to be efficient just as industrial production is expected to be. In the Rossi and others' conceptualization a comprehensive evaluation must include monitoring, impact assessment, and cost-benefit or cost-effective analysis, and one chapter is entitled "Measuring Efficiency." Production functions and econometrics are an extension of this type of analysis, although these authors do not go so far, choosing instead to emphasize both the desirability and difficulty of measuring the benefits and costs of social programs. However, other theorists have been less reticent in setting up equations for social programs that model the production processes, and the discovery of such production functions has at times been the object of considerable federal effort such as the evaluations of Title I of the Elementary and Secondary Education Act of 1965 (McLaughlin, 1975).

Rossi and others also repeatedly speak about social programs as being *effective, efficient, adequate,* and *useful.* This language suggests that there is a job to be done and that the program must accomplish this job. The notion of a particular job or task to be performed is congruent with the entire industrial production metaphor. Within the world defined by the fundamental metaphor, these terms become major evaluative terms. They indicate that the program is good if one can apply these terms and also suggest where to look to see if the program is good. They become major criteria of evaluation, criteria that are entailed by the general metaphors.

Targets and Goals as Metaphors

Although the industrial production metaphors dominate Rossi and others' view of evaluation, other metaphors also play a key role in their thinking. These are the metaphors of *target* and *goal*. The target metaphor is used extensively in the book in reference to *target problems, target populations*, and *impact*. The social program has impact on the targets. Presumably, the targets are social problems that social planners attack or alleviate.

Social Problems Are Targets

- Programs and projects are *aimed* at the target problems
- The program can be *misguided*
- The problems are located *in* the target population
- Problems are distributed and have location, extent, type, scope, and depth
- A needs assessment determines the nature, extent, and location of social problems
- Targets have boundaries and rules of inclusion and exclusion
- Programs have *impact* on the targets
- Impacts vary in magnitude
- An *impact evaluation* assesses the extent to which the program causes changes in the desired direction in the target population (Rossi and others, 1979, p. 16).

The underlying metaphoric conception is that social problems are targets, and that the social program is aimed at the target. Hitting the target results in the impact, and the magnitude of the impact is an indicator of how effective the program has been. The evaluator must measure the impact of the program on the target. The target population must be defined, and social services are directed not *to* the target population but *at* the target problems. The targeting metaphor entails quite a different image than the industrial production metaphors but one coherent with these. The target metaphor is employed when the authors discuss the ultimate effects of the program, and the industrial production metaphors are used in discussing the monitoring of the program itself. They use the pipeline or conduit image when discussing outcomes and the target image when discussing impact, which is the ultimate result.

Once again, the metaphors can be mixed to a certain degree.

Interventions can be delivered to the target or directed to the target population. *Coverage* is defined as the extent to which the program reaches the target population, combining the notions of both delivery systems and targets. The targeting metaphor maps out a particular aspect of social programs and their evaluation. And, according to Rossi and others, a comprehensive evaluation includes monitoring, impact assessment, and cost-benefit analysis.

A third possible metaphor employed extensively in the book is that of the *goal*. However, there is some question as to whether it should still be called a metaphor. That is, goal is used literally to mean *purpose*. The notion of goals appears to be derived originally from sports or games, but it has lost much of its metaphoric connotation. Concepts can be derived from metaphors and gradually transformed into literal meanings, thus losing their metaphoric meanings. The more the concepts are used, the more they take on the meaning of their new application. For example, the *foot* of the mountain is clearly metaphorical in origin but is close to meaning literally the bottom of the mountain. On the other hand, most of the terms and concepts of industrial production applied to social programs are clearly metaphorical, though some are more so than others. A term like *outcomes* is well on its way to literal usage in the evaluation community. Thus, there seem to be degrees of metaphoric meaning for particular concepts, and these meanings change over time. In a few years we may see literal dictionary definitions for terms that we now consider metaphoric. Their metaphoric nature will then reside only in their etymology. With that caveat I will proceed to a metaphoric analysis of *goal* and its connection to the other concepts, bearing in mind that these notions may have passed into literal usage.

The original definition of goal seems to be that of a physical distance, in which a goal is set along a course—such as a race course, a game, or a sport. In the course of the race, game, or sport, the player is supposed to reach or attain that goal.

Program Activities Are Goal-Directed Movements

- Goals are unattained standards
- Goals and objectives can be *set* and measured
- There are *gaps* between the goals and reality, between *where* one wants to be and where one it
- The intervention *closes* the *gap* between the two
- One seeks *convergence* between the program design and its implementation; there is *distance* between them
- Evaluations can *direct* the *course* of social life

- Evaluation can be a firm *guide*
- Surveys assess whether the target has been *reached.*

The latter statement is derived from a mixing of the goal and target metaphors and indicates the coherence between the dominant metaphors. This mixing of metaphors can be seen clearly in Rossi and others' definition of impact evaluation: "impact evaluation-assessment is the extent to which the program causes change in the desired direction in the target population" (p. 16). Although the basic metaphor is that of impact and target, impact is defined in terms of direction and physical distance, which is essentially goal language. Often in the assessment of goals and objectives, a land surveying metaphor of marking off the landscape, triangulating, and measuring distance is used. So again, even though these various metaphors do not present a single consistent image of evaluation, they form a coherent conception. Rossi and others' conceptualization of evaluation is so complex that several metaphors are necessary to highlight different aspects. No single metaphor will do, but both the target and the goal metaphors highlight the aim, direction, and purpose of the program.

Target is ultimately derived from war and sport. Originally a target was a light round shield used in combat, and this came to be the object one aimed at in target practice. The etymology of *goal* is less clear. Apparently, the term was derived from an ancient rustic sport *(Oxford English Dictionary).* In Old English it meant an obstacle, boundary, or limit. Eventually goal came to mean the terminal point of a race or the posts between which a ball is driven in a game or sport, as in football or soccer. And in archery the goal is the mark aimed at—that is, the target. But the notion that a game is nonserious, or just for fun, has not carried over from goal's original meaning. The goal metaphor has been stripped of its nonserious side and is used to mean a serious striving for achievement, or an earnest contest that is perhaps akin to war. Even though sports language is employed, social program evaluation is at least as serious as a game in the National Football League, which is serious indeed. Within this context, the player attains a goal in a sport or a game by scoring. Originally a score was a cut or a mark on something to keep count and eventually came to mean a line drawn for runners or marksmen to stand at. Ultimately, to *score* as a verb came to mean to make points in a game or contest *(Oxford English Dictionary).* Score also means one's performance on a test, as in a test score. Scores on outcome measures are very important in Rossi and others' framework: For example, net effects are measured in differences in scores on

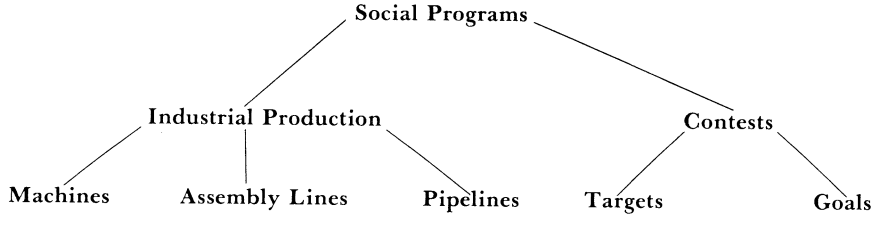
outcome measures. Apparently, both the target and goal metaphors, which are so pervasive in social program language, are derived from equating social programs with sports or games, or, more generally, contests (see Figure 2). Yet many of the metaphoric meanings are now lost, especially for goals.

In general, there is a strong directionality within all of these diverse yet coherent metaphors. Industrial production, such as in an assembly line or conduit, moves from one point to another, as does the trajectory traced by a missile as in archery, by a runner in a race, or the throwing of a ball through the goal as in a sport. Implicit in these metaphors is the movement of a physical object from one place to another. As more services are produced by the assembly line, more are delivered to the target population. As more services hit the target, there is more impact from the program. The more goals that are attained, the higher the scores and the more successful the program. Beneath these fundamental metaphors are the rather abstract notions of linearity and directionality—movement from one point to another. All of the basic metaphors share this abstract property and serve the purpose of indicating a certain kind of movement that is correlated with program success. Greater production, stronger impact, and more goals attained are all correlates of program success. Underlying the coherent metaphors, then, is a shared topoligical concept, a concept that remains invariant across metaphors.

The Building Metaphor in Program Evaluation

Yet another set of terms is applied directly to the evaluation itself rather than to the program. The evaluation must be a *firm assessment*, be a *firm guide*, produce *firm estimates of effects* and *solid information*, and not result in *faulty conclusions*. The construction terminology in evaluation is derived from such conventional metaphors as *Arguments Are*

Figure 2. An Extended Metaphoric Conception of Social Programs



Source: The figure is based on Rossi and others, 1979.

Buildings (Lakoff and Johnson, 1980). Evaluations, like arguments and theories, are conceptualized as physical structures, quite possibly because evaluations are recognized tacitly as arguments themselves. The building and construction metaphor is quite commonly applied to evaluations, regardless of the particular metaphors applied to social programs. Evaluations are expected to be firm, solid, well-constructed, and so on. They share the same basic societal metaphors as arguments, and these terms are applied not only in Rossi and others but in much of the evaluation literature.

Thus, some aspects of evaluation are derived from particular metaphors about what social programs are. In conceiving of social programs as industrial production, the evaluation takes shape from the nature of the object evaluated. However, other aspects of evaluation are rendered by more general metaphors, such as *Arguments Are Buildings*. These aspects of evaluation seem to be independent of notions of what social programs are supposed to be. And there are even more fundamental metaphors employed in the articulation of programs and evaluation. These include metaphoric structurings of time and labor as material resources, events as objects, and activities as substances. Although these ideas fit well into the overall conceptual scheme, they are not dependent upon it. They are readily available in everyday thought. Hence, the metaphors employed in evaluations of social programs are both special ones drawn specifically for this purpose and common ones used in many other settings.

Even this does not exhaust the metaphoric structure of the book by Rossi and others. The discussion of cost-benefit analysis draws upon the economic and accounting literature, which has its own metaphoric structure. But, although the metaphoric structure is pervasive and extremely important in shaping the ideas in the book, it is difficult to discover and make explicit. We share so much of the common experience of assembly lines, goals, and targets that the discussion seems literal rather than metaphoric. In this sense, the metaphoric structure is nearly invisible.

The Metaphoric Nature of Evaluation

The realization that a great deal of evaluative thought is metaphoric in nature will no doubt surprise and disturb many evaluators. Many see evaluation of social programs as applied social science and may wonder how metaphors could be so crucial to their thinking. The metaphoric analysis raises a number of questions: To what degree does

metaphor characterize all evaluative thinking? How does it work? Where do these metaphors come from? Are there conflicts between different schemes, depending upon one's underlying metaphors? Are all metaphors equally good? What is the scientific status of evaluation if this analysis is correct? Does such an analysis lead to relativism? Unfortunately, discussion of these issues is beyond the scope of this chapter. And, in general, the role of metaphor in thought is not well understood. (For further philosophical discussions of metaphor, see Sacks, 1978; Ortony, 1979; Johnson, 1981). This section briefly touches upon the origins of these metaphors, the values they embody, the purposes they serve, their scientific status, and their appropriateness.

Industrial production and sporting contests are often used as metaphors in evaluation because they are pervasive experiences in our society, and production and competition are primary values. Taken together, they entail winning. It is not surprising that we should evaluate our social programs from frameworks derived from such central experiences, and that these structural metaphors embody core values of American society. In employing these metaphors to evaluate social programs, we bring those values to bear upon social programs, sometimes explicitly but often tacitly.

Faced with the new task of evaluating social programs in the past two decades, evaluation theorists have turned to areas of their own experience that seem better defined. Evaluations therefore have been conceived and structured through concepts derived from other domains of experience. Differences in conceptions of evaluation often reflect differences in underlying metaphors, which are in turn derived from certain cultural experiences. The ultimate purpose of this metaphoric structuring is to tell us how to act as evaluators. In spite of the often expressed skepticism about the role of evaluation theory, without such conceptions to guide us we would not know how to act as evaluators. "In all aspects of life, not just in politics or in love, we define our reality in terms of metaphors and then proceed to act on the basis of the metaphors" (Lakoff and Johnson, 1980, p. 158).

The metaphors discussed to this point substantially define the reality of the evaluator's world. Once an evaluator has accepted the basic metaphors, certain entailments follow. Of course, our thinking is not entirely determined by the metaphors we use, and we are not enslaved by our own concepts. The relationship between metaphors and thinking is more one of likelihood—of probability—than one of determination. For example, it is very likely that an evaluator will be

led to certain types of evaluations if he or she sees social programs as industrial production. Furthermore, evaluators are taught certain metaphors as part of their training; it is part of their enculturation. Although they might conceivably overcome a particular way of viewing the world, as defined by certain metaphors, the pressure to be consistent is more likely to make them follow through with particular types of evaluations—to elaborate the metaphor, as it were. Such metaphoric structuring enables us to do a number of things in our evaluations and prevents us from doing others. Every way of viewing the world eliminates alternative possibilities. Metaphors highlight some things and shadow others, and the predominant views we have are necessarily partial and particular. Furthermore, metaphoric structures are derived from domains of our experience that are seldom logically consistent and fully coherent. This lack of consistency and coherence often carries over into our conceptions of evaluation.

Many evaluators and social planners see social programs as industrial production, targeting, and goal attainment and cannot see programs in any other way. Other theorists employ similar metaphors in their articulation of what evaluation is. In fact, these metaphors underlie one of the dominant views of social programs among professional evaluators in the United States, not because people adopt Rossi and others' point of view but because theorists draw upon common experiences and a common intellectual framework. However, as common as this point of view is, there are yet other evaluation theorists and planners who adopt different views of social programs and evaluations. They employ different metaphors—with different results in their conceptualization of evaluation. For example, responsive, illuminative, and stakeholder-based evaluation suggest different metaphors at work.

Not just any metaphor will do in structuring the concept of an evaluation. A former student of mine once write a paper in which she developed an evaluation system based upon the beliefs of a tribe of Plains Indians. Such a scheme is intriguing but unlikely to have much application in contemporary America, just as metaphors of industrial production would not have much appeal to the Plains Indians. Appropriate metaphors must be rooted in the experiences of the culture to be applicable. Metaphors used to evaluate social programs necessarily will be close to our core social values, although some theorists have attempted to create new evaluation approaches by deliberately developing different metaphors (Smith, 1981).

Embracing a particular set of metaphors not only expresses certain values but also promotes them. It is in the nature of metaphor that

certain things are emphasized and others deemphasized. Efficiency, effectiveness, goal seeking, and values of industrial production are promoted in the Rossi and others framework. The authors explicitly advocate these values which are embodied in their conceptual apparatus. Conceptions of evaluation are not value-neutral, and much of this inherent evaluation is embedded within the metaphoric structure. Different conceptions emphasize different values or weight the same values differently. Also the more common the metaphors employed to structure evaluation, the more persuasive and invisible the metaphors will be. Unusual metaphors are creative, but conventional metaphors shape most of our thinking and therefore seem natural.

Employing certain metaphors allows us not only to promulgate certain values but to do a number of other things as well, such as to refer and identify causes. For example, the employment of ontological metaphors, such as defining labor and time as substances, allows us to quantify things (Lakoff and Johnson, 1980). Defining a territory or putting a boundary around something is an act of quantification. Bounded objects, like social programs and social problems, have scope, dimension, and size. Within such a framework; an evaluator can locate social problems and measure them. This is usually accomplished through a *survey,* the original purpose of which was to determine the form, extent, and situation of parts of a tract of ground by linear and angular measurement (*Oxford English Dictionary).*

Other entities can be thought of as containers. For example, the participants are *in* the program, but they cannot be *in* the problem, although they can be *part* of the problem. Containers define a limited space, with a bounded surface, a center, and periphery, and can be seen as holding a substance, which may vary in amount. If one sees the program as a container object, it can be measured.

Programs Are Containers

- That is *not much* of a program.
- The program *does not have any content.*
- The program *lacks substance.*
- That is the *core* of the program.

Machines, assembly lines, and pipelines can all be viewed as container objects. Things can be located in them or be a part of them. The notion of a container object is abstract and deeply embedded in our thinking. In addition, one can conceive of the outcomes of a program as substances—which issue from the program container. The program has outcomes (a

substance). For example, when discussing the outcomes of programs, Rossi and others often switch to their *Social Programs Are Conduits* metaphor. Contamination and purification are of primary concern, so that one can measure the net outcomes: "An outcome is always a change in the level of a measurable variable" (Rossi and others, p. 164). The *gross* outcome effects are the measures of overall impact but the *net* outcome effects are those left after confounding effects have been removed (a mixing here of the conduit metaphor and an accounting metaphor, which they also use).

Both social programs and program outcomes can be quantified and measured via their metaphorical conversion into objects and substances, but the nature of their measurement differs. As metaphorical objects, programs have size, scope, and dimension, and require different methods of measurement than does the metaphorical substance of the outcomes. Objects may be described, and program description has received much attention. But, measurements of programs themselves have been limited compared to measurement of outcomes. Therefore, social programs normally must be converted into other categories, such as the time, costs, and procedures of the assembly line, before measurements become possible. In contrast, outcomes lend themselves more easily to direct measurement. An object may be dissimilar in its different parts but any quantity of a substance is like any other part of the substance. Hence, conceiving of outcomes as substances permits *cardinal* measurement—that is, the use of an interval scale. To be measurable in this way means that every instance of a commodity is a sum of perfectly identical parts or units. This is not literally true of the outcomes of social programs, but they often are treated that way in order to be quantifiable. In any case, quantification of programs and their outcomes is greatly facilitated by their metaphorical conversion into concrete objects and substances.

If outcomes are quantifiable we can define them as members of a particular statistical distribution, such as a normal curve. We might infer from the degree of overlap between the pre- and postmeasured distributions the likelihood of the postmeasure coming from a different statistical population. Hence, we begin employing statistical models, in which one treats the outcome scores as member of particular populations. A statistical treatment of impact data is a logical next step for Rossi and others to take, but the preliminary conceptual apparatus for doing this resides in the fundamental metaphors that they employ. For certain purposes, programs and activities are treated as if they were objects and substances. Obviously such conversions are useful.

The statistical model might be called a metaphor, but there is a significant difference between it and the overall metaphoric framework of Rossi and others. The statistical model is internally consistent: There is a single representation from which one can draw logical inferences that do not contradict each other. This is more similar to a scientific or mathematic model than the overall evaluation conceptualization of Rossi and others. But, there is no question that metaphoric thinking plays an important role in scientific thinking. For example, Kurt Lewin's theories draw heavily on analogies with physical theories in the use of certain concepts, such as *field, sector, force,* and *fluidity* (Black, 1962). More recently, cognitive psychology has conceived of the human mind as a computer, employing such concepts as *information processing, feedback, encoding,* and *memory storage* (Boyd, 1979). Metaphors play a constitutive role in scientific theories, although exactly how this role is performed is a matter of dispute (Kuhn, 1979). Of course, the use of metaphors does not mean that a conception is nonscientific. The traditional view of science as a clear, unambiguous, testable rendering of external reality in literal language has given way to a view of knowledge as based upon menal constructions (Ortony, 1979). Perhaps the significant difference between scientific theories and conceptions of evaluation is their internal consistency. Formal scientific theories can be seen as attempts to extend a set of metaphors consistently, whereas metaphors underlying evaluation are rarely consistent (Lakoff and Johnson, 1980).

However, there is another important difference between conceptions of evaluation and scientific theories—a difference of purpose. One might imagine that *minds are computers* and investigate the way in which information processing is done by the mind. According to Boyd (1979), a term like this provides "epistemic access" to the phenomenon being investigated. Other investigators may extend the concept until it becomes descriptive of how the mind functions—and eventually far removed from what the term means in the study of computers. But metaphors in conceptions of evaluation are not quite like this. The purpose of *Social Programs Are Conduits* is not to arrive at a finer definition of social programs (though one may do so). The researchers in the field do not investigate the extent to which social programs really resemble conduits. Rather, the main purpose is to impose the metaphor so that one knows how to act—that is, how to evaluate. Given the fundamental metaphors, certain investigations and judgments become possible. The judgments are about whether the social programs are any good, not about whether the metaphors fit and not even about finer descriptions

of the programs themselves. The difference is between describing and evaluating: These are fundamentally different acts. In both cases metaphors are employed but to different ends.

Perhaps this difference can be seen more clearly if the roles are reversed. Suppose that the *Minds Are Computers* metaphor is used for evaluation purposes. One can imagine trying to assess the information processing capacity, the memory storage, and the encoding processes of the mind—even comparing different minds on these dimensions. No doubt various criteria for evaluating would emerge from our experience with computers, and no doubt one could develop reliable procedures for assessment. One might end up saying that the information processing of a particular mind was very strong but the feedback mechanisms were poor. One would use concepts similar to those in cognitive psychology, but the purpose would be quite different than that of trying to describe the mind by computer analogies or judging the goodness of fit. In the act of evaluating, the metaphor is used to generate criteria for making judgments of worth. Conversely, if one used the metaphor *Social Programs Are Assembly Lines* in a descriptive investigation, one would investigate the degree to which social programs actually resemble assembly lines, modifying one's notions of industrial production to fit the operation of social programs. This is not what evaluation theorists or evaluators do.

In general, these underlying metaphors provide some of the basic concepts that instruct us on how to proceed. If one sees arguments as wars, one will argue in a certain fashion. If one sees social programs as industrial production, then one will evaluate in a certain fashion. Once one is committed to a particular metaphor, certain entailments arise for both thought and action. Thus, the dominant metaphors shape our actions. But not all metaphors are equally good for the purposes they are supposed to serve. There can be good and bad and appropriate and inappropriate metaphors, just as there can be good and bad social programs (Binkley, 1981; Booth, 1978; Loewenberg, 1981). The sense in which a metaphor is true, correct, or appropriate is beyond the limits of this chapter, but what can be said briefly is that the underlying metaphors must be considered within the context of the overall conception of evaluation. That is, one must judge the consequences of the overall conception. These judgments must be based upon criteria broader than being simply true or false as the notion is commonly understood. Evaluators of social programs must embrace comprehensive notions of correctness, including rightness and wrongness. The obligation of the evaluator is broader than that of the describer.

In retrospect, perhaps it is not so surprising that metaphoric thinking is important in evaluation. Black (1962) has explored the similarity between scientific models and metaphors and concludes that both models and metaphors play an indispensable role in scientific thinking. In fact, all intellectual pursuits rely upon such "exercises of the imagination. . . . Perhaps every science must start with metaphor and end with algebra; and perhaps without the metaphor there would never have been any algebra" (Black, 1962, p. 242).

References

Binkley, T. "On the Truth and Probity of Metaphor." In M. Johnson (Ed.), *Philosophical Perspectives on Metaphor.* Minneapolis: University of Minnesota Press, 1981.

Black, M. *Models and Metaphors:* Ithaca, N.Y.: Cornell University Press, 1962.

Booth, W. C. "Metaphor as Rhetoric: The Problem of Evaluation." In S. Sacks (Ed.), *On Metaphor.* Chicago: University of Chicago Press, 1978.

Boyd, R. "Metaphor and Theory Change: What Is 'Metaphor' a Metaphor For?" In A. Ortony (Ed.), *Metaphor and Thought.* Cambridge, England: Cambridge University Press, 1979.

Johnson, M. (Ed.). *Philosophical Perspectives on Metaphor.* Minneapolis: University of Minnesota Press, 1981.

Kuhn, T. S. "Metaphor in Science." In A. Ortony (Ed.), *Metaphor and Thought.* Cambridge, England: Cambridge University Press, 1979.

Lakoff, G., and Johnson, M. *Metaphors We Live By.* Chicago: University of Chicago Press, 1980.

Loewenberg, I. "Identifying Metaphors." In M. Johnson (Ed.), *Philosophical Perspectives on Metaphor.* Minneapolis: University of Minnesota Press, 1981.

McLaughlin, M. W. *Evaluation and Reform.* Cambridge, M.A.: Ballinger, 1975.

Ortony, A. (Ed.). *Metaphor and Thought.* Cambridge, England: Cambridge University Press, 1979.

Rossi, P. H., Freeman, H. E., and Wright, S. R. *Evaluation: A Systematic Approach.* Beverly Hills, Calif.: Sage, 1979.

Sacks, S. (Ed.). *On Metaphor.* Chicago: University of Chicago Press, 1978.

Schön, D. A. "Generative Metaphor: A Perspective on Problem-Setting in Social Policy." In A. Ortony (Ed.), *Metaphor and Thought.* Cambridge, England: Cambridge University Press, 1979.

Smith, N. L. (Ed.). *Metaphors for Evaluation.* Beverly Hills, Calif.: Sage Publications, 1981.

Turner, V. *Dramas, Fields, and Metaphors.* Ithaca, N.Y.: Cornell University Press, 1974.

4

The Preconditions for Successful Evaluation

Is There an Ideal Paradigm?

Dennis J. Palumbo and David Nachmias

The field of evaluation research is undergoing an identity crisis. From its initial surge in the 1960s when it was dominated by a single paradigm and researchers believed that its potential was unlimited, it has been undergoing a metamorphosis. Instead of a dominant paradigm, several alternative approaches to evaluation have emerged and skepticism about its potential contributions to public policy has been raised. House

An earlier draft of this paper was presented at the International Political Science Association Meeting, Rio de Janerio, Brazil, August 1982. We are indebted to Egon Guba, Paula Wright and Peter deLeon for their most perceptive and helpful comments.

From Dennis J. Palumbo and David Nachmias, "The Preconditons for Successful Evaluation: Is There an Ideal Paradigm?" *Policy Sciences,* 1983, 16, 67–79.

(1980:11), for example, maintains that "The current evaluation scene is marked by vitality and disorder. The scale, ubiquity, and diversity of evaluation activities make comprehension difficult, even for those operating within the field."

Such commentaries on the state of the field are very recent. In the 1960s and early 1970s, a dominant paradigm of evaluation research crystallized in the evaluation community. Leaning heavily on Campbell's (1969) vision of the "Experimenting Society," it was aimed at determining whether program goals were being achieved through the employment of rigorous quantitative research methodologies. Findings obtained in this manner were expected to find their way almost instantaneously in the policy process leading to marked improvements in societal conditions. The challenge was to establish "truth"; the problem of *Speaking Truth To Power* (Wildavsky, 1979) remained mute.

Research anomalies, the personal experiences of both evaluators and policymakers, and changing conceptions on the capabilities of government to ameliorate social problems, have raised second thoughts about the dominant paradigm. This article examines the dimensions of the identity crisis in the field of evaluation research, and advances the case for a greater congruence between theory, methods and practice. Four analytically distinct aspects of evaluation research are discussed: the relationship of evaluation to decisionmaking processes; the methodologies for conducting evaluations; the congruence between methodologies and organizational behavior; and the relationship between the evaluators and program managers. Obviously, these are not the only significant aspects of an ideal evaluation paradigm, but they are sufficient to convey the major developments and the inherent complexities of the field. There is also some overlap among the four aspects as is pointed out in the following sections.

The Ideal Role of Evaluations in Decisionmaking

The somewhat naive, apolitical notion that once the "truth" as revealed by evaluation research was known, programs would be changed in accord with the evaluation findings has been replaced with a welter of different and conflicting ideas about the ideal role of evaluations in decisionmaking.

There are at least four distinct roles that evaluations can play in decisionmaking: (1) they can result in terminate–continue decisions about programs; (2) they can lead to decisions on program improvements; (3) they can constitute only one informational component in decisionmaking, and not necessarily the most important; and (4) they can inform and educate society. The first and second roles are similar in that they both lead to program improvements, but in the first, the improvement occurs when programs that fail to achieve their objectives are terminated, and in the second improvements result because programs are made more effective.

There are proponents for each of these four roles. The traditional view has been that evaluations should lead to terminate–continue decisions (summative). There is, however, an increasing emphasis on the program improvement role (formative evaluation)

of evaluations. This shift reflects an increased awareness that many evaluation findings are inconclusive (Bernstein and Freeman, 1975), and that social programs tend to exert a differential impact on their target population depending on the degree of need (Hofferbert, 1982). Still others suggest that evaluation research, while being one source of information for policy decisions, is not necessarily the best guide (Wildavsky, 1979). The idealized evaluator, Brandl (1978:8) writes, seeks the truth; but elected officials are concerned with the "good," "Or, more likely, an accommodation of competing ideals of what is good, or perhaps just re-election." Brandl, who is a state legislator, maintains that legislators see evaluators as just another interest group advocating its value preferences, but the evaluation community is not among the most powerful of interest groups. The fourth role of evaluation is the most ambiguous one in the sense that it should be used primarily for "influencing and instructing society" (Cronbach, 1980).

Concerns over the proper role of evaluations in decisionmaking have remained unresolved and the confusion seems to be growing rather than receding. For example, in their Introduction to the 1981 *Evaluation Studies Review Annual*, which includes some of the best research produced in the latter half of the 1970s, Freeman and Solomon claim that the principal trend in evaluation research is a concern with the role of evaluation in decisions, and that there is a growing emphasis on understanding and modifying the front and back ends of the evaluation process so as to improve utilization.

The "front-end" refers to pre-evaluation activities such as evaluability assessment. The purpose here is "to build a shared understanding and, if possible, to achieve consensus on evaluation requirements and strategies to maximize the applicability of results and increase the likelihood of program improvement" (p. 16). Accordingly, the ideal role of evaluations is to increase the likelihood of program improvement and not lead to terminate–continue decisions. The "back end" refers to the utilization of evaluations. It is here that there is a great deal of confusion as to whether evaluations are used, and over what organizational structures and incentives promote utilization. Freeman and Solomon (p. 17) contend that many of the generalizations on utilization ". . . are tentative and untested." The utilization of evaluation findings continues to be an important problem in evaluation research (Nachmias and Felbinger, 1982). In an attempt to increase research utilization, there has been a growing interest in implementation research and process evaluations. The ideal role of evaluation in this research is sequential, incremental program improvement.

Perhaps most important, as Freeman and Solomon point out, there are disquieting indicators about the extent to which evaluation research is likely to be supported by government. The 1980s have been marked by a tightening of federal expenditures in all departments and agencies, and growing skepticism about the worth of evaluation research. Officials in the Reagan Administration and in Congress are not supportive of evaluation research in particular nor of social science research in general (Whyte, 1982). Ideology, not research, plays a greater role in policymaking (deLeon, 1983).

This raises serious dilemmas about the future of evaluation research and its use.

Consider the position of Rossi and Freeman (1982:15), for example, who view the field as "a robust area of activity devoted to collecting, analyzing, and interpreting information on the need for, implementation of, and impact of intervention efforts to better the lot of humankind by improving social conditions and community life." In referring to "intervention efforts," the authors leave off the word *public*, thereby implying that evaluation research can be used to evaluate private sector interventions as well. But, for the most part, program evaluation involves *public* programs, and as the authors acknowledge, the growth of government was an important stimulus for the development of evalution research. Implicit in this view is the belief that government can improve societal conditions. In the dominant paradigm, evaluation research is to be used as a way of deciding whether or not a program should be continued. In the long run, societal conditions will be improved because ineffective, unworkable programs will be terminated and effective ones will be continued. This view takes for granted an expanded role for government and excludes the possibility that programs could be terminated because government should not have intervened at all in certain areas. This latter, of course, is the ideological persuasion of the Reagan Administration, and a conclusion reached by some scholars with regard to governmental regulatory practices in several policy areas (e.g., Wilson, 1980; Bardach and Kagan, 1982; Savas, 1982). Evaluation research that concludes that the private sector, in particular voluntary associations, can provide services more effectively and at less cost than government agencies is quite different from program evaluation aimed at improving the operations of existing government programs. The former, broader kind of policy evaluation is concerned with the macro-question of finding the most effective institutional arrangement for allocating society's scarce resources, not with the micro-problem of whether a specific public program can be improved.

In many ways, evaluation research is an analytic continuation of reformist traditions. Political scientists, in particular, believe that they can find ways of making government both more effective and more responsive, and their evaluations most often are of governmental institutions and processes rather than of specific policies or programs. As Rossi and Freeman (1982:31) observe, by the 1980s, evaluation research became more "than an isolated academic concern; it thrived in the context of the social policy and public administration movement." These movements have stimulated evaluation research by contributing a new army of professionals trained in public policy, policy analysis and public administration programs across the country. The programs are primarily concerned with improving public management and public policy. They attempt to do this not only by imparting conceptual and analytical skills but also by training and socializing people for professional careers in the public sector. The latter aspect may well have the stronger impact.

The shift that occurred in governmental finances in the 1980s is ominous for evaluation research because program evaluation is likely to be the first cutback when programs are being retrenched. In the 1960s and 1970s, evaluations were used to

legitimize government intervention by demonstrating that government was concerned with the accountability of new programs. But now that public programs are being curtailed, often irrespective of their effectiveness, evaluation research is less justified and may even become a liability. Furthermore, if government turns over social programs to the private sector and volunteers, what role can evaluation research play? This is not a rhetorical question because it is unclear to us, at least, whether evaluation research can play *any* role in a system in which the "invisible hand" is supposed to make public policy decisions. For example, if consumers are given a choice of what schools they want to send their children to through a voucher system, then there is no need for independent evaluations to determine which programs are working; the parents themselves do the evaluating and they vote with their feet, so to speak, by taking their children out of the schools they do not like and sending them to the ones they prefer. Not all of these choices will be made on the basis of whether or not the programs are effective; they will be made for reasons of status, class, race, and other social factors. Whether a voucher system will eventually be used in education is not the point here. The point is that there is an increasing body of policy research that advocates a diminished role for government in areas such as health delivery (Olsen, 1981), social regulation (Wilson, 1980), and urban services (Savas, 1982).

The Ideal Methodology for Conducting Evaluations

There is much more consensus in the field about the most appropriate methodology for conducting evaluations. The dominant paradigm is what House (1980) terms the "behavioral objectives" approach. This approach is goal-oriented, uses experimental and quasi-experimental research designs, and heavily leans on quantitative data. For example, Freeman and Solomon (1981) in their *Evaluation Studies Review Annual*, include 25 evaluation articles of programs in education, human resources and social services, law and public safety, health, mental health, substance abuse and the environment. Twenty-three of the twenty-five use quantitative data and most of these are experimental or quasi-experimental evaluations. Although the articles in this volume are not representative of all evaluation research, Freeman and Solomon claim that their sample includes work that is "higher in methodological quality than those produced by the overall evaluation effort" (p. 13). The articles they include are in a very real sense, ideal, and the ideal here undoubtedly is quantitative research.

Now, more than in the past, there is some acceptance for an alternative methodology sometimes called "qualitative" evaluation research and sometimes called "naturalistic" inquiry or ethnographic research (Guba and Lincoln, 1981; Patton, 1980). This methodology is becoming somewhat more popular because it resolves the evaluator's predicament by attempting to represent all significant value positions in the evaluation (House, 1980). At the same time, some find it more effective for purposes of utilization. Patton (1978) suggests that the "personal factor," that is, the relationships that develop between the program manager and the evaluator, are critical for utilization.

Although the naturalistic methodologies have added an important dimension to evaluation research, there are no signs that a paradigm shift will occur.

Congruence Between Methodologies and Organizational Decisionmaking

The methodologies used by evaluators must be congruent with organizational behavior for the findings to be utilized. Guba and Lincoln (1981:26) call this requirement for method and theory congruence "value-boundedness." They argue that "the values undergirding the substantive theory selected to guide the inquiry [must be] resonant with the values undergirding the methodological [paradigmatic] theory." Thus, if organizations are conceived of as bureaucracies in the Weberian sense, "conventional methodological approaches serve quite well, since they, like bureaucratic theory, are very systems oriented" (p. 27).

At present, evaluators assume that decisionmakers analyze the situation first, then act. More explicitly, that decisionmakers, *before they act*, identify goals, specify alternative strategies for attaining them, assess the alternatives against a standard, such as costs and benefits, and then select the "best" alternative (the synoptic paradigm). But if organizations in fact do the opposite - if they act first and then analyze, evaluate and rationalize what they did - then evaluations based on the synoptic paradigm will be out of resonance.

The dominant evaluation paradigm is synoptic both in its methodology and in the assumptions it makes about the way organizations behave. For example, a great deal of evaluation research is based on the assumptions of micro-economic theory. That is, individuals and organizations are assumed to be rational, and to behave so as to maximize or at least "satisfice" some identifiable goal or a set of goals. Accordingly, the principal job of the analyst is to model organizational choices, to deduce desired objectives, and to estimate the relative effectiveness of different strategies for attaining them. But what if decisionmakers pay little attention to the relative effectiveness of different strategies for attaining the objectives stipulated in an evaluation? Suppose that they are more likely to act first and only then analyze why it is they did what they did. Suppose further that this kind of analysis is done informally and intuitively and that it is stored in personal memories, so that when the organization is faced with a similar situation in the future, its members recall previous experiences and try to apply these to the new situation. They do *not* re-analyze, search for alternatives, nor establish desired objectives. Instead, they repeat the cycle of acting then analyzing why they acted as they did. Obviously, in such a situation, a priori, micro-economics methods will be of little use in helping organizational actors. All the same, the economists' approach - with its claim for rationality and efficiency - has been legitimized by government officials, and since there are far more economists in government than there are other social scientists (e.g., political scientists, sociologists, anthropologists), it is not surprising that the micro-economics approach to evaluation is prominent. The synoptic, "problem-information-decision" cycle has seldom been

supported by empirical studies assessing the utilization of information in the policy-making process (Dutton et al., 1980; Rothman, 1980; Hargrove, 1980; Weiss, 1981). Furthermore, contemporary organization theory advances the thesis that organizations are not best conceived as rational instruments for achieving societal goals but as organized anarchies, or loosely-coupled systems (Cohen et al., 1972; Dunsire, 1978; Wieck, 1976). Still, the synoptic evaluation model continues to dominate.

If organizations behave in the latter manner, what are the implications for evaluation research? One thing seems clear: synoptic and micro-economic evaluations are likely to miss the mark because organizations (decisionmakers and individuals in organizations) are not looking for the one best way or most efficient alternative for solving a problem. They are instead searching for support for actions already taken, and for support that serves the interest of various components of the policy-shaping community (Walker, 1981). Thus, evaluators concerned with utilization should attempt to discover what societal needs have been met by the decisions undertaken; they should determine which stakeholders' interests are served by organizational actions.

The ideal paradigm of evaluation congruent with the reversed decision cycle is quite different from the dominant, synoptic paradigm. This can be demonstrated by examining the preconditions for successful evaluation based on the synoptic, goal-directed paradigm, and then contrasting them with the preconditions for successful evaluation based on the reversed decision cycle. Before describing these preconditions, we should note that this is a heuristic device, meant to illuminate the argument. Not everyone will agree with all of these preconditions, nor do we claim that we have described every conceivable one.

For the goal-directed paradigm, the preconditions of ideal evaluations are as follows:.

Precondition 1: The program to be evaluated must have clearly stated, operational goals on which all relevant participants agree.

Precondition 2: An explicit technology for achieving these goals must exist and be implemented.

Precondition 3: The methodology for determining the extent to which the program produces the outcomes and for controlling exogenous factors must be available.

Precondition 4: The managers of the program being evaluated must be committed to working toward achieving program goals.

Precondition 5: Decisionmakers must be committed to utilize the results of the evaluation.

Only rarely are all these preconditions met. Program goals tend to be amorphous, multiple and contradictory. Legislation is almost always ambiguous because it is politically expedient to be ambiguous. Moreover, policy goals do not become less ambiguous or more uniform when they are delegated to administrators for implementation. Administrative discretion allows for different interpretations of goals. Blau (1955) reported that two employment agency units, which had similar official goals, actually were very different in what they really were attempting to accomplish. One

unit was highly competitive, with members striving to outproduce each other in terms of the numbers of individuals placed. In the other unit, cooperation and quality of placement was stressed. Broad goals become more specific in implementation but the more specific goals take a variety of possible forms. Within the implementing organization, units can move into divergent directions which even if they are consistent with the stated goals could interfere with each other. Furthermore, often the technology for achieving policy goals is unavailable. We do not know how to eliminate the causes of poverty and crime nor how to design the perfect implementing organization. Although we have a great many sophisticated methodological tools at our disposal, we still do not know how to determine cause and effect relationships in *uncontrolled* environments which are the most typical in policy evaluation research. The ideal program manager is committed to working toward achieving program goals, but we do not often have ideal managers. Finally, decisionmakers are not committed to using the results of evaluations, especially when the results are either inconclusive or at variance with political objectives and ideological dispositions. The goal-directed paradigm derives from preconditions that seldom, if ever, can be realized. Why, then, should we build evaluation research on these preconditions? Why not, instead, build it on more realistic preconditions?

For the reverse decision cycle (i.e., action first, then analysis) the preconditions for evaluation are the following:

Precondition 1: Some of the activities engaged in by program administrators lead to positive outcomes valued by some stakeholders.

Precondition 2: The positive outcomes are related, even if only indirectly to the formally stated goals of the agency.

Precondition 3: The evaluation focuses mainly on positive outcomes.

Precondition 4: Program managers trust the evaluators.

Precondition 5: The evaluation may or may not be utilized depending on the findings.

Let us consider each of the preconditions and see how they contrast with those of the goal-directed paradigm. The first precondition is compatible with the ideas of "goal-free" evaluation (Scriven, 1972). The evaluator does not assume that there are preconceived goals the program is to achieve but, instead, looks at what is being done and identifies the positive outcomes produced by these activities. These outcomes are positive in the sense that they are beneficial or important to some of the stakeholders associated with the program (Stake, 1974). Since the agency will have formally stated goals for which it will be held accountable by some stakeholders, the positive outcomes must be correlated with the formally stated goals even if they are not identical.

The third precondition is especially important and may be why most evaluations based on the synoptic paradigm fail. The major challenge faced by a program manager charged with the responsibility of implementing a new program is to generate and maintain enthusiasm on the part of those who are implementing the program (Ripley and Franklin, 1982). The program manager has to convince the implementors that the

new goals and objectives are worthwhile, and that they will serve their best interests. For those who are new to the program it is easier for the manager to instill a sense of purpose and commitment to the program. In all cases, however, it is essential for the program manager to emphasize the positive aspects of the program.

All programs inevitably have both positive and negative aspects. To maintain a high level of activity and implementation, the positive aspects must be emphasized. But evaluation conducted under the synoptic, dominant paradigm inevitably turns up negative aspects of programs. This is because programs seldom, if ever, will meet their original intentions perfectly. Program objectives change and evolve as they are being implemented (Majone and Wildavsky, 1979; Palumbo and Harder, 1981), and we usually do not have the social technology, adequate resources and administrative know-how, for a program to achieve all of its original objectives. Thus, a program manager has to be wary of, and even downright hostile, toward a synoptic evaluation because it invariably points out negative aspects. Most programs operate in a politically volatile climate. Thus, no matter how solid it is methodologically, a synoptic evaluation cannot help a program manager because parts of it can be used against him or her.

This brings us to the fourth precondition. The program manager must be willing to trust the evaluator, and the likelihood of trust will increase with the conviction that the evaluation will produce helpful and useful information for the program manager. Thus, it is impossible for the evaluator to be totally independent or to engage in detached scientific research if his/her objective is utilization. This may sound as if we are saying evaluations should engage in whitewashing, but this is not what this precondition requires. Whitewashing occurs when wrongdoing is covered up or if the evaluation is perfunctory or gives a biased presentation of data. What precondition four requires is that evaluators become advocates for the programs and emphasize its positive outcomes. Similar to lawyers in the adversarial process of a trial, evaluators taking this approach should make the best case they can for the program they are evaluating. Given that values cannot be eliminated from evaluating (Guba and Lincoln, 1981), they may as well become an open, explicit aspect of the evaluation. Indeed, if the program is producing only negative or dysfunctional outcomes, it is the evaluator's obligation to report this. Obviously, he or she should not expect the evaluation to be used if all that is found are negative outcomes (Precondition 5). The program still may continue if it serves the political interests of decisionmakers and benefits some key stakeholders. Some programs exist only because they are part of a pork barrel, and negative evaluation findings are likely to be discredited or ignored.

The preconditions of the reverse decision cycle might be characterized as political evaluation (Stufflebeam and Webster, 1981). Political evaluation has the following advantages: it is acceptable to program managers; it helps build a positive image of social programs; it helps point to aspects of the programs that achieve positive outcomes; and it helps build and maintain constituencies' support for the program. At the same time, political evaluation has some inherent problems: it is biased toward emphasizing the positive aspects of programs; it is susceptible to cooptation by

program managers; and it is scientifically and (perhaps) professionally less credible than synoptic evaluations. We make no ethical judgement about whether political evaluation is morally inferior to scientific evaluation. Nor do we argue that it would produce more responsive public policies than summative evaluations. All we suggest here is that it is more congruent with the actual policy process. It is based on more realistic assumptions about organizational behavior. Incremental decisionmaking and organized anarchies (i.e., organizations) are not as idealistic as the synoptic, goal-oriented paradigm is, but they more realistically convey the complexities of policy behavior. Idealistically, we might like to increase the amount of rationality in organizations, but until this objective is reached, the political evaluation paradigm is a more realistic alternative for the conduct of policy evaluation.

The Ideal Evaluator-Manager Relationships

If there is disagreement about the ideal role of evaluations in decisionmaking, about the ideal methodologies for conducting evaluations, and about the need for congruence between evaluation methodologies and organizational behavior, there is almost total confusion about the ideal relationship between evaluators and those being evaluated. To a large extent, the relationship depends upon the methodologies used. The synoptic paradigm requires detachment and independence between the two parties. This, however, raises serious problems with respect to utilization. In order for evaluations to be of use to and be used by program managers, it is essential for a great deal of both "front end" and "back end" interaction to take place (Patton, 1978; Cronbach, 1980; Sproull and Larkey, 1979; Conner, 1979). But close interactions of evaluators and program managers may impinge on the validity and credibility of the results.

As it evolved during the 1970s, evaluation research moved out of the universities and the majority of evaluations were done by private firms in response to Requests-for-Proposals (RFPs) issued by government agencies. This development changed the nature of the enterprise as well as the relationship between the evaluator and the program manager. Because most evaluations are done on the basis of RFPs written by government agencies, not only are the goals and objectives of evaluations somewhat dictated by government agencies, but the entire enterprise has been put at the mercy of the government and, therefore, of the political process. Evaluations conducted under these conditions are political not only in the sense that they are subject to the political pressures of stakeholders (Guba and Lincoln, 1981:22), but also in the sense that they are subject to the political whim of change in government regimes. In contrast to evaluation research conducted in universities where the objective is to build knowledge and where the interests of the researchers tend to predominate (Coleman, 1972), evaluation research conducted by consulting firms is dominated by the interests of those being evaluated. House (1980:11) observes that "Too often evaluators do what sponsors want them to do. Too often evaluators misconceive the nature of their tasks

and do injustice to the social programs that they evaluate."

Many evaluations conducted under these conditions may be of interest to practitioners but they are of little interest and have scant impact on academic research. Consider, for example, the kinds of problems addressed in the *Evaluation Studies Review Annual* mentioned above (Freeman and Solomon, 1981). One can almost take a Proxmire-like attitude about some of the topics reviewed, when viewed from the halls of academe: "Decreasing Dog Litter: Behavioral Consultation to Help A Community Group"; "Evaluation of the Costs and Benefits of Motorcycle Helmet Laws"; "Homeowner Warranties: A Study of the Need and Demand for Protection Against Unanticipated Repair Expenses." These, of course, are eminently practical concerns, but of minimal interest to academic social scientists. If the government did not fund this research, who would do it and why? Certainly no one in the traditional disciplines in universities. Instead it would be done by service agencies such as Consumer's Union.

Thus, the problem of the ideal relationship between evaluator and program manager is a much broader one than the question of whether or not those being evaluated should participate in the formulation and execution of an evaluation. It touches upon institutional relationships, profit motives, and the resulting credibility and validity of evaluations.

Conclusions

The dominant paradigm of current evaluation research is goal-directed, views its role in decisionmaking in a narrow sense, and is in the logical positivistic tradition. Whereas these attributes contributed to the acceptance of policy evaluation research in academia, the overall contributions of research to the policy process have been limited. A major reason for this is that the dominant paradigm is more congruent with research traditions in the social sciences than with the actual policy process. It is patterned along assumptions and norms conducive to the scientific estate, not the political-bureaucratic estate.

Recent developments in the policy evaluation field pose serious challenges to the prevailing paradigm. The proposition that evaluations should lead to terminate-continue decisions is being questioned. There is an increasing use of qualitative, case study methods as a methodology for conducting evaluations, and there is a shift in organization theory away from the rational, top-down bureaucratic model toward the loosely-coupled, reticular approach.

Although these developments reinforce each other, they will not be reflected in the practice of policy evaluation research for some time. Paradigmatic shifts are slow processes, especially when multiple participants with conflicting objectives are involved. Clark and McKibbin (1982:672) have described how views of school administrators are changing from the top-down model to the loosely-coupled approach, but added a caveat:

> The newer organizational perspectives . . . will eventually permeate the field of educational administration because they are useful. Their spread will be slow, however, because the language, politics, and psychology of rationalism will make it difficult for practitioners to espouse the new perspectives or to abandon safe, rational structures.

The dominant principles are difficult to dislodge, because as these authors point out, it is impractical for an administrator to say to legislators or boards of trustees that redundancy and waste cannot be eliminated, or that goals are determined after one has acted, not before, or that the technology to accomplish desired goals is unavailable. Managers need to have mission statements and formal goals in order to hold their subordinates accountable and to protect themselves from challenges from outside.

As we have argued above, there is at present no ideal evaluation paradigm: the dominant model is both methodologically and institutionally inadequate; the proposed model is currently unattainable. But the tension between the two is not irreconcilable even though the dominant paradigm of evaluation research and its assumptions about how organizations behave are likely to remain unchanged for some time. It is more likely that the dominant paradigm will be adjusted in an incremental, loosely-coupled manner to be more congruent with the policy process, and then be given a new name to show that in fact that is where the field was headed all along. Like all Holy Grails, the ideal evaluation paradigm in all its pristine trappings might well be eternally beyond our grasp. That does not imply that the rewards are not worth the quest. Indeed, the implicit assumption of this article is that they surely are. Thus, the challenge before both evaluation practitioners and sponsors is to define what this ideal paradigm might be – most critically, how it resolves the rigor vs. relevancy rivalry – and begin to move in those directions.

References

Bardach, Eugene, and Robert A. Kagan (1982). *Going By the Book*. Philadelphia: Temple University Press.

Bernstein, Ilene, and Howard Freeman (1975). *Academic and Enterpreneurial Research*. New York: Russel Sage.

Blau, Peter (1955). *The Dynamics of Bureaucracy*. Chicago: University of Chicago Press.

Brandl, John E. (1978). "Evaluation and politics," *Evaluation*. Special Issue.

Campbell, Donald T. (1969). "Reforms as experiments," *American Psychologist* 24: 409–429.

Clark, David L., and Sue McKibbin (1982). "Free orthodoxy to pluralism: New views of school administration," *Phi Delta Kappa* 18: 669–672.

Cohen, Michael D., James G. March, and John P. Olsen (1972). "A garbage can model of organizational choice," *Administrative Science Quarterly* 17: 1–25.

Coleman, James S. (1972). *Policy Research in the Social Sciences*. Morristown, N.J.: General Learning Press.

Conner, Ross F. (1979). "The evaluator–manager relationship: An examination of the sources of conflict and a model for a successful union," in H. C. Schulberg and J. M. Jerrell (eds.), *The Evaluator and Management*. Beverly Hills, CA: Sage, pp. 119–139.

Cronbach, Lee J., et al. (1980). *Toward Reform of Program Evaluation*. San Francisco: Jossey-Bass, 1980.

deLeon, Peter (1983). "Policy evaluation and program termination," *Policy Studies Review*, 2.

Dunsire, Andrew (1978). *Implementation in a Bureaucracy*. New York: St. Martin's Press.

Dutton, W. H., J. N. Danziger, and K. L. Kraemer (1980). "Did the policy fail? The selective use of automated information in the policymaking process," in H. J. Ingram and D. E. Mann (eds.), *Why Policies Succeed or Fail*. Beverly Hills, CA: Sage Publications, pp. 163-184.

Freeman, Harold, and Ann Solomon (eds.) (1981). *Evaluation Studies Review Annual*. Volume 6, Beverly Hills, CA: Sage.

Guba, Egon and Yvonne S. Lincoln (1981). *Effective Evaluation*. San Francisco: Jossey-Bass.

Hargrove, F. C. (1980). "The bureaucratic politics of evaluation," *Public Administration Review* 40: 151-159.

Hofferbert, Richard I. (1982). "Differential program impact as a function of target need: Or why some good policies often seem to fail," *Policy Studies Review* 2: 279-292.

House, Ernest R. (1980). *Evaluating With Validity*. Beverly Hills, CA: Sage.

Majone, G., and Aaron Wildavsky (1979). "Implementation as evolution" in Pressman, Jeffrey and Aaron Wildavsky, *Implementation*. Berkeley, CA: University of California Press.

Nachmias, David, and Claire Felbinger (1982). "Utilization in the policy cycle: Directions for research," *Policy Studies Review* 2: 300-308.

Olsen, Mancur (ed.) (1981). *A New Approach to the Economics of Health Care*. Washington: American Enterprise Institute.

Palumbo, Dennis J., and Elaine Sharp (1980). "Process versus impact evaluation of community corrections," in David Nachmias (ed.), *The Practice of Policy Evaluation*. New York: St. Martin's Press, pp. 288-304.

Palumbo, Dennis J. and Marvin A. Harder (eds.) (1981). *Implementing Public Policy*. Lexington, MA: D.C. Heath and Co.

Patton, Michael Q. (1980). *Qualitative Evaluation Methods*. Beverly Hills, CA: Sage.

Patton, Michael Q. (1978). *Utilization Focused Evaluation*. Beverley Hills, CA: Sage.

Ripley, Randall B., and Grace A. Franklin (1982). *Bureaucracy and Policy Implementation*. Homewood, IL: The Dorsey Press.

Rossi, Peter, and Harold Freeman (1982). *Evaluation*. 2nd ed., Beverly Hills, CA: Sage.

Rothman, Jack (1980). *Using Research in Organizations*. Beverly Hills, CA: Sage.

Savas, E. S. (1982). *Privatizing the Public Sector*. Chatham, NJ: Chatham House.

Scriven, M. (1972). "Pros and cons about goal-free evaluation," *Evaluation Comment* 3: 1-4.

Sproull, Lee, and Patrick Larkey (1979). "Managerial behavior and evaluator effectiveness," in H. C. Schulberg and J. M. Jerrell, *The Evaluator and Management*. Beverly Hills, CA: Sage, pp. 89-105.

Stake, R. E. (1974). "Program Evaluation: Particularly Responsive Evaluations," Unpublished manuscript.

Stufflebeam, Daniel L., and Willim J. Webster (1981). "An analysis of alternative approaches to evaluation," in Howard Freeman and Marian Solomon (eds.), *Evaluation Studies Review Annual*. Vol. 6, Beverly Hills, CA: Sage, pp. 70-86.

Walker, Jack L. (1981). "The diffusion of knowledge, policy communities and agenda setting: The relationship of knowledge and power," in J. Tropman, Milan Dlucy and R. Lind (eds.), *New Strategic Perspectives on Social Policy*. New York: Pergamon Press, pp. 75-96.

Weick, K. E. (1976). "Educational organizations as loosely-coupled systems," *Administrative Science Quarterly*, 21: 1-19.

Weiss, Carol (1981). "Measuring the use of evaluation," in James A. Ciarlo (ed.), *Utilizing Evaluation; Concepts and Measurement Techniques*. Beverly Hills, CA: Sage, pp. 17-33.

Whyte, William Foote (1982). "Social inventions for solving human problems," *American Sociological Review* 47: 1-13.

Wildavsky, Aaron (1979). *Speaking Truth to Power*. Boston: Little Brown.

Wilson, James Q. (ed.) (1980). *The Politics of Regulation*. New York: Basic Books.

II

THE CONTEXT SURROUNDING EVALUATION

In this part of the *Annual* we move from a discussion of different evaluation paradigms to a discussion of contextual (or setting-specific) factors affecting evaluation. Evaluation is not simply a technical activity involving the design and implementation of sound scientific procedures. Evaluators continually strive to achieve a balance between the need for rigorous models and the concerns of various constituencies who utilize, and who are affected by, evaluation data. Since evaluation research is not conducted within a social vacuum, consideration of how contextual factors affect the design, implementation, interpretation, and utilization of evaluation is essential. Successful evaluations are those that are able to strike a balance between the rigors of science and the constraints posed by the people, programs, and settings in which the evaluation is conducted. These constraints include interpersonal conflict, economic limitations, political pressure, and environmental limitations.

The articles in this section focus primarily on contextual issues surrounding national evaluations. While the context of national evaluations may differ to some extent from narrower evaluations, many of the issues are common to all types of evaluations. Some of these contextual issues concern the appropriate role of the federal government in setting the evaluation agenda, the influence of political factors in affecting who and what is evaluated, and the use of evaluation data in establishing public policy.

In most evaluation research, the relationship between evaluators and those being evaluated is a factor affecting the degree of success of the evaluation. Travers and Light review early childhood educational evaluations and point out how these programs are shaped by forces external to the evaluation. They note that program policies evolve as a function of objective environmental conditions as well as various constituency concerns. Therefore, it is critical to view policies within the context of general societal change. To be successful, evaluations must be responsive to continual shifts in program priorities. Travers and Light describe how evaluation research and program policies exist in a dialectical relationship; evaluations affect policies and policies affect the conduct of evaluations. They make an important distinction between evaluations informing policymakers and evaluations being coopted by policymakers (e.g., research versus advocacy). The authors argue that as much as possible, evaluators should maintain their scientific research perspective.

The federal government is a primary provider of funds for and consumer of evaluations. With increasing frequency, mandates for evaluation of social programs are legislated. The extent to which the government (in this case the

U.S. Congress) should be involved in defining the parameters of evaluation research is addressed by St.Pierre. After reviewing several national-level evaluations, St.Pierre concludes that congressional involvement is desirable in proposing evaluation questions, defining relevant target audiences, and establishing the temporal parameters of the evaluation. In contrast, congressional involvement in specifying particular evaluation methodologies or research designs is unnecessary and, in many cases, detrimental to carrying out successful evaluations.

Havender reviews another aspect of the role of the government in evaluation research. He provides a concrete example of how the relationship between the scientist (in this case, evaluator) and policymakers (the government) affect public policy. The article discusses the ban of cyclamates by the FDA and the conditions under which this ban occurred. For the most part, the ban was based on the government's interpretation of the scientific data that cyclamates were harmful to health. Havender suggests, however, that data utilized to make this decision were incorrectly assessed and interpreted. This view is based on the lack of replicable data that were available to make an accurate interpretation, the likelihood that Type I errors were made due to the use of liberal p values (up to .20), and the unrealistic expectations of the FDA regarding the statistical sensitivity of the data. This article provides a clear example of how evaluation data are used to make important policy decisions in ways that may not be appropriate. With recognition of this fact, evaluators will be better able to assess the context under which their data might be used or misused. In turn, this will foster the development of appropriate strategies for dealing with the utilization of evaluation data.

As pointed out by St.Pierre, Congress often commissions evaluations of policy issues it considers. An interesting example of this is a request for evaluation data by the House Subcommittee on International Security and Scientific Affairs for its hearings on chemical warfare. Specifically, the subcommittee requested "information describing deterrence against use of chemical weapons, Soviet and U.S. chemical warfare capabilities, binary chemical weapons, and disarmament." The agency charged with providing this information was the U.S. General Accounting Office's (GAO) Institute for Program Evaluation.

Reprinted here is a series of items related to the chemical warfare debate and the GAO evaluation report. These include the congressional letter requesting the study, the introductory and concluding chapters of the GAO evaluation report, and communications between the Department of Defense and the GAO about the report. In addition, segments of the *Congressional Record* covering this issue and the evaluation report (including, surprisingly, a methodological discussion) as well as two articles from the *Washington Post* commenting on this issue are included. This section of the *Annual* provides an extremely interesting view of factors influencing the design, implementation, interpretation, and utilization of evaluation data. What makes it even more

important is the extreme relevance of the data to international foreign policy. According to Lois-ellin Datta, Associate Director of the GAO Program Evaluation and Methodology Division,

> The study could be credited, if one is persuaded by its conclusions, with saving about $114 million in 1983 alone in development and cumulatively with far greater savings. While many factors influenced the Congressional decision ... the study was among the major influences in the decision, read widely and debated hotly.

We expect that the evaluation community will find this section particularly enlightening, since the influence of contextual factors on evaluation comes through clearly.

Rossi concludes this section with a paper on the prospects for social science under the "Reagan Regime." While the paper was written soon after Reagan took office, the points raised about how the context surrounding evaluation affects its conduct are still relevant. Overall, Rossi's assessment of the situation is mixed. His discouragement comes from the general reduction in resources for social science research and intervention. On the positive side, however, is the fact that these reductions have brought attention to the social sciences, attention that has led some policymakers to reaffirm their commitment to some types of social science research. In a metaphorical summary of his views, Rossi states,

> It has become clearer that we are not the weak, incompetent, and superfluous pussycats the conservatives thought we were.... Social science research and social scientists [are] more like percherons—strong, competent, able to take on a big load and do a good job. ... It does not look as if we will be given ... a chance at glorious martyrdom. (p. 26)

5

Evaluating Early Childhood Demonstration Programs

Jeffrey R. Travers and Richard J. Light

INTRODUCTION

During the last two decades, public and private programs for young children and their families have undergone profound changes. Programs and philosophies have proliferated. Program objectives have broadened. Federal support has increased: Projected expenditures for child care and preschool education alone neared $3 billion several years ago. Target populations have expanded and diversified, as have the constituencies affected by programs; such constituencies reach beyond the target populations themselves.

A sizable evaluation enterprise has grown along with the expansion in programs. Formal outcome measurement has gained increasing acceptance as a tool for policy analysis, as a test of accountability, and to some extent as a guide for improving program practices. Programs have been subjected to scrutiny from all sides, as parents, practitioners, and politicians have become increasingly sophisticated about methods and issues that once were the exclusive preserve of the researcher. At the same time, evaluation has come under attack--some of it politically motivated, some of it justified. Professionals question the technical quality of evaluations, while parents, practitioners, and policy makers complain that studies fail to address their concerns or to reflect program realities. Improvements in evaluation design and outcome measurement have failed to keep pace with the evolution of programs, widening the gap between what is measured and what programs actually do.

This report attempts to take modest steps toward rectifying the situation. Rather than recommend specific instruments, its aims are (1) to characterize recent

From Jeffrey R. Travers and Richard J. Light, "Evaluating Early Childhood Demonstration Programs," pp. 3-53 (selected pages) in *Learning from Experience: Evaluating Early Childhood Demonstration Programs*, edited by J. R. Travers and R. J. Light.

Editors' Note: Due to space limitations, we have reprinted only those parts of this chapter that focus on issues of general interest to evaluation researchers.

developments in programs and policies for children and families that challenge traditional approaches to evaluations and (2) to trace the implications for outcome measurement and for the broader conduct of evaluation studies. We have attempted to identify various types of information that evaluators of early childhood programs might collect, depending on their purposes. Our intent is not so much to prescribe how evaluation should be done as to provide a basis for intelligent choice of data to be collected.

Two related premises underlie much of our argument. First, policies and programs, at least those in the public domain, are shaped by many forces. Constituencies with conflicting interests influence policies or programs and in turn are affected by them. Policies and programs evolve continuously, in response to objective conditions and to the concerns of constituents. Demonstration programs, the subject of this report, are particularly likely to change as experience accumulates. Consequently, evaluation must address multiple concerns and must shift focus as programs mature or change character and as new policy issues emerge. Any single study is limited in its capacity to react to changes, but a single study is only a part of the larger evaluation process.

Second, the role of the evaluator is to contribute to public debate, to help make programs and policies more effective by informing the forensic process through which they are shaped. Though the evaluator might never actually engage in public discussion or make policy recommendations, he or she is nevertheless a participant in the policy formation process, a participant whose special role is to provide systematic information and to articulate value choices, rather than to plead the case for particular actions or values.

Note that we distinguish between informing the policy formation process and being co-opted by it--between research and advocacy. Research is characterized by systematic inquiry, concern with the reduction and control of bias, and commitment to addressing all the evidence. Nothing that we say is intended to relax the need for such rigor.

There are many views of the evaluator's role. Relevant discussions appear in numerous standard sources on evaluation methodology, such as Suchman (1967), Weiss (1972), Rossi et al. (1979), and Goodwin and Driscoll (1980). Some of these views are consonant, and some are partially contrasting with ours. For example, one widely held view

is that the role of the evaluator is, ideally, to provide definitive information to decision makers about the degree to which programs or policies are achieving their stated goals.[1] Though we agree that evaluation should inform decision makers (among others) and should strive for clear evidence on whether goals are being met, we argue that this view is insufficiently attuned to the pluralistic, dynamic process through which most programs and policies are formed and changed.

Sometimes the most valuable lesson to be learned from a demonstration is whether a particular intervention has achieved a specified end. Often, however, other lessons are equally or more important. An intervention can succeed for reasons that have little import for future programs or policies--for example, because of the efforts of uniquely talented staff. Conversely, a demonstration that fails, overall, may contain successful elements deserving replication in other contexts, and it may succeed in identifying practices that should be amended or avoided. Or a demonstration may shift its goals and "treatments" in response to local needs and resources, thereby failing to achieve its original ends but succeeding in other important respects.

By the same token, a randomized field experiment, with rigorous control of treatment and subject assignment, is sometimes the most appropriate way to answer questions salient for policy formation or program management. In such situations, government should be encouraged to provide the support necessary to implement experimental designs. There are situations, however, in which experimental rigor is impractical or premature, or in which information of a different character is likely to be more useful to policy makers and program managers. Preoccupation with prespecified goals and treatments can cause evaluators to overlook important changes in the aims and operations of programs as well as important

[1]Strictly speaking, this view applies only to "summative" evaluations, as distinguished from "formative" evaluations, which are intended to provide continuous feedback to program participants for the purpose of improving program operations.

outcomes that were not part of the original plan. If demonstrations have been allowed to adapt to local conditions, thoughtful documentation of the process of change can be far more useful in designing future programs than a report on whether original goals were met.

Even if change in goals and treatments is not at issue, understanding the mechanisms by which programs work or fail to work is likely to be more helpful than simply knowing whether they have achieved their stated goals. These mechanisms are often complex, and the evaluator's understanding of them often develops gradually. To elucidate mechanisms of change, it may be necessary to modify an initial experimental design, to perform post hoc analyses without benefit of experimental control, or to supplement quantitative data collection with qualitative accounts of program operations.

In short, we believe that evaluation is best conceived as a process of systematic learning from experience--the experience of the demonstration program itself and the experience of the evaluator as he or she gains increasing familiarity with the program. It is the systematic quality of evaluation that distinguishes it from advocacy or journalism. It is the need to bring experience to bear on practice that distinguishes evaluation from other forms of social scientific inquiry.

THE PROGRAM AND POLICY CONTEXT OF THE 1980s

Public policy both creates social change and responds to it. The evolution of policies toward children and families must be understood in the context of general societal change. Demographic shifts in the number of young children, the composition of families, and the labor force participation of mothers in recent years have increased and broadened the demand for services. They have also heightened consciousness about policy issues surrounding child health care, early education, and social services. Policy makers and evaluators in the 1980s are coping with the consequences of these broad changes. Contemporary policy issues and program

characteristics constitute the environment in which evaluators ply their trade, and they pose challenges with which new evaluations and outcome measures must deal.

To understand the policy context surrounding demonstration programs for children in the 1980s, it is useful to begin by outlining some general considerations that affect the formation of policy. These generic considerations apply to virtually all programs and public issues but shift in emphasis and importance as they are applied to particular programs and issues, at particular times, under particular conditions. The most fundamental consideration is whether the program or policy in question (whether newly proposed or a candidate for modification or termination) accords with the general philosophy of some group of policy makers and their constituents. Closely related is the question of tangible public support for a program or policy: Can the groups favoring a particular action translate their needs into effective political pressure?

Assuming that basic support exists, issues of access, equity, effectiveness, and efficiency arise. Will a program reach the target population(s) that it is intended to affect (access)? Will it provide benefits fairly, without favoring or denying any eligible target group--for example, by virtue of geographic location, ethnicity, or any other characteristics irrelevant to eligibility? And will its costs, financial and nonfinancial, be apportioned fairly (equity)? Will it achieve its intended objectives (effectiveness)? Will it do so without excessively cumbersome administrative machinery, and will cost-effectiveness and administrative requirements compare favorably with alternative programs or policies (efficiency)?

Two related concerns have to do with the unintended consequences of programs and policies and their interplay with existing policies and institutions. Will the policy or program have unanticipated positive or negative effects? Will it facilitate or impede the operations of existing policies, programs, or agencies? How will it affect the operations of private, formal, and informal institutions?

Programs for children and families are not exempt from any of these concerns. Some have loomed larger than others at times in the past two decades, and the current configuration is rather different from the one that prevailed when the first evaluations of compensatory education were initiated. The policy climate of the early

1960s was one of concern over poverty and inequality and of faith in the effectiveness of government-initiated social reform. The principal policy initiative of that period directed toward children and families--namely, the founding of Head Start--exemplified this concern and this faith. Head Start was initially administered by the now defunct Office of Economic Opportunity (OEO), and many local Head Start centers were affiliated with OEO-funded Community Action Programs. Thus, while it was in the first instance a service to children, Head Start was also part of the government's somewhat paradoxical attempt to stimulate grass roots political action "from the top down." The national managers made a conscious, concerted effort to distinguish Head Start from other children's services, notably day care. The latter was seen as controversial--hence, a politically risky ally.

The early 1960s was a time of economic and governmental expansion. Consequently, questions of cost and efficiency did not come to the fore. The principal concerns of the period were to extend services--to broaden access--and to demonstrate the effectiveness of the program. As noted earlier, effectiveness in the public mind was largely equated with cognitive gains. Despite the political character of the program, studies documenting its effectiveness as a focus for community organization and political action received little attention or weight--perhaps because the political activities of OEO-funded entities, such as the Community Action Programs and Legal Services, were sensitive issues even in the 1960s. Yet it was precisely the effectiveness of Head Start at mobilizing parents (together with the political skills of its national leaders) that saved the program when the Westinghouse-Ohio study produced bleak results and a new administration dismantled OEO.

During the 1970s the policy climate changed markedly. Economic slowdown and growing disillusionment with what were seen as excesses and failures of the policies of the 1960s brought about a concern for accountability and fiscal restraint, a concern that is still present and growing. Head Start responded by establishing national performance standards in an effort at quality control. Expansion was curtailed as the program fought to retain its budget in the face of inflation and congressional skepticism. (In fiscal 1977 only 15-18 percent of eligible children were actually served by Head Start.) Policy makers and program managers began to demand that

evaluations focus on management information and cost accounting.

At the same time, other policies and programs for children and families were gaining national attention. Economic pressures, the increased labor force participation of women, and the rise of feminism brought day care into prominence. Federal investment in day care increased under Title XX of the Social Security Act and numerous other federal programs for the working poor, backed by a curious alliance of feminists, liberals, child advocates, and "workfare conservatives." Although anti-day-care, "pro-family" forces remained strong, public subsidy of day care was gradually, if sometimes grudgingly, accepted as a reality. Most of the policy controversy surrounding day care in the 1970s centered on the trade-off of cost and quality: Should day care be viewed primarily as a service designed to free (or force) mothers to work--and therefore be funded at minimum levels consistent with children's physical and psychological safety? Or should it be viewed as a developmental service, akin to Head Start, or as a vehicle for delivering other services, such as health care and parent counseling, with attendant increases in cost? The controversy took concrete form in the debate over the Federal Interagency Day Care Requirements--purchasing standards that specify the type and quality of care on which federal dollars can legally be spent.

As we move into the 1980s, new, or more precisely latent, issues are likely to become prominent with respect to day care. The financing of day care is likely to become an ever more pressing problem, as the service becomes increasingly professionalized. Day care workers, among the nation's lowest paid, are likely to seek higher wages. Informal, low-cost care by friends or relatives may absorb less demand than it has in the past, as women who have heretofore provided such care either enter the work force in other capacities or begin to seek increased recognition and compensation for their services. At the same time, the importance of relatively informal care arrangements, such as family day care, have come to be recognized in policy circles. Informal arrangements are in fact the most prevalent forms of out-of-home care, especially for children of school age and for children under three. With this recognition will come new debates about the proper role of government: Should it regulate? Provide training? Invent new subsidy mechanisms? Major demonstrations examining alternative funding and regula-

tory policies for both center and family day care have already been undertaken by the state of California. Novel ways of funding child care, such as "tuition" vouchers, have been urged and studied, and a child care tax credit has already been legislated.

Day care is of course not the only type of children's program that underwent major change in the 1970s. Important new initiatives arose in the areas of child health and nutrition. For example, the Department of Agriculture established the Supplementary Food Program for Women, Infants, and Children and the Child Care Food Program; these provide low-cost nutritional supplements to low-income families and to the child care programs serving them. The Early and Periodic Screening, Diagnosis, and Treatment program was established to ensure that children from low-income families would be examined for problems of health, vision, hearing, etc.

Another initiative, sweeping in its implications, was the federal mandate under the Education for All Handicapped Children Act of 1975 (P.L. 94-142) that handicapped children be provided with a "free, appropriate public education," interpreted to mean education in the "least restrictive environment" feasible given their handicaps. The consequences for public schools have been enormous, and federal programs for younger children have also responded by building in provisions for the handicapped. The Head Start Economic Opportunity and Community Partnership Act of 1976 requires that 10 percent of Head Start slots in each state be set aside for handicapped children.

Although P.L. 94-142 is linked to federal funds to aid the handicapped, the law has the character of an entitlement rather than being a service program per se. The law establishes very broad rights and guidelines, not particular machinery for service delivery. Entitlements greatly broaden the constituencies affected by federal policy, for they extend far beyond the children of the poor. They highlight questions of access and equity for those charged with enforcement at the federal level. In the case of P.L. 94-142, questions of effectiveness and efficiency have largely been delegated to the local level: Local experts and practitioners are confronted with the task of devising programs that work at reasonable costs under local conditions. Questions having to do with overall effects of the policy on children, schools, and families have not been addressed at a national level. However, federal funds have been made available under

other legislative authorization for the establishment and evaluation of small-scale model programs for serving handicapped children.

Another major development with profound consequences for the schools is the bilingual education movement. The movement has been reinforced by the courts, most notably by the case of Lau v. Nichols, in which a California federal district court, later upheld by the U.S. Supreme Court, declared that it is discriminatory for schools to provide instruction only in English to students whose primary language is not English. Although the case was brought on behalf of Oriental children, its primary effects are being felt in those states where Hispanic children constitute a large and growing segment of the student population. And, like P.L. 94-142, the bilingual education movement has generally trickled down to the preschool level, where bilingual programs are rapidly being established in Head Start and other programs. The bilingual movement poses basic questions about federal and state policies toward minority subcultures--questions of pluralism versus integration that have never been fully addresssed. At the local level, these highly controversial issues are fueled with additional controversies over what are seen as federal rights of encroachment and the responsibilities of local governments.

Concurrent with these specific legislative and judicial initiatives, more diffuse but no less important policy issues have arisen in connection with certain federal demonstration programs. Two characteristics of these programs are particularly salient: an emphasis on the family and the community institutions with which it interacts, rather than on the child in isolation, and a stress on localism--on the diversity, rather than the uniformity, of programs and on their adaptation to local values and conditions. Programs exemplifying these emphases include Head Start's spinoff demonstrations, such as the Parent-Child Development Centers and the Child and Family Resource Program. These projects have acquired new strategic importance, in part as a result of a recent General Accounting Office report (General Accounting Office, 1979) that holds them up as models for future delivery of services to children from low-income families. Some nonfederal programs also emphasize multiservice support for families; an example is the Brookline Early Education Project, a privately funded program within a public school system. Other important

examples are day care programs funded under Title XX of the Social Security Act, which provides grants to states to purchase social services. These programs often provide a wide range of services that go beyond direct care of the child. And Title XX itself represents an attempt to decentralize decision making by allowing states considerable latitude in the use of federal funds.

These policy emphases have multiple roots. In part they stem from a reaction against what has been seen as an intrusive, excessively prescriptive federal posture vis-a-vis local programs and their clients. In part they represent an assertion of the family's central role and responsibility in child rearing. In part they have a theoretical base and reflect an ecological perspective on child development--one that sees changes in the child's immediate social milieu, the family, and family-community relations as the best way to create and sustain change in individual children. In part they arise from practical experience with and applied research on earlier programs, which repeatedly showed dramatic differences in practices and effects from site to site, even when they were allegedly committed to implementing some prescribed treatment or model.

Family support programs raise issues that have not been prominent with respect to earlier demonstrations. They focus attention on the relationships between children's programs and other service agencies in local communities. They also focus attention on relations between programs and informal institutions, such as extended families, which in some subcultures have traditionally provided the kind of global support that some demonstration programs aim to provide. They raise basic questions as to whether ecological approaches in general are more effective than interventions aimed at the child alone. Finally, they highlight issues having to do with the prerogatives and responsibilities of different levels of government and of government vis-a-vis private program sponsors, service providers, and clients. A tension is created by pressures for accountability at the federal level and conflicting pressures for delegation of responsibility to the state or local level. Evaluation often plays a role in struggles among the various levels of government, usually as a device by which federal program managers attempt to exert some control over local practices.

In short, the policy context surrounding early child-

hood demonstration programs in 1980 has become very complex. Old issues have remained, and new or resurgent issues have been overlaid on them. The need to measure program effects on children has not diminished--witness the current effort by Head Start to develop a new, comprehensive battery of outcome measures. Concerns about cost, efficiency, and equity have become acute, as the federal government has expanded the scope of its responsibilities. Broad entitlements and new initiatives have increased the competition for finite resources in the face of widespread resistance to further taxation and bureaucratic expansion. There is increased pressure for centralized accountability and cost and quality control. At the same time there has been a broadening of the constituencies affected by early childhood programs as well as increased emphasis on pluralism of goals and values; decentralized, local decision making; and the individualization of services. Fortunately, no single evaluation will ever have to address all of these policy concerns simultaneously. Nevertheless, their complexity and antithetical value premises pose staggering challenges for the evaluator who hopes to influence policy. Although evaluators can address only a small subset of these concerns, they must constantly be aware of the larger picture or run the risk that the information they provide will be irrelevant or misleading in light of the full configuration of issues bearing on the future of a particular program.

These last observations lead to a final point about the policy climate of the 1980s: the role of evaluation itself in policy determination. An evaluation industry was born with the Great Society programs of the 1960s, which often included evaluations as integral parts. That enterprise has continued to grow and its audience has expanded, as clients, advocacy groups, and practitioners as well as policy makers and social scientists have learned to use evaluation results for their own diverse purposes. Congress has explicitly written evaluation requirements into the authorizing legislation for major programs, such as Title I of the Elementary and Secondary Education Act and the Education for All Handicapped Children Act.

As evaluation has grown in prevalence and importance, some of its limitations have also become apparent. By their very nature, evaluative studies must be restricted in scope and therefore can address broad policy issues only in a partial and fragmentary fashion. The injection

of rational, systematic, analytic perspective into policy formation does not dispense with value conflicts; the choice of questions in evaluations is partly a matter of values, and findings are always subject to interpretation from multiple perspectives. Evaluation itself has costs, not only financial but also in terms of respondent burden and potential invasion of privacy. There are concrete manifestations of resistance to evaluation, in the form of increased restrictions on data collection.

Despite these limitations we believe that evaluation can contribute to policy. Particular findings may mesh with the immediate information needs of policy makers and thus affect decisions directly. Boruch and Cordray (1980) provide some striking case studies illustrating this sort of direct contribution. Perhaps more typically, findings from many studies over time can create a general climate of belief, for example, belief that early intervention in some sense "works," which in turn subtly and gradually shapes the questions that policy makers ask, shifting their attention, for example, from questions of effectiveness to questions of access, equity, and efficiency. Evaluation can also reveal unintended consequences of programs and point to new policy questions and new directions for program development. Sophistication about the multiple concerns of policy makers and their own limited roles in the process of policy determination may breed in evaluators a salutary humility, but it should not breed despair. And awareness should make their contribution even greater.

IMPLICATIONS FOR OUTCOME MEASUREMENT AND EVALUATION DESIGN

The programs and policy issues that have evolved over the past two decades, particularly in the late 1970s, pose serious challenges for evaluators. However, experience in performing evaluative studies has been accumulating since the early 1960s, and that experience offers contemporary evaluators some lessons about how to deal with at least some of these challenges. In this section we discuss specific characteristics of contemporary programs for young children that confront evaluators with problems of design and measurement and lessons drawn from past experience that may help improve future evaluations.

Challenges to the Evaluator

Many of our concepts of outcome measurement and evaluation design were, as already suggested, shaped by the compensatory education and cognitive enrichment programs of the early 1960s. These programs were initiated under private auspices, often with government funding, at one or a few sites. While these programs were to become models for public policy and in many cases were consciously intended as such, they were not immediately concerned with issues of administration and implementation on a large scale or with links to other public service delivery systems, such as nutrition or health care. Nor were they much concerned with questions of cost or cost-effectiveness. The question on everyone's mind was, will preschool education work? That is, will it improve the school functioning and test scores of low-income children?

The early programs were new and relatively small, their goals were relatively clear and circumscribed, and comparable services were not widely available. The individual child was typically the recipient of treatment, and the programs were implicitly conceived as operating in relative isolation from other social institutions and forces. Consequently, it was possible to devise simple evaluations, in which test scores and school performance of children in the program were compared with those of similar children in the same communities who received no services. The program itself was viewed as a unitary "treatment," and children in the control or comparison group were assumed to receive no treatment. Such evaluation designs were straightforward extensions of laboratory paradigms, although the children in control groups were often selected by post hoc matching rather than random assignment, thus making many evaluations designs quasi-experiments rather than true experiments. Of course, not all early programs were rigorously evaluated, and not all evaluations were as limited as we have suggested; for example, diffusion of effects to siblings and neighbors was a topic of interest in some of the early evaluation studies.

As suggested earlier, experimental designs are ideal for answering certain kinds of evaluation questions, because they provide the most direct means of establishing linkages of cause and effect. Children's academic skills and performance are often important program outcomes, and standardized tests, properly interpreted, measure aspects

of these skills. However, experience with the demonstrations that have evolved over the past two decades has made three points clear: First, a wider range of outcome measurement is necessary to do justice to program goals. Second, measurement of outcomes alone does not show why a program achieved or failed to achieve its intended goal--often the most significant lesson to be learned from a demonstration. Third, the conditions necessary for successful experimentation are often not met when demonstrations are conducted on a relatively large scale. Treatments tend to be multifaceted and variable. Often the pairing of client and treatment is beyond the experimenter's control. Extremely complex designs may be needed to tease out complex chains of causation.

We amplify these points in the pages that follow. It should be clear, however, that we are not opposed to experimental approaches, controlled assignment, or formal designs. We discuss program characteristics that pose barriers to formal experimentation in order to make a case for supplementing, not supplanting, experimental approaches with other scientifically defensible forms of investigation. Similarly, we recognize the value of outcome measures focused on individual development, including academic skills and achievement. However, we emphasize program characteristics that point to the need for other kinds of data--measures of outcomes that go beyond the individual child and measures of context and process that illuminate why and how a program works or fails to work.

IMPLICATIONS FOR THE EVALUATION PROCESS

Some of our suggestions about design and measurement have indirect implications for the way in which applied research is organized and conducted, for the way in which its results may be presented most effectively, and even for the relationship between applied research and basic social science.

Involving Multiple Constituencies in Selecting Outcome Measures

Given that demonstration programs affect many constituencies that have a stake or a say in the program's future, ways must be found to involve these groups or at least take account of their concerns in selecting outcome measures. Actual involvement is preferable, because it creates a commitment to the evaluation process, which may not otherwise be present on the part of some constitutent

groups, even if the outcome measures used in an evaluation are relevant to their concerns.

To say that constituents should somehow be involved in identifying salient concerns or potential program outcomes of course does not mean that the outcomes can or should be selected on the basis of a survey. Constituencies differ in the salience that they accord to different outcomes. In some cases, outcomes valued by different constituencies may conflict. For example, when parents of handicapped children exercise their rights to change their children's educational placement, there is no guarantee that the educational experiences of the child will in fact be improved, either by the lengthy process of appeals that may be involved or by the ultimate outcome. In such a situation, legitimate values compete: Is it more important for parents to have such rights or for children to have steady, uninterrupted, and relaxed educational experiences? Such conflicts create delicate situations in which evaluators, sponsors of evaluations, practitioners, and clients must negotiate the choice and weighting of outcomes. Our point is that the scope of an evaluation, the breadth of the audience for which it provides at least some relevant information, and the likelihood that its findings will be put to use will all be enhanced if the perspectives of the various constituencies are considered.

Communicating with Multiple Audiences

We have argued consistently that if evaluation is to accomplish its goal of helping to improve programs and shape policies, it must be attuned to practical issues, not only to the interests of discipline-based researchers and methodologists. Beyond this first and most important step, evaluators can, by virtue of the way in which they present their work, take further measures to ensure the dissemination and utilization of their results.

Basic researchers are usually trained to speak only to other researchers. Buttressed with statistics and hedged with caveats, their reports typically have a logic and an organization aimed at persuading professional critics of the accuracy of careful delimited empirical claims. However, applied researchers must address many audiences who make very different uses of their findings. Policy makers, government program managers, advocacy groups, practitioners, and parents are among their many audiences. Each group has its own concerns and requires a special form of communication. However, all these groups have

some common needs and aims, quite different from those of the research audience. They all want information to guide action, rather than information for its own sake. They have limited interest and sophistication with respect to research methods and statistics.

This situation poses practical and ethical problems for the evaluator. The practical problem is simply that of finding ways to communicate findings clearly, with a minimum of jargon and technical detail. One strategy that has proved effective in this regard is organizing presentations around the questions of concern to non-technical audiences, rather than around the researcher's data-collection procedures and analyses. Adoption of this strategy of course presumes that the research itself has been designed at least in part to answer the questions of policy makers and practitioners. In addition, the impact of a report, however well written, can be enhanced by adroit management of other aspects of the dissemination process--public presentations, informal discussions with members of the intended audience, and the like--which can help create a climate of realistic advance expectations and appropriate after-the-fact interpretation.

The ethical problem is that of drawing the line between necessary qualification and unnecessary detail. One can always write a report with a clear message by ignoring inconsistent data and problematic analyses. The difficulty is to maintain scientific integrity without burying the message in methodological complexities and caveats. There is no general formula for solving this problem, any more than there is a formula for writing accurately and forcefully. It is important, however, that the problem be recognized--that researchers do not allow themselves to fall back on comfortable obscurantism or to strain for publicity and effect at the price of scientific honesty.

Building in Familiarity and Flexibility

The considerations about design and measurement discussed above have practical implications for the way in which applied research is conducted. One implication is that both researchers and the people who manage applied research--particularly government project officers and perhaps even program officers in foundations--need to develop intimate familiarity with the operations of service programs as well as basic understanding of the policy context surrounding those programs. Technical virtuosity and substantive excellence in an academic

discipline do not alone make an effective evaluator. Over and above these kinds of knowledge, a practical, experiential awareness of program realities and policy concerns is essential if evaluation is to deal with those realities and to address those concerns. When third-party evaluations are conducted by organizations other than the service program or its funding agency, a preliminary period of familiarization may be needed by the outside evaluator. Moreover, that individual or organization should remain in close enough touch with the service program throughout the evaluation to respond to changes in focus, clientele, or program practices.

A second, related implication is that the evaluation process must be flexible enough to accommodate the evolution of programs and the researcher's understanding. Premature commitment to a particular design or set of measures may leave an evaluation with insufficient resources to respond to important changes, ultimately resulting in a report that speaks only to a program's past and not to its future. Such a report fails disastrously in meeting what we see as the primary responsibility of the evaluator, namely to teach the public and the policy maker whatever there is to learn from the program's experience.

There is danger, too, in the evaluator's being familiar with programs and flexible in responding to program changes as we have advocated. Too much intimacy with a program can erode an evaluator's intellectual independence, which is often threatened in any case by his or her financial dependence on the agency sponsoring the program in question. (Most evaluations are funded and monitored by federal mission agencies or private sponsors that also operate demonstration programs themselves.) We see no easy solution to this serious dilemma, but at the same time we can point to mechanisms that limit any distortions introduced by too close a relationship between evaluator and program. Most important among them are the canons of science, which require that the evaluator collect, analyze, and present data in a way that opens the conclusions to scrutiny. The political process can also act as a corrective force, in that it exposes the evaluator's conclusions to criticism from many value perspectives. Finally, as some researchers have urged, it may sometimes be feasible to deal with advocacy in evaluation by establishing concurrent evaluations of the same program, perhaps funded by separate agencies, but in any case deliberately designed to reflect divergent values and presuppositions.

This report does not discuss in detail the institutional arrangements that might lead to more effective program evaluations nor does it examine current arrangements critically. Such an examination would be a major report in itself. Relevant reports have been written under the aegis of the National Research Council, e.g., Raizen and Rossi (1981). However, we observe that many major evaluations are funded by the federal government through contracts with universities or private research organizations. The contracting process is rather tightly controlled. Subject to the approval of the funding agency, the contractor is typically required to choose designs, variables, and measures early in the course of the study, then stick to them. It is rare that contractors are given adequate time to assimilate preliminary information or to develop and pretest study designs and methods. Sometimes the overall evaluation process is segmented into separate contracts for design, data collection, statistical analysis, and policy analysis. It is perfectly understandable that the government is reluctant to give universities or contract research organizations carte blanche, especially in large evaluations, which may cost millions of dollars. Even the fragmentation of evaluation efforts may be partially justifiable, on the grounds that it allows the government to purchase the services of organizations with complementary, specialied expertise. Whatever the merits of these policies, it seems clear that in some respects the contracting process is at odds with the needs we have identified for gradual accretion of practical understanding and for flexibility in adapting designs and measures to changes in programs.

Drawing on and Contributing to Basic Social Science

In some respects, evaluation stands in the same relationship to traditional social science disciplines as do engineering, medicine, and other applied fields to the physical and biological sciences. Evaluation draws on the theories, findings, and methods of anthropology, economics, history, political science, psychology, sociology, statistics, and kindred basic research fields. At the same time, evaluation "technology" can also contribute to basic knowledge. The approach to the evaluation of children's programs set forth in this report has implications both for the kinds of basic social science that are likely to give rise to the most useful applications and for the kinds of contributions that evaluation can make to fundamental research.

Traditionally, evaluation has borrowed most heavily from basic research fields that emphasize formal designs and quantitative analytic techniques--statistics, economics, experimental psychology, survey research in sociology, and political science. The approach to evaluation we suggest implies that quantitative techniques can usefully be supplemented--not supplanted--by ethnographic, historical, and clinical techniques. These qualitative approaches are well suited to formulating hypotheses about orderly patterns underlying complex, multidetermined, constantly changing phenomena, although not to rigorous establishment of causal chains. There is nothing scientific about adherence to forms and techniques that have proved their usefulness elsewhere but fail to fit the phenomena at hand. Science instead adapts and develops techniques to fit natural and social phenomena. When a field is at an early stage of development, available techniques are likely to have severe limitations. But the use of all the techniques available, with candid admission of their limitations, is preferable to Procrustean distortion of phenomena to fit preferred methods in pursuit of spurious rigor.

Our proposed approach also suggests that global, systemic approaches to theory, of which the ecological approach to human development is an example, are potentially useful. Ad hoc empirical "theories" that specify relationships among small numbers of variables, whatever their merits in terms of clarity and precision, simply omit too much. Theories that explicate relationships among variables describing individual growth, family dynamics, and ties between families and other institutions have greater heuristic value, even if they are too ambitious to be precise at this early stage in their development.

It should be clear that we favor precision, rigor, and quantitative techniques. Each has its place, even given the present state of the evaluation art, and that place is likely to become larger and more secure as the art advances. We argue, however, that description and qualitative understanding of social programs are in themselves worthwhile aims of evaluation and are essential to the development of useful formal approaches.

We have indicated some of the directions in which we think evaluation technology is likely to lead social science. Because understanding social programs requires a judicious fusion of qualitative and quantitative methods, evaluation may stimulate new methodological work

articulating the two approaches. We may, for example, learn better ways to bring together clinical and experimental studies of individual children or ethnographic and survey-based studies of the family. Because understanding programs requires an appreciation of interlocking social systems, evaluation may contribute to the expansion and refinement of ecological, systemic theories. Thinking about children's programs may lead to a deeper understanding of the ways in which individual development is shaped by social systems of which the child is a part. Finally, because programs are complex phenomena that cannot be fully comprehended within the intellectual boundaries of a single discipline, evaluation may open up fruitful areas of interdisciplinary cooperation.

We are well aware that science often proceeds analytically rather than holistically; for example, it is useful for some purposes to isolate the circulatory system as an object of study, even though it is intimately linked to many other bodily systems. Nevertheless it is also useful now and then to examine interrelationships among previously defined systems to see if new insights and new areas of study--new systems--emerge. It is our hope that evaluation research can play this role vis-a-vis the social sciences. By focusing on concrete, real-world phenomena that do not fit neatly into existing theoretical or methodological boxes, evaluation may stimulate the development of both theory and method.

REFERENCES

Ainsworth, M. D. S., and Wittig, B. A.
(1969) Attachment and exploratory behavior of one-year-olds in a strange situation. In B. M. Foss, ed., Determinants of Infant Behavior, Volume 4. London: Methuen.

Anderson, S., and Messick, S.
(1974) Social competency in young children. Developmental Psychology 10:282-293.

Belsky, J.
(1980) Child maltreatment: an ecological integration. American Psychologist 35(4):320-335.

Belsky, J., and Steinberg, L. D.
(1978) The effects of day care: a critical review. Child Development 49:929-949.

Boruch, R. F., and Cordray, D. S.
(1980) An Appraisal of Educational Program Evaluations: Federal, State and Local Agencies. Report prepared for the U.S. Department of Education, Contract No. 300-79-0467. Northwestern University (June 30).

Brim, O. G.
(1959) Education for Child Rearing. New York: Russell Sage Foundation.

Bronfenbrenner, U.
(1974) A Report on Longitudinal Evaluations of Preschool Programs. Vol. II: Is Early Intervention Effective? U.S. Department of Health, Education, and Welfare, Publication No. OHD 75-25. Washington, D.C.: U.S. Department of Health, Education, and Welfare.
(1979) The Ecology of Human Development. Cambridge, Mass.: Harvard University Press.

Bureau of Education for the Handicapped
(1979) Progress Toward a Free, Appropriate Public Education. A Report to Congress on the Implementation of Public Law 94-142: The Education for All Handicapped Children Act. HEW Publication No. (OE) 79-05003. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Connell, D. C., and Carew, J. V.
(1980) Infant Activities in Low-Income Homes: Impact of Family-Focused Intervention. International Conference on Infant Studies, New Haven, Conn. (April).

Datta, L. E.
(1979) Another spring and other hopes: some findings from National Evaluations of Project Head Start. In E. Zigler and J. Valentine, eds., Project Head Start: A Legacy of the War on Poverty. New York: Free Press.

Farran, D., and Ramey, C.
(1980) Social class differences in dyadic involvement during infancy. Child Development 51:254-257.

General Accounting Office
(1979) Early Childhood and Family Development Programs Improve the Quality of Life for Low-Income Families. Report to the Congress by the Comptroller General. HR-79-40 (February).

Goodson, B. D., and Hess, R. D.
(1978) The effects of parent training programs on child performance and parent behavior. In B. Brown, ed., Found: Long-Term Gains from Early Education. Boulder, Colo.: Westview Press.

Goodwin, W. L., and Driscoll, L. A.
(1980) Handbook for Measurement and Evaluation in Early Childhood Education. San Francisco, Calif.: Jossey-Bass, Inc., Publishers.

Horowitz, F. D., and Paden, L. Y.
(1973) The effectiveness of environmental programs. In B. Caldwell and H. D. Ricciuti, eds., Review of Child Development Research. Vol. 3: Child Development and Social Policy. Chicago, Ill.: University of Chicago Press.

Johnson, O. G.
(1976) Tests and Measurements in Child Development: Handbook II. Vols. 1 and 2. San Francisco, Calif.: Jossey-Bass, Inc., Publishers.

Johnson, O. G., and Bommarito, J. W.
(1971) Tests and Measurements in Child Development: A Handbook. San Francisco, Calif.: Jossey-Bass, Inc., Publishers.

Kirschner Associates, Albuquerque, N.M.
(1970) A National Survey of the Impacts of Head Start Centers on Community Institutions. (ED045195) Washington, D.C.: Office of Economic Opportunity.

Lazar, I., and Darlington, R. B.
(1978) Lasting Effects After Preschool. A report of the Consortium for Longitudinal Studies. U.S. Department of Health, Education, and Welfare, Office of Human Development Services, Administration for Children, Youth, and Families.

Lindsey, W. E.
(1976) Instrumentation of OCD Research Projects on the Family. Mimeographed report prepared under contract HEW-105-76-1120, U.S. Department of Health, Education, and Welfare. Social Research Group, The George Washington University, Washington, D.C.

Love, J. M., Nauta, M. J., Coelen, C. G., and Ruopp, R. R.
(1975) Home Start Evaluation Study: Executive Summary--Findings and Recommendations. Ypsilanti, Mich., and Cambridge, Mass.: High/Scope Educational Research Foundation and Abt Associates, Inc.

Raizen, S. A., and Rossi, P. H., eds.
(1981) Program Evaluation in Education: When? How? To What Ends? Committee on Program Evaluation in Education, Assembly of Behavioral and Social Sciences, National Research Council. Washington, D.C.: National Academy Press.

Ramey, C., and Mills, J.
(1975) Mother-Infant Interaction Patterns as a Function of Rearing Conditions. Paper presented at the biennial meeting of the Society for Research in Child Development, Denver, Colo. (April).

Rossi, P. H., Freeman, H. E., and Wright, S. R.
(1979) Evaluation: A Systematic Approach. Beverly Hills, Calif.: Sage Publications.

Ruopp, R., Travers, J., Coelen, C., and Glantz, F.
(1979) Children at the Center. Final report of the National Day Care Study, Volume I. Cambridge, Mass.: Abt Books.

Smith, M. S., and Bissell, J. S.
(1970) Report analysis: the impact of Head Start. Harvard Educational Review 40:51-104.

Sroufe, L. A.
(1979) The coherence of individual development: early care, attachment and subsequent developmental issues. American Psychologist 34:834-841.

Stallings, J.
(1975) Implementation and child effects of teaching practices in Follow Through classrooms. Monographs of the Society for Research in Child Development 40(7-8), Serial No. 163.

Stebbins, L. B., et al.
(1977) Education as Experimentation: A Planned Variation Model. Vol. IV. Cambridge, Mass.: Abt Associates, Inc. Also issued by the U.S. Office of Education as National Evaluation: Patterns of Effects. Vol. II of the Follow Through Planned Variation Series.

Suchman, E. A.
(1967) Evaluation Research: Principles and Practice in Public Service and Social Action Programs. New York: Russell Sage Foundation.

Walker, D. K.
(1973) Socioemotional Measures for Preschool and Kindergarten Children. San Francisco, Calif.: Jossey-Bass, Inc., Publishers.

Weber, C. U., Foster, P. S., and Weikart, D. P.
(1977) An economic analysis of the Ypsilanti Perry Preschool Project. Monographs of the High/Scope Educational Research Foundation. Series No. 5.

Weiss, C. H.
(1972) Evaluating Action Programs: Readings in Social Action and Education. Boston, Mass.: Allyn & Bacon, Inc.

Westinghouse Learning Corporation and Ohio University
(1969) The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development. Executive Summary. Report to the Office of Economic Opportunity (ED036321). Washington, D.C.: Clearinghouse for Federal Scientific and Technical Information.

Zigler, E., and Trickett, P.
(1978) IQ, social competence and evaluation of early childhood intervention programs. American Psychologist 33:789-798.

Zigler, E., and Valentine, J., eds.
(1979) Project Head Start: A Legacy of the War on Poverty. New York: The Free Press.

6

Congressional Input to Program Evaluation

Scope and Effects

Robert G. St. Pierre

The U.S. Congress is a major funder and user of program evaluation. Through studies performed by federal administrative agencies or by the General Accounting Office under direct congressional mandate, and through studies performed by federal, state, or local administrative agencies under general congressional authorization, much of the program evaluation done in the United States is based on congressional requests for information. In many cases Congress's call for evaluation is broad which reflects a general concern for accountability rather than a specific informational need. For example, the General

From Robert G. St.Pierre, "Congressional Input to Program Evaluation," *Evaluation Review*, 1983 7(4), 411-436.

Education Provisions Act applies to all federally funded education programs (Section 1221b) and specifies that state and local applicants for federal education funds must include an evaluation component in order to "determine the effectiveness of covered programs in meeting their statutory objectives" (Section 1232d). These provisions are stated in general terms, and while they "establish the cornerstone of federal policy on education evaluations, the influence of these requirements remains unclear" (Boruch and Cordray, 1980: 2-18).

Stating evaluation requirements in broad terms is in keeping with Congress's approach to programmatic legislation that typically involves a rather general mandate for a program to ameliorate a given problem. The appropriate federal administrative agency is then charged with the task of preparing regulations in order to implement the program. This approach is necessitated by the political compromises involved in passing legislation (a general program that is applicable to a wide range of constituencies is more politically viable than a program targeted to a small subpopulation) as well as by the great amount of work involved in preparing regulations for program implementation. Congress deals with national problems in which "policy is blocked out with broad brushstrokes, and operational planning is left to lower levels" (Cronbach et al., 1980: 102). The "broad brush" approach has also been a major tool in creating legislation evaluation.

In other cases Congress has quite targeted information needs regarding particular programs and prepares legislation mandating that federal agencies addres those needs. For example, as part of the legislation that creates or reauthorizes programs, Congress often includes a specific call for evaluations to be conducted for program oversight. In the recent past, major congressionally mandated studies in education have concerned compensatory education, vocational education, bilingual education, special education, and school finance, as well as other areas.

In spite of the increasing number of congressional mandates for evaluation, discontent has been expressed about the impact of evaluation on policy both by those involved in the legislative process and by evaluators. Many evaluators believe that their work has had insufficient impact on policy, and a literature has arisen on the topic (Florio et al., 1979; Leviton and Hughes, 1981; Patton, 1978). Those in the legislative arena complain that evaluators do not present information in clear, concise, and understandable form; do not provide information in a

timely fashion; do not provide unequivocal or concrete answers; do not work directly enough with congressional staff; and do not address the issues faced by Congress (Florio et al., 1979). The point is not that evaluations are of little intrinsic value, but that the process of getting evaluations used has been neglected by evaluators. Furthermore, Florio et al. argue that it is up to the evaluators to remedy the situation by justifying the utility of evaluation research to congressional users.

One way that evaluators have tried to solve this problem is to encourage legislators to be clear about what is wanted from an evaluation as the legislation calling for the evaluation is being prepared. In a comprehensive study of educational program evaluations, Boruch and Cordray (1980) address the complaint that evaluators do not deal with the issues faced by Congress, that information is not targeted to congressional needs, and recommend that "evaluation statutes identify the specific questions which need to be addressed and specific audiences for results" and that "higher quality research designs, especially randomized experiments, be authorized explicitly in law for testing new programs." In a companion study, Raizen and Rossi (1981) also call for greater specificity in congressional requests for information and note that "a call for evaluation that does not specify what questions are being asked can lead to the mismatching of expectations and performance by Congress and the evaluators" (p. 55). A similar message is relayed by Levine who holds that "the 'ideal legislation' will specify the scope of the evaluation activity, the questions to be addressed, and the procedure for reporting these results to Congress, and to the responsible program and oversight groups" (p. 19).

Thus, evaluators are responding to complaints about the utility of evaluation in part by asking Congress for more direction—by saying that more targeted information will result from more targeted questions. Still, some of the same evaluators are worried that Congress will go too far in terms of specifying the nature of evaluation studies. On the same page as their call for greater specificity of evaluation questions, Raizen and Rossi note that "though we recommend that it be specific with respect to question and audience, legislative language regarding evaluation should refrain from specifying details of method (such as sampling procedure or use of control groups) or of measurement" (p. 55). So, the clear call for input on evaluation questions and relevant audiences also raises a concern that such requests for direction may

result in overspecification, such as Congress giving directions in methodological areas.

This article considers the advantage and disadvantages of having evaluation users increase the specificity of evaluation requirements through a review of the legislative requirements behind several congressionally mandated national-level evaluations, and an in-depth examination of one congressionally mandated evaluation in which very specific design parameters were included in the authorizing legislation. The evaluations that form the basis of this article are drawn from education and agriculture—two areas that have received substantial congressional attention in recent years and that have been the subject of many evaluations. It is likely that the conclusions reached here apply to many (although perhaps not all) federal programs, and especially to those that have received relatively little evaluation. The conclusion is that specific congressional input in the areas of evaluation questions, audiences, and timeliness is warranted and important, while explicitness in areas of research design is unnecessarily restrictive and can diminish the quality of evaluations.

SOME RECENT CONGRESSIONALLY MANDATED EVALUATIONS

The legislation that accompanies the creation or reauthorization of social programs often calls for studies of some sort to be conducted in preparation for the next reauthorization. The directions given by Congress to federal agencies in charge of implementing the evaluations (and hence to contracted evaluators) come in two forms: (1) the initial legislation that includes a mandate for research or evaluation; and (2) subsequent refinement of the mandate through discussions between research staff in federal agencies and responsible congressional committee staffers. This article concentrates on the scope and effects of the first type of congressional input to evaluation: the impact of the written legislative mandate and related materials from sources such as committee reports and the *Congressional Record.* Though many of the examples used here draw on educational evaluations, some examples outside education are included to show that the conclusions drawn apply to evaluations of human service delivery programs in general.

The education program that has received the most attention in terms of congressionally mandated evaluation efforts is Title I of the Elementary and Secondary Education Act of 1965 (now called Chapter I of the Educational Consolidation and Improvement Act of 1981). Title I is the cornerstone of federal aid to elementary education and was the first federal education act to require annual evaluations at the local level. By requiring local evaluations, Congress has created pressure on itself to make periodic national evaluations. For example, as part of Public Law 93-380 (Title I, Part B, Section 821), in 1974 Congress requested that the National Institute of Education (NIE) conduct a comprehensive study of compensatory education programs. A somewhat abstracted version of the relevant legislation follows.

(a) The National Institute of Education shall undertake a thorough evaluation and study of compensatory education programs including . . .
  (1) examination of the fundamental purposes of programs and the effectiveness of programs in attaining such purposes;
  (2) analysis of means to identify accurately the children who have the greatest need for programs;
  (3) analysis of the effectiveness of methods for meeting the educational needs of children, including the use of individualized written educational plans for children and programs for training the teachers of children;
  (4) exploration of alternative methods, including the use of procedures to assess education disadvantage, for distributing funds to State and local educational agencies. . .;
  (5) not more than 20 experimental programs, which shall be reasonably geographically representative . . .; and
  (6) findings and recommendations, including recommendations for changes in such title I or for new legislation.
(b) The National Advisory Council on the Education of Disadvantaged Children shall advise the Institute with respect to the design and execution of such study.
(c) . . . interim report to the President and to the Congress not later than December 31, 1976, and . . . final report nine months after the . . . interim report. . . . Such reports shall not be submitted to any review outside of the Institute before their transmittal to the Congress.
(d) Sums made available pursuant to . . . the Elementary and Secondary Education Act of 1965 shall be available to carry out the provisions of this section.
(e) (1) The Institute shall submit to the Congress, within one hundred and twenty days after the date of the enactment of this Act, a plan for its study to be conducted under this section.

In 1978, the NIE responded to this mandate with a coordinated effort consisting of more than 35 research projects which addressed a great

range of issues and cost some $15 million. As can be seen from the legislation, some parts of the congressional mandate were quite broad, calling for an examination of the fundamental purposes of compensatory education. However, according to Brown et al., the provision of issue-specific information was of much greater interest to Congress than research on more global processes: "Congress was interested in the likely effects of making marginal changes in its use of familiar policy instruments; it did not need speculative research on more fundamental changes" (1979: 11). For example, as part of the mandate Congress requested specific information on the effects of proposed changes in program legislation by asking NIE to investigate Representative Albert Quie's proposal to allocate Title I funds to school districts based on numbers of low-achieving rather than low-income children.

In a review of the process used by NIE in conducting the study, Hill identified several technical and tactical problems, one of which was "how to move from the broad research objectives set by Congress to specific statements of researchable problems" (1980: 59). Given the nonspecific nature of most of the congressional mandate, Hill and his colleagues spent a great deal of time across a two-year period building a research strategy. This included reaching agreement with Congress on a proper response to the requirements of the mandate, asking Congress to identify topics that the mandate had omitted, and proposing additional areas of research to Congress. In short, a large amount of effort was required to specify the questions that were of most importance to Congress. A good part of the success of NIE's study can be attributed to this up-front work.

As part of the Education Amendments of 1978 (Public Law 95-561, Section 102) Congress called for a study of Title I that differed in many key respects from the previously discussed Title I evaluation. In the present case, Congress was concerned about complaints that Title I regulations dealing with the comparability of Title I and non-Title I services were preventing schools from making the best use of Title I funds. In response, Congress mandated a study of the effects of alternatives to the comparability provision that would provide greater flexibility to schools. The study mandate is abstracted below.

(a) The Commissioner shall, not later than September 30, 1981, make a study of the feasibility and desirability of alternative criteria for demonstrating the comparability of services . . . in each project area . . . to those provided outside such areas. . . .

(b) The Commissioner may select all the local educational agencies in one State and not more than twenty such agencies in other States which are reasonably representative of the various geographical areas of the Nation for participation in the study. . . .

(c) Local educational agencies selected for participation in the study provided for in this section shall demonstrate comparability through the use of alternative criteria, which, at a minimum, meet the conditions of the following paragraphs. . . .
(d) In order to provide a basis for comparison, local educational agencies participating in the study under this section shall continue to make reports under existing criteria for comparability of services.

The congressional call for a study of the comparability provision is much more specific than the legislation cited earlier. The mandate goes so far as to prescribe the number of school districts that should be involved in the study as well as the distribution of the sample. Furthermore, the legislation contains direct language (not reproduced here) as to the conditions that must be met by schools participating under alternative criteria for demonstrating comparability. As will be apparent from subsequent examples, this level of congressional input occurs when a genuine political difference is being negotiated. Due to the high degree of prescription, the evaluation report (Ellman et al., 1981) noted two major limitations imposed on the study by the congressional mandate. First, Congress narrowed the nature and scope of the alternatives in such a way that only alternatives very similar to the existing provision were eligible. Second, as a result of the mandated sampling plan, one of the alternatives was overrepresented in the sample.

In Public Law 95-561 (Part C, Section 742) Congress called for the Office of Bilingual Education and the NIE to establish a program of research in bilingual education. As was the case with the mandated Title I studies, some of the language in the abstracted legislation is quite broad and some is fairly targeted.

(a) (1) Through competitive contracts provide financial assistance for research and development proposals.
(2) Carry out a program of research in bilingual education in order to enhance the effectiveness of bilingual education.
(3) Coordinate research activities of . . . appropriate agencies in order to develop a national research program for bilingual education.
(b) Authorized research activities include:
(1) studies to determine and evaluate effective models for bilingual-bicultural programs;
(2) studies to determine (A) language acquisition characteristics and (B) the most effective method of teaching English within the context of a bilingual-bicultural program to students who have language proficiencies other than English;
(3) A five-year longitudinal study to measure the effect of this title on the education of students who have language proficiencies other than English;
(4) studies to determine . . . methods of identification of students who should be entitled to services;

(5) the operation of a clearinghouse on information for bilingual education;
(6) studies to determine the most effective methods of teaching reading to children and adults who have language proficiencies other than English;
(7) studies to determine the effectiveness of teacher training preservice and inservice programs;
(8) studies to determine the critical cultural characteristics of selected groups of individuals.

(c) Provide for periodic consultation with representatives of State and local education agencies and appropriate groups.
(d) Publish and disseminate all requests for proposals.
(e) Through competitive contracts develop and disseminate instructional materials and equipment suitable for bilingual education programs.
(f) Authorized for fiscal year 1979 and for each succeeding fiscal year ending prior to October 1, 1983, $20,000,000 to carry out these provisions.

The response to this mandate included a formal research plan prepared by the Office of the Assistant Secretary for Education (1979) which outlined a comprehensive agenda of research and evaluation for bilingual education. The plan included some 24 research activities responding to all aspects of the congressional mandate. As of this writing, some of the research has been completed while other parts are just beginning.

In 1976, Congress issued a quite general request for information as part of the reauthorization of the Vocational Education Act. An abstract of the relevant legislation (Public Law 94-482, Title V, Part B, Section 523) is given below.

(a) Carry out a study of the extent to which sex discrimination and sex stereotyping exist in vocational education programs assisted under the Vocational Education Act of 1963, and of the progress made to reduce or eliminate discrimination and stereotyping in programs and in the occupations for which programs prepare students.
(b) (1) The National Institute of Education shall undertake a thorough evaluation and study of vocational educational programs. . . . Such a study shall include:
(A) a study of the distribution of vocational education funds in terms of services, occupations, target populations, enrollments, and educational and governmental levels and what such distribution should be in order to meet the greatest human resource needs for the next 10 years;
(B) an examination of how to achieve compliance with, and enforcement of, the provisions of applicable laws of the United States;
(C) an analysis of the means of assessing program quality andeffectiveness;
(D) no more than three experimental studies to be administered by the Institute. . .;
(E) findings and recommendations, including recommendations for changes in such Acts or for new legislation.

(2) Make an interim report to the President and to the Congress not later than September 30, 1979, and make a final report no later than September 30, 1980. Such reports shall not be submitted to any review outside of the Institute before their transmittal to the Congress.

(3) Funds to carry out the provisions of this section shall not exceed $1,000,000 per year for each of the fiscal years ending prior to October 1, 1979.

(4) (A) Submit to the Congress, within 10 months after the date appropria tions become available, a plan for the study to be conducted under this section.

The NIE responded to this mandate by preparing a study plan and contracting for a series of six research and evaluation projects. The overall structure of the vocational education study was patterned after the Title I study that NIE had done in 1977: A team of researchers at NIE prepared the study plan, wrote requests for proposals, selected research contractors, guided the research, and used the contracted research results to prepare reports for Congress.

As part of the Education Amendments of 1978, Congress mandated a study of the Department of Defense's overseas education system for the children of defense personnel. Written into Public Law 95-561 (Title XIV, Section 1412) the call for a study, as shown below, is very broad.

(a) (1) The Director shall provide for a comprehensive study of the entire defense dependents' education system, which shall include a detailed analysis of the education programs and the facilities of the system.

(2) The study . . . shall be conducted by a contractor selected by the Director after an open competition. . . . The contractor shall submit a report to the Director not later than one year after the effective date of this title.

(b) In designing the specifications for the study . . . and in selecting a contractor . . . the Director shall consult with the Advisory Council on Dependents' Education.

(c) The Director shall submit to the Congress not later than one year after the effective date of this title the report . . . describing the results of the study . . . together with the recommendations of the contractor for legislation or any increase in funding needed to improve the defense dependents' education system. . . . Such report shall not be submitted to any review before its transmittal to the Congress.

(d) The Director may provide for additional studies of the defense dependents' education system to be conducted in accordance with the provisions of this section, but such studies shall not be conducted more frequently than once a year.

In spite of the mandate to submit a report to Congress within one year of the legislation (which was dated November 1978), a contract for the study was not awarded until May 1982.

The final example to be presented is the congressional call for an evaluation of the U.S. Trustee program. In 1978, as part of Public Law 95-598 (Section 408), Congress included the following mandate:

> (a) The Attorney General shall conduct such studies and surveys as necessary to evaluate the needs, feasibility, and effectiveness of the United States trustee system and shall report the result of such studies and surveys to the Congress, the President, and the Judicial Conference of the United States, beginning on or before January 3, 1980, and annually thereafter during the transition period.
> (b) Not later than January 3, 1984, the Attorney General shall report to the Congress, to the President, and the Judicial Conference of the United States, as to the feasibility, projected annual cost and effectiveness of the United States trustee system, as determined on the basis of the studies and surveys respecting the operation of the United States trustee system in the districts, together with recommendations as to the desirability and method of proceeding with implementation of the United States trustee system in all judicial districts of the United States.

The evaluation is under under way and, in fact, is scheduled to be completed by the end of 1982 rather than the mandated date of January 3, 1984, to allow adequate time for congressional review and deliberation. Unless continued or modified by legislation, the U.S. Trustee program will terminate on April 1, 1984.

COMMON CHARACTERISTICS OF MANDATED EVALUATIONS

Several common characteristics of congressionally mandated evaluations can be identified by examining the above sections of legislation. One series of legislative requirements deals with operational issues involved in the process of getting the evaluation under way, with funding, and with reporting of results. Except for the bilingual education research mandate, each piece of legislation specifies a date for reporting the results of the study or studies. Though reporting dates are sometimes selected so that information will be available in time to provide input for annual appropriations hearings for the next congressional reauthorization, such well-planned timing is not standard (Boruch and Cordray, 1980). In some cases, Congress simply requires annual reports of progress and findings.

Another characteristic shared by most of the mandates is specification of funding authority. Some of the studies are funded through monies set aside for evaluation in which a proportion of the funds spent on the

program in question—one-half of one percent in the case of ESEA Title I—is allocated for annual evaluation activities. In other cases, Congress authorizes specific dollar amounts for evaluation. Several of the evaluation mandates also contain provisions for reporting results directly to Congress without review or revision by the Executive Branch. Hill (1980) documents both the appropriateness of this requirement as well as the problems it can cause. Finally, all of the reviewed mandates delegate responsibility for the evaluation to an appropriate federal agency.

In addition to these operational issues that occur with some frequency, several other related issues were observed less often. Included here are requirements for the evaluation to be conducted by a competitively selected independent contractor; for Congress to review and approve the evaluation's methodology; for affected groups to have input to the process of planning the evaluations; and for the implementation of demonstration projects when appropriate.

A second series of requirements contained in evaluation legislation has to do with specification of research areas. Those who have criticized Congress for being overly vague in their calls for evaluation are in part supported by the evidence presented here. The examples include some of the largest and most important mandated educational evaluations over the past several years, yet in few cases were evaluation questions clearly specified. The situations where pointed questions were asked are those in which Congress wanted information on the effects of specific proposed changes in legislation—for example, Quie's proposal for changing the formula for the way in which ESEA Title I funds are allocated, and the study of changes in the Title I comparability provision. Thus, when there is a division in Congress or when there has been a public debate over the adequacy or fairness of a particular regulation or aspect of a program, Congress can prepare legislation calling for an investigation of alternative ways of addressing the issue. However, the more common situation—as demonstrated by these examples—involves a general mandate for the purpose of program oversight to "conduct a comprehensive study," to "undertake a thorough evaluation," or to "carry out a program of research." This is not to say that these general mandates are inappropriate; rather, it is important to reinforce the fact that studies that are clearly focused on issues of concern to Congress have a better chance of providing information that is useful for assisting in the policy process than studies focused on issues selected primarily by evaluators. The latter set of issues may be perfectly reasonable, but without congressional input they may not be of maximum utility in program oversight.

A third set of requirements that could potentially be contained in legislative mandates deals with the methodology of conducting evaluations. Reviewing the legislation presented here reveals little in this area. The study containing the most direct language on methodology is the study of ESEA Title I comparability, which specified a sample of all school districts in one state plus up to 20 school districts spread across other states. Furthermore the sample was to be "reasonably representative of the various geographical areas of the Nation." The other mandates reviewed here were much less concerned about methods. Some of them authorized the sponsoring agency to initiate experimental programs if they are needed in order to conduct the research. In the case of the 1974 ESEA Title I legislation, NIE conducted demonstration programs in 13 school districts to test the effects of implementing Representative Quie's proposal. The bilingual education legislation called for a five-year longitudinal study. Still, with the exception of the Title I comparability study, few evaluators would regard this level of input to methodology as a major infringement on the way in which any of the mandated evaluations were to be conducted.

This review has shown that most congressional evaluation mandates consist of general calls for evaluation rather than including specific evaluation questions or methods. On the other hand, there are examples in which Congress has been quite targeted in terms of its information needs, and we reviewed one example in which authorizing legislation included a good deal of language on methodology. The evaluation literature cited contains complaints about a lack of congressional direction in terms of questions to be addressed, but has expressed concern that congressional mandates should not move in the direction of specifying evaluation methods—a concern that appears well-founded based on the Title I Comparability Study. To investigate this issue further we now turn to an in-depth examination of a congressionally mandated evaluation in which the legislation includes quite detailed language both in terms of evaluation questions and evaluation methods.

BACKGROUND TO THE COMMODITY DONATION DEMONSTRATION AND EVALUATION

In December 1980, Congress directed the U.S. Department of Agriculture (USDA) to design and implement a demonstration program and an associated evaluation in order to estimate the effects of two alteratives to the donation of agricultural commodities to schools participating in the National School Lunch Program. The USDA distributes

foods bought under price support and surplus removal legislation to needy recipients through the Commodity Donation Program. In fiscal year 1981, about 90% of all donated commodities (some $900 million worth of food) was given to schools participating in the National School Lunch Program. These donated commodities comprised about 20% of all foods served to children in the school lunch program.

The Commodity Donation Program has been the subject of some controversy in recent years, in part because it has the mandate to satisfy two competing yet somewhat contradictory objectives: to aid American farmers by stabilizing farm prices through the purchase of excess agricultural commodities, while at the same time to improve the nutritional well-being of needy adults and the nation's schoolchildren. The key problem is that the program's major stakeholders (farmers, food processors, food distributors, and schools) have different goals and thus different views on how the Commodity Donation Program should operate. Proponents of the program argue, for example, the following: (1) Commodities are used to help provide children with a wholesome and nutritious meal, and moving away from the current program would result in the use of less nutritious foods in school lunches. (2) Food provided to schools by the USDA costs less than locally purchased food of the same quality. (3) The program provides the agricultural support needed to maintain the viability of the American farmer (any other system could not direct assistance to the specific markets currently included in the USDA's purchase activities). (4) The quality of donated foods is higher than food schools would purchase on their own. (5) The Commodity Donation Program provides assistance to recipients other than schoolchildren (e.g., the elderly, Indians) and provides food supplies that can be used to respond to a national emergency or natural disaster; however, without the donation of commodities to schools, it is unlikely that the remaining distribution system could be operated efficiently.

On the other hand, opponents of the program claim the following: (1) Due to the high cost of transporting, storing, handling, and processing donated commodities into usable products, the cost of donated foods is actually higher than that of locally acquired products. (2) some donated foods are difficult to use in preparing meals and increase the cost of operating a food service program. (3) The uncertainty of delivery dates and the bunching of deliveries at the end of the school year overload local storage capacity, increase costs, and make menu planning difficult. (4) Donated foods are often packaged in ways that are unusable by the schools, include items that children simply do not like, and often arrive in damaged condition. (5) Serving donated commodities lowers student

participation and increases waste in the school lunch program. (6) Current regulations impose an excessive burden on school districts. (7) Agricultural support provided by the program could be achieved by alternative systems that are better for schools and children.[1]

Critics of the program have proposed two basic alternatives to the donation of commodities in order to remedy the problems cited above. The alternatives would transfer some or all of the responsibility for food purchasing from the USDA to the local level, thus giving school districts more freedom in deciding what foods to purchase for school lunches. In addition, both alternatives were proposed as ways to maintain services to children in the face of federal funding cuts by reducing operation costs. One alternative, first proposed in the mid-1970s, consists of simply providing schools with the cash value of the donated commodities they would have received under the Commodity Donation Program, this program is referred to as "cash in lieu of commodities." The USDA currently provides schools with an average of 11 cents worth of donated commodities for each meal served in the school lunch program; under the cash-in-lieu option, schools would be allowed to use these funds in their school lunch program to buy whatever foods they desire. Thus, the cash system gives schools complete control over all foods used in the school lunch program. Some small-scale pilot studies have been conducted to test the effects of using cash instead of commodities (Erickson 1982; USDA 1980), but, because of limited samples and other methodological deficiencies, the results of these studies are inconclusive.

A second alternative was proposed by Congressmen Ford (Michigan) and Goodling (Pennsylvania) when in 1980 they sponsored legislation to implement a "commodity letter-of-credit" system in place of the Commodity Donation Program. This system would permit schools to purchase locally the same general food items that would be donated by the USDA under the Commodity Donation Program. School districts would be given commodity letters of credit (commodity vouchers) for a generic product that the the USDA intended to buy—for example, apples. The schools could then purchase the product locally in a form best suited to their needs—for example, apple sauce, apple juice, or raw apples. The intent of the letter-of-credit system was to enable the USDA to retain some control over the types of foods purchased while giving school districts discretion over both actual purchases and delivery. In some quarters this sytem was seen as more viable politically than the cash alternative because it retains ties to the agricultural sector. Legislation to implement a national letter-of-credit system came very close to reaching the Senate floor, being rejected on a tie vote in the House

Education and Labor Committee, Subcommittee on Elementary, Secondary, and Vocational Education.

Moved by complaints about the program as noted above and as documented by the U.S. General Accounting Office (1977, 1981), by the inconclusiveness of past research, and by the close vote on the Ford-Goodling legislation, Congress passed legislation requiring a demonstration project and associated evaluation to compare the two alternatives against the existing program. The legislation did not consist of a general call for evaluation of the Commodity Donation Program, as was the case in most of the congressionally mandated evaluations reviewed earlier. Rather, since there was a history of debate over the program and alternatives had been proposed, the mandate was very specific about several aspects of the demonstration and evaluation, including the nature of the demonstration alternatives (the treatments), the research questions, and the length of the study. The mandate also contained language relevant to the evaluation design, including the level at which the treatments were to be implemented, the sample size in each treatment group, and the process by which the sample was to be selected.

THE DEMONSTRATION AND EVALUATION MANDATE

In the Appropriations Act for Agriculture, Rural Development and Related Agencies for Fiscal Year 1981 (PL 96-528, December 15, 1980) Congress directed the USDA to examine these two alternatives to commodity donation in the National School Lunch Program: cash in lieu of commodities, and commodity letters of credit. As the case in the earlier referenced studies of Title I comparability and of Representative Quie's proposal to change the mechanism for allocating Title I funds, Congress needed information about how these specific changes would affect a program that had received considerable debate. The legislation required that

> the Secretary shall conduct a 3-year pilot project study in 60 school districts of all cash assistance and all commodity letter of credit assistance in lieu of commodities for the school lunch programs operated in such districts.

The call for a pilot project study was elaborated in three documents: the House Conference Report (No. 96-1519, December 2, 1980), the House Appropriations Report for Fiscal Year 1982, and the *Congressional Record* (December 4, 1980). The House Conference Report stipulated that

the school districts shall be selected by stratified random sample to represent a nationwide variety.... The Secretary shall report the results of the pilot projects to Congress by December 15, 1984, and any school district participating in the pilot projects shall be permitted to continue to participate during the 1984-1985 school year.

The House Appropriations report added the following requirement:

The Committee will expect the Secretary to establish a group of school districts, similar in size, number, and other characteristics to the 60 school districts being studied in order to serve as a control group against which comparisons will be made. In addition, the Committee will expect the Secretary to either withhold, charge for, or earmark in some fashion the value of bonus commodities which may be received by those school districts participating in the pilot study. The Committee feels that this is necessary in order to avoid any bias in the study.

Finally, the *Congressional Record* contains a discussion between Senators McClure and Bellmon which further clarifies the study.

MR. McCLURE: Reference is made to 60 school districts with all-cash assistance and all-commodity letter of credit assistance. I think the intention of the conferees was clearly that there be 30 of each, making a total of 60.

MR. BELLMON: A number of conditions were established which the conferees believed the pilot projects should be conducted.... These conditions are important to assure that the pilot projects are, in fact, conducted fairly and accurately....

First, The Secretary shall publish in the Federal Register a notice of the opportunity for participation in the pilot projects.

Second. Federal and State authorities shall monitor the activities carried out during the pilot projects to ensure the objectivity of these projects.

Third. The Secretary shall allow school districts... to apply for participation in pilot projects conducted under this subsection. If the applications . . . [do] not constitute a stratified random sample of all school districts, the Secretary shall solicit the participation of appropriate school districts.

Fourth. The pilot projects conducted shall include only those school lunches that satisfy the meal pattern requirements promulgated by the Secretary.

Fifth. The Secretary shall submit the methodology that shall be used in the pilot projects to the House Committee on Education and Labor and the Senate Committee on Agriculture, Nutrition, and Forestry at least 15 days prior to the commencement of such pilot projects.

Sixth. The Secretary shall conduct a study to analyze the effect of the pilot projects. The study shall include an assessment of

(a) The administrative feasibility and nutritional effect of cash and letters of credit in lieu of donated foods, cost savings, if any, that may be effected thereby at the Federal, State, and local levels, any additional costs that may be placed on programs and participating students; and

> (b) the effect on farmers, the quality of food served, plate waste in the school lunch program, local economies, and local, regional and national marketing commodities used in the school lunch program, with special emphasis on milk and other dairy products and beef and other meat products.

In addition to this conversation, the *Congressional Record* contained the following statement:

> As part of the pilot study evaluation, USDA will be expected to evaluate the impact of this program on other departmental progams that involve commodity support.

Several features differentiate this mandate for the establishment of demonstration projects and an associated evaluation from the mandates reviewed earlier. The remainder of this paper identifies the common and unique aspects of this call for evaluation and discusses the implications of the mandate for the conduct of the demonstration and evaluation, focusing on areas where the mandate was of great assistance, as well as on other areas where the mandate has proved to be a hindrance.

ANALYSIS OF THE MANDATE AND IMPLICATIONS FOR THE EVALUATION

Perhaps the most striking feature of the mandate for the demonstration and evaluation is its overall high degree of specificity. Whereas the mandates cited earlier were often specific with respect to operational provisions, the present mandate is also prescriptive in areas of evaluation questions and methdology.

STANDARD OPERATIONAL PROVISIONS

The mandate includes all of the standard operational provisions contained in the congressional calls for evaluation that were reviewed earlier. It specifies the length of the study (three years), the date a report is due to Congress (December 15, 1984), and the amount of funds available for the study ($1,975,000).[2] It also indicates that the methodology used to conduct the demonstration and evaluation should be submitted to interested legislative committees for their review. It is important to have all of these conditions written into the legislation. They define the scope of the study in broad terms and specify a date for

providing information to Congress. The call for a three-year study, a report on the methodology of the demonstration and evaluation, and a final report by a given date, as well as the appropriation of funding for the study, all have been valuable in the process of setting bounds for the demonstration and evaluation.

EVALUATION QUESTIONS

In addition to specifying operational provisions, the mandate also is prescriptive with respect to the areas to be covered by the evaluation. The questions to be addressed by the demonstration and evaluation are thus based upon the areas identified in the *Congressional Record* and upon the arguments cited by opponents and proponents of the Commodity Donation Program. The evaluation objectives are to

(1) estimate and compare the costs (food, nonfood, and administrative) and effectiveness associated with the alternative systems;
(2) examine changes in food purchases associated with the alternative systems;
(3) estimate the impact that the alternative systems would have on agricultural commodity markets, farm incomes, the existing food distribution system, government price support and surplus removal programs, and the goal to stabilize prices;
(4) examine the administrative feasibility of implementing the alternative systems nationally;
(5) estimate the impact that the alternative systems would have on participation in the National School Lunch Program, on plate waste, on the quality of food purchased, and on other USDA feeding programs.

As can be seen from the discussion in the *Congressional Record*, Congress's areas of interest were clear. In part this reflected the up-front involvement of the USDA staff who helped frame the evaluation questions. USDA staff members had criticized the earlier studies of cash in lieu of commodities because of selection bias, small sample sizes, and limited measurement. Since these criticisms were made during the March 1980 oversight hearings on child nutrition, committee staff members responsible for preparing legislation for the present evaluation were made aware of the problems faced by the prior studies and subsequently contacted the USDA staff and asked for assistance in drafting legislation that would ensure a sound evaluation. This decision to ask for up-front assistance had a profound effect on the evaluation.

Several outcome areas were defined while the legislation was being prepared and so the job of specifying evaluation objectives did not necessitate the extended, after-the-fact interaction with committee

staffers that Hill (1980) discusses with respect to Title I. Rather, the problem was more one of determining the relative priorities to be placed on the objectives, so that resources could be allocated in order to do the best job of addressing the most important questions. Such an ordering of priorities was important since fully addressing all areas of interest would have necessitated a tremendously expensive study. Further, some areas of congressional interest dealt with issues that should be affected only slightly by the treatments and hence deserved less attention than other areas where treatment effects are more likely.

EVALUATION METHODOLOGY

In contrast with the congressional mandates reviewed earlier, the legislation for the Commodity Donation Program evaluation addresses several key methodological areas. Before discussing the nature and implications of the methodological content of the congressional mandate, it is important to note the desire of the legislators involved to provide for a fair, unbiased, and statistically defensible study. This issue arises three times in the small amount of text quoted earlier; once when Senator Bellmon related the intent of Congress to provide a "statistically sound study," again when Senator McClure discussed conditions that the committee imposed in order to assure that the demonstration and evaluation are "conducted fairly and accurately," and a third time when the *Congressional Record* specified that federal and state authorities should monitor the demonstration in order to "ensure the objectivity of the projects."

Whereas some other evaluation mandates have authorized experimental projects if they were needed, the present leislation mandated a pilot study and prescribed three treatments: cash in lieu of commodities; commodity letters of credit; and use of the Commodity Donation Program as a control group. Specification of the treatments in the congressional mandate is appropriate and warranted in this case. The general intent of the demonstration is to test the effect of giving school districts more freedom in deciding what foods to buy for consumption in the National School Lunch Program, and several different policy options could have been devised to accomplish this intent. However, since the debate about the Commodity Donation Program has centered on the two policy alternatives of cash in lieu of commodities and commodity letters of credit, it made sense for the congressional mandate to specify these two alternatives as those to be implemented in the demonstration. This ensured that the demonstration would test the options of most

interest to Congress—not those of most interest to some particular group of stakeholders.

Since different stakeholders have different views of what should be implemented under the cash in lieu of commodities and commodity letters of credit programs, and since the rules for implementing the two alternatives have a direct impact on the types of statements that could be made by the evaluation, the process of defining the treatments engendered a great deal of debate during the planning stage of the demonstration, and the exact language used in the legislation turned out to be quite important in this debate.

To take this further, a key issue in defining the treatments was related to the fact that school districts receive two forms of support from the commodity program. First, they are given a certain amount of "entitlement" commodities for each meal served in the National School Lunch Program: For the 1982-1983 school year, the value of this entitlement was 11 cents per meal. In addition to entitlement foods, school districts receive "bonus commodities." These are foods that are in such oversupply that the USDA gives school districts as much of them as can be used without waste. In recent years, dairy products have been the principal bonus commodities, and groups supporting the dairy industry are strong supporters of the commodity program inasmuch as it provides the chief outlet for price-supported dairy products. Since school districts are given all the bonus commodities they can use, a district with creative menu planners can obtain a substantial benefit from these free foods.

The USDA's plan for defining the two alternatives was to substitute cash and letters of credit for both entitlement commodities and bonus commodities in order to provide a test of "pure" cash and letter-of-credit systems. However, this plan did not satisfy dairy supporters whose chief interest was to use the school lunch program to reduce the stockpiled supply of dairy products. Providing school districts with bonus cash (which would not be targeted to any specific product) or with bonus letters to credit (which would be targeted to dairy products currently on the market—not those in storage) would do nothing to reduce the embarrassingly large stockpiles. Furthermore, inclusion of bonus cash or letters of credit as part of the alternative treatments would mean that findings from the demonstration project could lead to changes in the method of donating bonus commodities. The counter proposal to the USDA's plan was therefore one in which school districts participating under the cash and letter-of-credit systems would receive cash and letters of credit in place of their entitlement commodities, but would receive bonus commodities just as they had done in the past.

In attempting to resolve the conflict over whether to provide bonuses in the form of actual commodities or in the form of cash and letters of credit, the exact language used in the legislative mandate for the study was invoked. It specified that the demonstration projects would use "all cash assistance" and "all commodity letter-of-credit assistance" (USDA, 1980). This wording was used to maintain the purity of the two alternatives, and thus the USDA's plan for defining the treatments was implemented.

Congressional interest in the demonstration and especially in the definition of treatments remained at a high level during the conduct of the study. Even after the cash and letter-of-credit treatments had been implemented for several months, the debate over how the treat bonus commodities had not subsided. Supporters of the current program and dairy supporters continued to argue for providing bonuses in the form of actual commodities. Senator Paul Traxler made this case in the *Congressional Record* (August 18, 1982), and though a House/Senate Conference Committee upheld the right of the USDA to continue implementation of the demonstration as initially planned (distributing bonus cash and bonus letters of credit to school districts instead of bonus commodities), the Conference Report accompanying the fiscal year 1982 Supplemental Appropriations Bill contained language that could have had a profound effect on the evaluation:

> The pilot products may proceed as planned by the Department, with distribution of bonus cash or bonus letters of credit in lieu of bonus commodities, but the Department is directed to eliminate such bonuses from consideration in the evaluation phase of the school lunch demonstration projects (p. 8).

Exactly how this could have been accomplished is not clear. However, it proved unnecessary to precisely adhere to this mandate since the language satisfied neither opponents nor proponents of the current program. Actors on all sides could only lose by having bonus commodities—a key part of the program—included in the treatments but completely eliminated from consideration in the evaluation. Calls from program supporters to committee staff members quickly revealed that confusion existed over the bonus issue, and that the language in the Supplemental Appropriations Bill did not reflect the intentions of the program supporters. This finally led to the solution: The USDA would prepare a memorandum of understanding on how the congressional mandate on bonus commodities would be interpreted. This memorandum had the effect of allowing the evaluation to proceed as planned—as if the language in the 1982 Supplemental Appropriations Bill had never existed.[3]

To review, initial specification of the number and nature of the treatments is a key feature of the congressional mandate for the commodity donation study. It focused the demonstration on the two policy options of greatest interest to Congress, and called for explicit comparisons to be made against the current program by including a control group of school districts.[4] Finally, it is clear that the issue of treatment definition, as exemplified by the problem of how to deal with bonus commodities, will continue to be debated throughout the demonstration.

The original mandate also specified that the treatments are to be implemented by school districts rather than by some other level of the commodity system (such as schools or states). This makes sense since school districts are the main implementors of the Commodity Donation Program. Though schools actually prepare the meals for the school lunch program, and though funds for the program are allocated to school districts at the state level, the school lunch program is organized at the school district level. Funds flow to the school district rather than to individual schools, and in most districts the school lunch program is run centrally rather than from individual schools. Since the intent of the demonstration is to test the effects of transferring the power to decide what foods are bought for the school lunch program from the Department of Agriculture to the local level, and since school districts already purchase the majority of food used in the school lunch program using local funds, cash receipts from children, and federal cash reimbursments, it is reasonable for school districts to also do the purchasing of foods using resources provided through the Commodity Donation Program.

The legislation specified the overall sample size as well as the distribution of school districts across treatment groups—30 in each of three groups for a total of 90. On the face of it there is no compelling reason to specify the overall sample size or the distribution of school districts across treatment groups in the congressional mandate. This decision should not be prespecified. Rather, the sampling design should be developed by evaluation contractors (in conjunction with the federal agency responsible for the evaluation) with the appropriate technical skills to make tradeoffs between the size of a sample needed in order to assure adequate power for the statistical analyses and the sample size that can be afforded. The decision to distribute the sample equally among the three groups should also be left to those with technical skills. Though equal sample size per group may make the analysis simpler, there may well be an argument for including more school districts in the

two alternative treatments in order to obtain better information on the effects those treatments have on different types of school districts.

In defense of the congressional mandate, it should be pointed out that the rationale for including specifications on the sample size was to ensure a statistically sound study. Unfortunately, in asking for advice on what size sample to specify, the authors of the mandate contacted staff in the USDA who had only limited experience in conducting large-scale program evaluations. By the time the USDA agency responsible for the demonstration and evaluation reviewed the legislation, it was too late to change the sample size.

It should be noted that the power afforded by the sample of 90 school districts, 30 in each treatment group, is neither particularly high nor particularly low. Though a larger sample would be desirable from the viewpoint of increasing statistical power, the legislated sample should allow reasonable power for statistical tests of the main treatment effects. The real weaknesses of the sample are that it is not large enough to withstand attrition of school districts and still retain reasonable power,[5] and that it is not large enough to allow for more than a cursory investigation of the effects of the alternatives on subsets of school districts—for example, large versus small districts or rural versus urban districts.

Though the intent of specifying the sample size in the congressional mandate was to ensure implementation of a statistically valid study, the fact remains that specification of the sample size should be done by those with technical training. This could have been done in time for inclusion in the legislation if the writers of the legislation had been able to contact USDA staff members with appropriate training, or it could have been omitted from the legislation and done after the fact by the federal agency in conjunction with a contractor.[6] We would opt for the latter approach simply because the time pressures accompanying the preparation of legislation may not allow for sufficient consideration of the sample size, and because, as demonstrated here, the technical advice obtained at this point may not be of the highest quality.

The mandate also included language bearing on the sample selection process, including the somewhat contradictory stipulations that (1) the school districts should represent a national stratified random sample and that (2) school districts should be able to apply for participation in the demonstration. This language was included with the hope of avoiding the problems of prior USDA studies of the cash-in-lieu-of-commodities program which had been criticized due to small and geo-

graphically limited samples. Yet, to stipulate that the sample contain volunteers and also represent a nationally representative random sample is contradictory and has led to more complications than are necessary.

CONCLUSION

Evaluators have called for Congress to be more specific in its requests for information, claiming that more targeted questions would result in more useful evaluations. Concern has also been expressed that congressional mandates should refrain from specifying methodological details. This article contains a review of the legislative mandates behind several congressionally mandated evaluations and identifies three areas in which requirements have been specified: (1) standard operating provisions, (2) evaluation questions, and (3) evaluation methodology.

The typical congressional call for evaluation contains several standard operating provisions that set forth the general topic of the evaluation, the responsible federal agency, the length of the study, the date(s) a report is due, and the funding authority. Also included in some mandates are provisions for having Congress review plans for the evaluation, for reporting results directly to Congress, and for having the evaluation done by an independent contractor. The inclusion of any or all of these standard provisions is valuable. They help frame and bound the activity, and give guidance that helps the responsible federal agency and its contractor(s) to punctually provide information to the appropriate audiences.

Evaluation mandates are much less often prescriptive with respect to evaluation questions. In most cases, the mandate contains only a broad call for evaluation, and staff from the responsible federal agency typically have had to spend a great deal of time working with congressional committee staff members to determine more specific information needs. This process can be shortened and simplified considerably (though not eliminated) through up-front contact between committee staff members and federal agency staff so that the mandate can include research areas, outcomes, or objectives of greatest interest to Congress. Being specific about areas of interest in congressional calls for evaluation will lead to more useful evaluations by improving the evaluators' understanding of Congress's information needs and by shortening the process of negotiation over evaluation objectives.

A third set of specifications—those related to evaluation methods—are included in some congressionally mandated evaluations. Though language in this area is relatively infrequent, its inclusion has proved unnecessarily restrictive at best, and at worst can diminish the quality of evaluations. It is not possible to recommend simply that congressional mandates ignore methodological questions because of the experience in the Commodity Donation Evaluation in which Congress wanted to be certain that the study was conducted in a fair, unbiased, and statistically sound manner. This is a legitimate and commendable concern that could be satisfied by other, less restrictive means. In particular, rather than including specifics on the details of defining the treatments, sample size, distribution of sites, or sampling methods in the congressional mandate, it would be preferable to delegate such judgments to someone with the appropriate technical skills (usually a contractor). The evaluation mandates could specify (as several have) that the resulting evaluation plan be submitted to the appropriate Congressional committees and/or to their consultants for review and approval. This would allow evaluation experts to make the initial decisions about methodological issues while providing Congress with a vehicle for ensuring that the methods are adequate from their point of view.

In sum, improved evaluations will result from continued congressional interest in specifying information needs, as well as from Congress's willingness to leave decisions about evaluation methods to those with appropriate training.

NOTES

1. These arguments have been abstracted from testimony given before the Committee on Small Business, House of Representatives, April 28, April 29, May 13, and July 15, 1981, and before the Committee on Education and Labor, House of Representatives, March 18, 1981.

2. These funds were appropriated specifically for the evaluation. Additional funding is coming from discretionary evaluation funds within the USDA.

3. Since the time this article was prepared for publication, the debate over bonus commodities has continued, with important impacts on the evaluation. The full story, however, is more appropriate for a different article.

4. The congressional mandate did not initially specify a control group. After reviewing the mandate, the USDA evaluation staff contacted committee staff members and arranged for the House Appropriations report to call for a control group.

5. The actual sample included 98 school districts in order to provide some protection against attrition.

6. In fact, USDA evaluators wanted to ask the congressional committee to authorize a larger sample size. These efforts were abandoned on political grounds.

REFERENCES

BORUCH, R. F. and D. S. CORDRAY (1980) An Appraisal of Educational Program Evaluations: Federal, State, and Local Agencies. Evanston, IL: Department of Psychology, Northwestern University.

BROWN, L. L., A. L. GINSBURG, and B. J. TURNBULL (1979) "Mandated studies and information needs in the policy system: the case of bilingual education." Presented at the annual meetings of the American Educational Research Association, San Francisco.

CRONBACH, L. J. et al. (1980) Toward Reform Evaluation: Aims, Methods, and Institutional Arrangements. San Francisco: Jossey-Bass.

ELLMAN, F., L. FERRARA, J. MOSKOWITZ, and S. STEWART (1981) Utilization and effects of alternative measure of comparability. Washington, DC: AUI Policy Research.

ERICKSON, D. B. (1982) "Cost of producing school lunches: Using USDA donated commodities versus cash in lieu of commodities." School Food Service Research Rev. 6, 1: 26-31.

FLORIO, D. H., M. M. BEHRMANN, and D. L. GOLTZ (1979) "What do policymkaers think of educational research and improvement? Or do they?" Educ. Evaluation and Policy Analysis 1,6: 61-87.

HILL, P. T. (1980) "Evaluating education programs for federal policymakers: lessons from the NIE Compensatory Education Study," in J. Pincus (ed.) Educational Evaluation in the Public Policy Setting. Santa Monica, CA: Rand Corporation.

LEVINE, R. A. (1981) "Program evaluation and policy analysis in western nations: an overview," in R. A. Levine et al. (eds.) Evaluation research and practice: Comparative and international perspectives. Beverly Hills, CA: Sage.

LEVITON, L. C. and E. F. HUGHES (1981) "Research on the utilization of evaluations: a review and synthesis." Evaluation Rev. 5, 4: 525-548.

National Institute of Education (1978) Compensatory Education Study: A Final Report. Washington, DC: Author.

PATTON, M. Q. (1978) Utilization-Focused Evaluation. Beverly Hills, CA: Sage.

RAIZEN, S. A. and P. H. ROSSI [eds.] (1981) Program Evaluation in Education: When How, to What Ends? Washington, DC: National Academy Press.

U.S. Department of Agriculture (1980) A Study of Cash in Lieu of Commodities in School Food Service Programs. Washington DC: Author.

U.S. General Accounting Office (1981) More Can Be Done to Improve the Department of Agriculture's Commodity Donation Program. Washington, DC: Author.

——— (1977) The Impact of Federal Commodity Donations on the School Lunch Program. Washington, DC: Author.

7

The Science and Politics of Cyclamate

William R. Havender

Most of us pay little attention to the body of Federal law that governs the safety of our food supply. This is because these laws function smoothly and harmoniously most of the time. However, there is one portion of these regulations—the Delaney clause of the Food, Drug, and Cosmetic Act—that is repeatedly in the news, and is, it seems, constantly coming up against hard cases.[1]

It is received wisdom among most specialists and the general public that this clause is a special source of trouble, the repeal of which would simplify the production of diet pop, hot dogs, and bacon. There is good reason for this view, since the Delaney clause did play a central role in several unpopular decisions of the Food and Drug Administration (FDA), such as the 1969 ban on cyclamate, and the abortive attempts to ban saccharin and nitrite. Most

[1] Section 409 of the Federal Food, Drug, and Cosmetic Act reads, in part, that a food additive petition shall not be issued if a fair evaluation of the data "fails to establish that the proposed use of the food additive, under the conditions of use to be specified in the regulation, will be safe: Provided, that no additive shall be deemed to be safe if it is found to induce cancer when ingested by man or animal, or if it is found, after tests which are appropriate for the evaluation of the safety of food additives, to induce cancer in man or animal . . ." The phrase preceding the colon is the "general safety" clause; that which follows is the Delaney clause.

From William R. Havender, "The Science and Politics of Cyclamate." Reprinted with permission of the author and publisher from *The Public Interest*, No. 71 (Winter 1983), 17–32.

of the public debate has focused on how the Delaney clause ties the hands of the FDA, not permitting it leeway to judge the appropriate human risk to be inferred from animal cancer tests, nor the discretion to weigh a very substantial health benefit against a slight or hypothetical cancer risk (for instance, nitrite prevents botulism).

But Peter Hutt, FDA chief counsel from 1971 to 1975, has long argued that the Delaney clause is redundant: that the FDA is already legally empowered by the "general safety" clause to take any action sanctioned by Delaney. However, this has not traditionally been the FDA's own interpretation of its legal mandate. In the past, the FDA has usually argued that Delaney demanded a far more stringent regulatory response to cancer findings than did the general safety clause: that, absent Delaney, weak or merely suggestive indications of carcinogenicity would not by themselves necessarily compel a ban, and the benefits of an additive could still be balanced against its risks. But a recent decision shows that the FDA has now changed its views, and that the general safety clause may prove a greater source of trouble than the Delaney clause.

In September 1980 the FDA formally denied Abbott Laboratories' 1973 petition for the reapproval of cyclamate. Especially notable about this decision is that, while the question of cancer was the principal ground for the denial, the Delaney clause was explicitly *not* invoked.[2] Instead, the FDA relied exclusively on a new and expansive interpretation of the general safety clause. This decision thus established an ambitious precedent that may guide future decisions on food additives—for instance, when the ban moratorium on saccharin runs out on August 14, 1983—and so is worth examining in some detail.

A cyclamate primer

Cyclamate is a chemical with thirty times the sweetening power of sugar but with none of the calories. (The plural form that is sometimes used—cyclamates—refers simply to the various salts of

[2] A secondary ground was the alleged mutagenicity of cyclamate. These experiments will not be discussed in detail here. But their interpretation does express the same attitude toward the use of scientific data that will be documented here concerning carcinogenicity. In particular, not a single study decisively demonstrated that cyclamate is mutagenic, although a few were "suggestive." The role that "suggestive" results played in this decision will be made clear in the discussion of cancer.

cyclamic acid, such as sodium cyclamate and calcium cyclamate). It was discovered in 1937 by Michael Sweda, a chemist then working for Du Pont, which licensed the substance to Abbott Laboratories. Cyclamate has two useful properties: It lacks the bitter aftertaste of saccharin, and it synergizes with saccharin so that a mixture of the two tastes sweeter than their simple sum. The optimal product was found to be a 10:1 cyclamate/saccharin mixture (hereafter referred to as the "10:1 mix"), in which each component contributed about half of the final sweetening power. This mix was introduced in 1953 and soon dominated the diet food industry.

In 1959, cyclamate and saccharin were classified "Generally Recognized As Safe" (GRAS) by the FDA under the Food Additive Amendment of 1958. This classification was based on animal studies and human experience with each sweetener separately. In order to evaluate the safety of the combination of the two substances, Abbott commissioned a rat feeding study on the 10:1 mix to be conducted by Food and Drug Laboratories (FDRL) under the direction of Dr. Bernard Oser. (This laboratory is not connected with the FDA despite the similarity in names.)

Toward the end of this two-year study, another researcher reported that cholesterol pellets impregnated with cyclamate and surgically implanted in the bladders of rats caused much higher tumor incidence than cholesterol pellets did alone. The relevance to human oral intake of this method of administration was obscure (and indeed, such experiments have never been taken as indicative of human risk). But this finding caused Abbott to ensure that the animals in the FDRL study were examined for bladder cancer during, and at the termination of, the experiment. In early October 1969, before all the histological examinations had been completed, Abbott promptly communicated to the FDA partial results that indicated the presence of bladder tumors. A group of pathologists convened by the FDA confirmed these diagnoses. The final test result was that twelve of the seventy rats that had received the highest dose (5 percent of the diet) were found to have incipient bladder cancer, while none of the undosed (control) animals did. This difference was statistically significant.

Within a week of Abbott's notification, the FDA called a press conference and, citing the Delaney clause, announced it was removing cyclamate from the GRAS list and banning its use in general purpose foods and non-prescription drugs. Eleven months later its use in flavoring prescription drugs was also stopped. In 1973, Abbott petitioned the FDA to have cyclamate reapproved on the

basis of new studies. After many exchanges, the FDA issued its final decision on September 16, 1980. Just this past fall Abbott submitted a new petition for the approval of cyclamate; this petition is currently under review, but a decision is not expected before the spring of 1983.

Irregular statistics: (I) lack of replicability

In choosing not to cite the Delaney clause in this case, and instead reinterpreting its mandate under the general safety provision, the FDA decided that cyclamate, like Caesar's wife, would have to be above all suspicion. This stance allowed the FDA unprecedented license to rummage through the vast array of data now available about cyclamate, finding bits of evidence here and there, none of it secure enough to support a determination of carcinogenicity under the Delaney clause, but sufficient to raise "questions" under the general safety clause. Inherent in such a stance was the temptation to be capricious in the interpretation of the data, a temptation the FDA did not avoid. Three instances stand out.

One irregularity was the unsound way in which the FDA handled the normal scientific requirement that an experimental result be shown to be replicable before being accepted as valid. To be sure, there was a difficult regulatory decision to be taken in 1969, when it was first reported that the 10:1 mix induced bladder cancer in rats. Even though these tumors were minute (most were only visible microscopically) and had no significant deleterious effects on the health of the affected animals, the fact that tumorigenic activity was seen at all was understandably alarming, given the wide use of artificial sweeteners. Besides, saccharin could still be used by those—chiefly diabetics and dieters—who had a true health need for a non-nutritive sweetener.[3] So the policy choice was largely between improved palatability on the one hand, and a widespread potential health hazard on the other.[4] And deciding to withhold judgment pending replication of the tumor findings would mean waiting at least two years more, during which time public ex-

[3] It was conceivable, of course, that the tumors were induced by the saccharin in the mix (or impurities). But cyclamate was clearly the main component, and saccharin had evidently been used safely for decades. It was not unreasonable, therefore, that suspicion focused mainly on cyclamate.

[4] Manufacturers' costs were not considered, but they were quite high. Estimates of lost inventories were in excess of $100 million (see editorial, *Barrons*, October 7, 1974). A lawsuit brought against the government for $23 million by the California Canners and Growers Association for losses suffered as a result of the sudden ban is before the courts now, and a decision is expected soon.

posure would continue. Therefore, one can understand the decision taken in 1969 to withdraw cyclamate from the market, even in the absence of any independent verification of the tumor findings.

But the implementation of this ban stimulated a large number of supplementary animal tests to confirm the original report of bladder tumors. These newer results included ten experiments on rats, seven on mice, one on hamsters, one on beagles, and two on monkeys. Among the rat studies, two were conducted on the 10:1 mix (and hence directly attempted to reproduce the FDRL test), five were on cyclamate itself, and three were on cyclohexylamine (the principal metabolite of cyclamate produced in the body, hereafter called CHA). All of these were high-dose, long-term feeding studies, and none produced a statistically significant incidence of bladder tumors.

Ordinarily in science, the inability to repeat a singular result, despite a sustained and deliberate effort, should lead to a judgment that the initial finding was wrong. Things are not that simple in the regulatory world, however. There *were* a few odd bladder tumors scattered among the treated rats in these newer tests, but not enough to establish statistical significance within any single experiment. Yet instead of viewing these tests as several independent confirmations of cyclamate's *lack* of carcinogenicity—which scientists would normally do, and which they would interpret as strong exoneration—the FDA lumped together the various negative tests involving the same strain of rats. Then they did a statistical analysis, *not* against the lumped control animals from the corresponding tests (which still would not have yielded a statistically significant result), but against the average "spontaneous" incidence of bladder tumors based on "historical" data derived from experiments carried out under other conditions at other times and in other laboratories. By these means statistical "significance," of a sort, was achieved for the *three* bladder tumors seen among the more than 400 treated animals in the three tests involving Sprague-Dawley rats, and as it was for the three tumors seen in the more than 200 treated rats in the two studies on Wistar rats.

Now the unreliability of this procedure owes to the fact that the occurrence of spontaneous tumors in test animals is not fixed but varies from experiment to experiment, from time to time, and from laboratory to laboratory.[5] Sometimes this variability can be traced

[5] Task Force of Past Presidents of the Society of Toxicology, "Animal Data in Hazard Evaluation: Paths and Pitfalls," *Fundamental and Applied Toxicology* 2 (July, 1982): 103-104.

to identifiable differences in the maintenance conditions of the animals (such as differences in the feed, or the drinking water, or whether animals were caged individually or in groups), or to differences in the thoroughness of the search for tumors (whether or not the bladders were inflated after autopsy, or whether they were scrutinized microscopically rather than merely scanned visually). Often, however, this variability is unexplained. To avoid biases and inaccuracies that these uncontrolled influences can introduce, careful experimental design requires that dosed and undosed animals be kept in the same laboratory under identical treatment conditions, that the animals be randomly assigned to the treated and untreated groups, and that they be examined for tumors by uniform procedures. Without such care to eliminate (or at least even out) the possible influence of extraneous factors, there is no way to know if the "historical" spontaneous incidence can be used validly as the standard against which to compare the dosed animals in these specific experiments. Hence, the results of such a statistical analysis, however "significant," can easily be spurious.

The FDA was not unaware of the vulnerable nature of this analysis, which is why the agency did not invoke the Delaney clause on the basis of these "significant" incidences. Under its new "Caesar's wife" reading of the general safety provision, however, it no longer needed to. All that was needed was for a "question" to be raised, and the FDA considered this analysis, however inherently suspect and deviant from the proper statistical comparison with the simultaneously-run controls, to be ample for this purpose. Specifically, the FDA concluded that "the lack of a statistically significant effect in each of these studies when considered alone does not rebut the question about cyclamate's safety raised by the comparison between the combined incidence of bladder tumors found in cyclamate treated Sprague-Dawley and Wistar rats and the background rate for such tumors based on historical data."[6] So it was that the FDA, by a statistical sleight-of-hand, converted five clearly negative and mutually reinforcing studies into five experiments that "raised a question about cyclamate's safety."

The statistical background

The second irregularity in the FDA's analysis of the cyclamate data concerns the science of statistical hypothesis testing. To under-

[6] J. Goyan, "Cyclamate (Cyclamic Acid, Calcium Cyclamate, and Sodium Cyclamate): Commissioner's Decision," *Federal Register* 45, Number 181 (September 16, 1980): 61491.

stand the import of the FDA's radical deviation from normal practice, a brief detour into this science will prove helpful. The usual point of departure in statistically evaluating an animal cancer test is to see whether or not a true "treatment effect" exists. After all, even if a chemical is not a carcinogen, tumors often occur spontaneously (particularly in aged animals), and there is a small assortive chance that most of the animals that are fated to develop spontaneous tumors will be placed in the dosed groups, thus making the tested chemical look like a carcinogen when it is not. Statistical tables give the exact probability that such an extreme outcome might come about purely by chance. This probability has a technical name: the "p-value." The lower the p-value of an experimental outcome, the more unlikely it is that the result could have been obtained just by chance. Only when an apparently carcinogenic result is highly unlikely by assortative chance—when the p-value is very low—does one accept the result as "significant" in the statistical sense.

One must, of course, adopt some precise and uniform rule for deciding when a p-value is low enough to be called significant. Very commonly in cancer testing (and in many other experimental situations) this cutoff value is set as $p \leq .05$ (on a scale ranging from 0 to 1). This means that we only accept results that have a 5 percent probability (or less) of occurring simply by assortative chance.

There are strong reasons for not choosing values that are very much higher or very much lower than this. To begin with, note that choosing this value as the decision boundary means that, on the average, there is no more than one chance in twenty that, were the chemical under test truly a non-carcinogen, one would obtain a result that would mistakenly lead one to judge it to be a carcinogen. (This is a useful way to view it: Were the test on the non-carcinogen repeated twenty times, we would expect to classify the substance, falsely, as a carcinogen no more than once). But the obverse is also true: One *would* anticipate making that mistake on as many as 5 percent of the non-carcinogens.[7] If one set the decision boundary higher (say, $p \leq .10$) then one would expect as many as one in ten such experiments to lead to a false judgment of carcinogenicity. If one set the bound lower (say $p \leq$

[7] To see experimentally a false positive rate as high as 5 percent presupposes that enough spontaneous tumors do in fact arise that they *could* be assorted by chance across dose groups in patterns suggestive of carcinogenicity having p-values $\leq .05$. This is not, however, a very restrictive presupposition for cancer bioassays that run for the natural lifetime of the test animals. It is com-

.01) then as many as one in a hundred such tests would be expected to lead to a false judgment of carcinogenicity. (Such wrong judgments are termed "false positives.") The decision bound one chooses, then, limits the fraction of non-carcinogens that might be judged as false positives.

But saying that an experimental result meets our $p \leq .05$ criterion is *not* equivalent to saying that the result has at least a 95 percent chance of being true (i.e., that 19 out of 20 such positive judgments would be valid on average). For example, if all of a group of chemicals under test were, in fact, non-carcinogens, then *every* judgment of carcinogenicity would be incorrect (even though no more than 5 percent of the group would be so judged), while if the group consisted only of carcinogens, then every judgment of carcinogenicity would be correct. Thus, looking only at the *sub*-group of chemicals that *have* been judged to be carcinogens by the $p \leq .05$ rule, the fraction of these that will be false positives can range from 0 to 100 percent, depending on the *true* frequency of carcinogens in the group of chemicals under test. This is an *extremely* common mistake in the interpretation of p-values.

Suppose we have a situation where there is one true carcinogen among a group of 100 chemicals selected for testing. (This is not unrealistic, since it has long been an article of faith that carcinogenicity is a fairly rare property of chemicals.) Our animal test would presumably detect the one true carcinogen. But in addition, our $p \leq .05$ decision rule would expectedly lead to as many as 5 percent of the 99 non-carcinogens (or about five of them) being falsely judged to be carcinogens. Of the six positive results, then, five would be false. If the true frequency of carcinogens were instead twenty among the 100 chemicals being tested, then one would presumably correctly identify these twenty (assuming there are no false negatives), but in addition, as many as 5 percent of the eighty non-carcinogens (or four) would also be judged to be carcinogens. Here, only four among 24 positive judgments would be false. Yet if the true frequency of carcinogens were one among 1000 tested chemicals, then the number of positive judgments could be as high

---

mon in aged animals for the spontaneous tumor incidence of at least one tissue site to be high enough to generate an expectation of false positive results on the order of 5 percent. For the expected false positive rates for rat and mouse strains used in the NCI/NTP cancer bioassay series see T.R. Fears, R.E. Tarone, and K.C. Chu, "False-Positive and False-Negative Rates for Carcinogenicity Screens," *Cancer Research* 37 (1977): 1941-1945; and J.J. Gart, K.C. Chu, and R.E. Tarone, "Statistical Issues in Interpretation of Cronic Bioassay Tests for Carcinogenicity," *Journal of the National Cancer Institute* 62 (1979): 957-974.

as 51, of which fifty would be false. Thus, *the fraction of positive judgments that are false can be very large when the true frequency of carcinogens in a group of tested chemicals is much lower than the p-value chosen as the decision boundary.*

Stated another way, as the proportion of true carcinogens declines, the chance of falsely judging carcinogenicity increases (for any given decision boundary). This also means that, for a given true incidence of carcinogens, raising the decision bound increases very rapidly the proportion of false positives among all positive judgments. Limiting these wrong judgments so that the true carcinogens are not swamped with false positives is a compelling reason why scientists do not usually use decision bounds much higher than $p \leq .05$ in animal cancer tests.

Of course, there is a danger of making the opposite sort of mistake, of concluding that a chemical is not a carcinogen when in fact it is (for example, if the animals that would develop spontaneous tumors ended up by chance in the undosed group, so that the proportion of *induced* tumors in the dosed animals showed no or very little increase over the undosed animals). This is called a false negative. If one were only concerned with false negatives, the decision boundary could not be set too high. But a balance must be struck, recognizing that the consequences of a false negative in a cancer test can be serious (because a carcinogen might be ingested by millions of consumers), but that the consequences of a high proportion of false positives can also be serious (because the economy would be hurt by the frequent, sudden banning of widely useful chemicals). The decision bound that is generally viewed as striking the best balance between these two opposing sorts of error, and which is currently predominant in the field of animal cancer testing, is $p \leq .05$. (This is the practice adopted, for example, by the government in its own large animal cancer testing program, which is carried out by the National Cancer Institute/National Toxicology Program (NCI/NTP).)

(II) Statistical boundaries

This should make clear the severe problems with FDA Commissioner Jere Goyan's decision to relax this standard decision bound in the cyclamate tests. Commissioner Goyan reasoned that an experimental result with a p-value of exactly .05 (which would just be significant by the usual criterion) is not really all that different from one with a p-value of .06 (which would just miss be-

ing considered statistically significant), and hence the latter should not be discounted merely because of the arbitrariness of a statistical rule. It is true, of course, that a result with $p = .06$ is hardly different from one with $p = .05$. But the problem with this argument is that there is no logical stopping point. And, in fact, Goyan did not stop. He rummaged through the data, discovering result after result with p-values greater than .05 which he interpreted to be "significant," undaunted even by a rat study with a very few bladder tumors in which the p-value was .2. (That is, even if there were no true carcinogenic effect, a result as extreme as this one would show up one time in five purely by chance.) Thus, in response to Abbott's contention that this study, which was not nearly significant at the $p \leq .05$ cutoff point, should be counted as a negative, Goyan wrote in his decision: "I disagree. The total tumor incidence in this study is significant at the $p = .2$ level. There is thus an 80 percent chance that the results of [this] study are due to cyclamate instead of a 95 percent probability necessary for statistical significance at the $p \leq .05$ level . . . I do not consider these results to be negative . . . the study cannot be considered proof of safety and indeed raises a question as to the potential carcinogenicity of cyclamate." (Here Commissioner Goyan makes the common mistake in interpreting p-values.) Once again, a study that would be clearly negative by normal criteria had been alchemically converted into its opposite: one that "raised a question" about cyclamate's carcinogenicity.

Perhaps what is most unsettling here is the thought that the FDA may intend in the future to accept as "significant" p-values as high as .2. For this rule would greatly increase the number of tested substances judged to be carcinogens, subjecting them to a ban or other stringent regulation. And if true carcinogens are as relatively rare as scientists believe, then a large fraction—perhaps a majority—of these supposed carcinogens will be falsely so identified. To most people, this would be carrying "prudence" rather far.

Such a thoroughgoing revision of statistical methodology might be expected to excite some remark from the statistics profession, and in fact it did. A letter in reference to the cyclamate decision was sent to the Commissioner of the FDA from C.R. Buncher, Chairperson of the Executive Committee of the Biopharmaceutical Section of the American Statistical Association, and Professor of Epidemiology and Biostatistics at the University of Cincinnati. This letter reads in part:

> . . . we are concerned about the extreme misrepresentation of our pro-

> fessional methodology. . . . Many of the published statements are profoundly fallacious. . . . The concept expressed [concerning the interpretation of p-values] is foreign to everything that is taught in the statistics profession. . . . We strongly encourage you to have the appropriate professionals prepare a new statement that correctly expresses the statistical principles that are involved in this issue. This new statement is needed to avoid the ridicule of knowledgeable scientists. . . . WE BELIEVE THAT A REVISION MUST BE PUBLISHED IN THE FEDERAL REGISTER AS A CORRECTION.

No such correction has yet been published.

(III) Unreasonable "sensitivity"

The third instance of the FDA's irregular use of statistics is the plaint, oft-repeated by the Commissioner in this decision, about the lack of "statistical sensitivity" in the animal studies under review. Time after time a calculation was offered that showed that studies of the size that were used had only a 50 percent chance of detecting at the $p \leq .05$ level a true difference of X (where X varied from a few percent up to some 30 percent, depending on the number of animals in the experiment). What this meant was that with the finite number of animals in any single experiment, one stood a fair chance of missing "small" carcinogenic effects entirely, and the smaller the experiment, the larger the effects that might be missed.

It would be perfectly reasonable to raise this objection if an experiment used an unusually small number of animals. But only two or three of the many cyclamate tests did so; the great majority were of a size comparable to, or larger than, the original FDRL study that had produced the first (and only) finding of bladder cancer (and hence, were statistically capable of contradicting it). They were also comparable in size to the exemplary NCI/NTP animal cancer test series, which seldom uses group sizes larger than fifty.

To be sure, such a group size has a 50 percent chance of detecting a tumor incidence at the $p \leq .05$ level only if the true incidence is as large as 8 or 9 percent, and hence stands a fair chance of missing lower incidences. But this is a general problem with animal cancer testing, not something unique to the cyclamate tests; besides, this limited statistical sensitivity is precisely why enormous doses far exceeding normal human exposure are used. The 5 percent dietary dose, for example, which is the typical maximal dose used in cyclamate studies, corresponds to hundreds of bottles of diet pop

daily, and this is maintained over the entire post-weaning lifetime of the animals. This huge dose is a giant precautionary factor *already* introduced into such experiments expressly to compensate for the limited numbers of animals that can be used in a routine test.

An example of the FDA's capriciousness is their treatment of one of the largest of the rat studies on cyclamate, carried out in Germany. Actually, three feeding studies were performed, one each on cyclamate, the 10:1 mix, and CHA. Only one bladder tumor was seen among the 208 animals dosed with cyclamate, none in the 208 treated with the 10:1 mix, none in the 104 animals treated with CHA, and none in the 104 controls. How did the FDA evaluate this outcome? Commissioner Goyan wrote: ". . . the occurrence of a bladder tumor in [this] study is consistent with a small treatment effect, even though it is not significant at the $p \leq .05$ level. . . . Accordingly, I cannot consider [this] study to be proof of cyclamate's safety." [8]

What sort of study *would* the FDA consider of adequate statistical sensitivity to rebut the "question" of cancer hazard raised by the occasional tumors seen in these tests? Commissioner Goyan again: "In the case of cyclamate, it is certainly possible that further adequate testing, such as the study proposed by the Temporary Committee, could resolve the current questions about cyclamate's possible carcinogenicity.[9] If such testing is done, it may yet be possible for FDA to conclude that there is a reasonable certainty that cyclamates does not cause cancer."

On the face of it this sounds like a reasonable stance, but it is instructive to see what lies behind the words. For what the Temporary Committee had discussed (not "proposed") was an idealized study with no limits on cost and technical resources, and with the following features: It should be large enough to have a 95 percent chance at the $p \leq .05$ level of detecting a 1 percent difference in incidence within each dose group, and should consist of both sexes of two species that would be tested with at least three doses, in addition to the controls. This would require 51,968 animals, and

[8] Commissioner Goyan gave the historical incidence of bladder tumors in this strain of rats (Sprague-Dawley) as .23 percent, that is, 2.3 per thousand rats. Thus, seeing about one such tumor among the 624 rats in this test would be expected. The probability that a bladder tumor, given that one occurs at all, would be found in a dosed animal rather than a control animal, is simply $520/624 = 5/6 = .833$, that is, $p = .833$. Here, the Commissioner treated a result with a p-value greater than .8 as of sufficient weight to invalidate counting this study as a negative.

[9] This was a committee set up by the National Cancer Institute to review the available cyclamate studies. Its report was issued in 1976.

if allowance were made for a possible 50 percent premature mortality rate—animals often die well before the end of a two year experiment from non-cancerous causes, and so never get old enough to be at risk of developing bladder cancer—then twice this number of animals would be needed to start the experiment.[10] So say 100,000 animals: Currently, such tests cost about $1000 an animal, which means an expenditure of $100 million, a sum most people would consider better spent on researching a *cure* for cancer. (Moreover, the largest animal study ever done—and this was very much an exceptional case—involved 24,000 mice and required an unprecedented logistical effort to carry out.[11]) Such a study is obviously impossible.

These three instances of the FDA's posture towards the cyclamate evidence by no means exhaust (but do accurately exemplify) the scientific curiosities in this decision. Suffice it to say that there are other results the FDA finds suggestive though not conclusive, but nonetheless—or rather, given its new interpretation of the general safety provision, therefore—sufficient to deny Abbott's petition. The statistical and biological significance of these depends on such technicalities (which you need not understand) as whether or not one can validly apply the Armitage test for linear trend to groups with N less than five, or whether the Bonferroni multiplier should be applied to the calculated p-values, or how tumors should be grouped for statistical analysis (liver tumors together with lung tumors? all lymphosarcomas together or only those seen in selected organs? all tumors of all kinds seen in the test?), or whether different generations within one experiment should be grouped together rather than analyzed separately, and so on. The evidence on which the 1980 cyclamate ban is based, far from being solid, is a stew of shifting, unreproduceable, and ephemeral "findings" all embedded in a startlingly *ad hoc* interpretation of scientific methodology.

Discretion is a two-edged sword

The magnitude of the change in the FDA's implementation of food safety policy is shown by contrasting the FDA's postures in

[10] National Cancer Institute, Division of Cancer Cause and Prevention, *Report of the Temporary Committee for the Review of Data on Carcinogenicity of Cyclamate*, Appendix V (1976): 55-59.

[11] N.A. Littlefield, et al., "Effects of Dose and Time in a Long-Term, Low-Dose, Carcinogenesis Study," *Journal of Environmental Pathology and Toxicology* 3 (1980).

the cyclamate case with the earlier saccharin and nitrite cases. The evidence for carcinogenicity in the case of saccharin, as weak as it is, is far more secure than it is for cyclamate.[12] Saccharin at least does replicably induce a small but statistically significant (at the $p \leq .05$ level) increase in non-lethal bladder tumors in second generation rats (i.e., rats whose dams were maximally dosed with saccharin throughout the conception, gestation, and nursing of their offspring). This has been shown three times, and there are no tests in which this effect has not been seen. An occasional tumor or two is seen in first generation rats as well, which could certainly be interpreted as being "consistent with a small treatment effect." Both of these features would be adequate under the new reading of the general safety clause to "raise a question" that, in the absence of a convincing rebuttal in the form of a mega-rodent test, would support a ban. This reasoning was not employed earlier; rather, the Delaney clause was invoked, and all public debate focused there. And as for nitrite, the fact that it forms a class of carcinogenic compounds called nitrosamines upon metabolism by the body could certainly be construed under the new policy as "raising a question." Again, this reasoning was not used earlier; instead, the Delaney clause was invoked.

This progression in policy, provided it stands, does make the Delaney clause entirely redundant (as Peter Hutt presciently foresaw), since any set of carcinogenic data strong enough to sustain a ban under Delaney would necessarily suffice to "raise a question" under the new interpretation of the general safety clause. In fact, one cannot envision a situation in which the FDA would have to rely *only* on Delaney to withhold approval in a future food additive decision. Those who have been urging Delaney's repeal have been outmaneuvered: For under the Delaney clause, the burden of proof at least was on those who made an affirmative declaration that a substance did cause cancer in animals. This need no longer be done. All that is needed now is a small reason to raise a suspicion that a substance *might* cause cancer, and then the burden of proof is on the petitioner to demonstrate that it does not. It is possible that one day we will look back on the Delaney clause with nostalgia.

Indeed, the critics of Delaney have been hoisted with their own petard. For it is the very "discretion" that was supposed to be a virtue of decision-making under the general safety clause—we were

[12] W.R. Havender, "Ruminations on a Rat: Saccharin and Human Risk," *Regulation* (March/April 1979): 17-24.

told it would allow the agency leeway to weigh benefits against risks, and to take the exaggerated conditions of animal cancer tests into consideration when estimating human risks—that has permitted the FDA unexpectedly to revise its application of this provision in a much *more* stringent and unreasonable manner. Discretion, it seems, is a two-edged sword.

It is possible, however, that the FDA's new policy may not endure. Several features suggest that it will prove impossible to generalize. For one, a $100 million animal cancer test cannot generally be required of every petitioner seeking approval for a food additive about which skimpy doubts can be raised. For another, "creative" statistics will drown the agency with "carcinogens" needing regulation, a substantial fraction of which are probably wrongly so classified. A hint of what may be in store is given by the fact that precious few *foods* could withstand the scrutiny of the FDA's new policy. Sugar, for example, has caused a statistically significant incidence of tumors (at the $p \leq .05$ level) in at least one test,[13] as has pepper,[14] as has Vitamin D,[15] as has a mixture of egg yolks and milk.[16] Perhaps the toughest near-term test of the new policy will be the saccharin decision which is due next summer (if it is not postponed again). Currently, saccharin is the only artificial sweetener approved for use in soft drinks, so the public will no doubt protest vigorously against any attempt to use this policy to ban saccharin. If it turns out that the new FDA policy cannot be so generalized, then it is unlikely that the current cyclamate denial can stand, either.

The cyclamate episode clearly points up the softness of the wording of the general safety clause. The FDA nominally interpreted it as requiring a petitioner to supply "proof of a reasonable certainty that no harm will result from the proposed use of an additive," which sounds perfectly sensible. But there is no clear meaning about what "reasonable certainty" of "no harm" means in operational terms, or what rules should apply for reaching a judgment.

---

[13] Hoffman Laroche Co., Ltd., "Tumorigenicity and Carcinogenicity Study with Xylitol in Long-Term Dietary Administration to Mice," Study Number HLR 25/77774 (January 30, 1978), prepared by Huntingdon Research Center, Huntingdon, Cambridgeshire, England. Available from the U.S. Food and Drug Administration, Rockville, Maryland.

[14] J.M. Concon, D.S. Newburg, and T.W. Swerczek, "Black Pepper (*Piper nigrum*): Evidence of Carcinogenicity," *Nutrition and Cancer* 1 (1979): 22-26.

[15] G.H. Gass and W.T. Alaben, "Preliminary Report on the Carcinogenic Dose Response Curve to Oral Vitamin D_2," *IRCS Medical Science* 5 (1977): 477.

[16] D. Nelson et al., "Hepatic Tumors in Rats Following Prolonged Ingestion of Milk and Egg Yolk," *Cancer Research* 14 (1954): 441-445.

The cyclamate decision shows how unbounded these terms can be, particularly when "questions" and suspicions, rather than facts, are enough to deny a petition. That cyclamate has been extensively tested, and that no secure, repeatable finding of cancer has been established, would seem to supply, by any normal application of scientific inference, a "reasonable certainty" that no harm would result from the intended uses of cyclamate. But as we have seen, the FDA was not governed by normal scientific criteria. With so little constraint on what the FDA can conjure up to "raise a question," and with the burden of proof so one-sidedly placed on the petitioner, the potential for arbitrariness is great.

This decision, then, makes the need for reform urgently clear. With firm direction from the top, the FDA could reform itself, and it could do so without delay (for example, by repudiating these tactics in its evaluation of Abbott's 1982 petition). But even this sort of reform might not be sustained over time: The law on the books is clearly soft, so policy at the FDA is likely to go through wide swings as personnel and presidential administrations change. Only a change in the law—not merely the Delaney clause but the general safety clause as well—will be lasting.

8

The Chemical Warfare Evaluation

A Case Study of Evaluation in Action

March 18, 1982

Mr. Charles A. Bowsher
Comptroller General of the United States
General Accounting Office
Room 7026
441 G Street, N.W.
Washington, D.C. 20548

Dear Mr. Bowsher:

The Subcommittee on International Security and Scientific Affairs is preparing for hearings on chemical warfare. Information describing deterrence against use of chemical weapons, Soviet and U.S. chemical warfare capabilities, binary chemical weapons, and disarmament would be very valuable to the Subcommittee in preparing for hearings. More specifically, the Subcommittee is interested in obtaining information on the fifteen questions presented in the attachment to this letter.

Discussion between my Staff Director, Ivo Spalatin, and staff from your Institute for Program Evaluation indicated that the Institute would be able to provide us with information in time for our hearings. It would be most helpful to us if the Institute staff could synthesize and assess the currently existing information on these fifteen questions and brief us on what they have learned no later than April 7, 1982 with a written report to follow as soon as possible thereafter.

The text in this chapter is excerpted from the following sources: U.S. General Accounting Office, *Chemical Warfare: Many Unanswered Questions,* Report to the Committee on Foreign Affairs, U.S. House of Representatives of the United States; *Congressional Record,* U.S. House of Representatives, June 15, 1983, H3990–H4011. Colman McCarthy, "Defending Nerve Gas," *Washington Post,* July 30, 1983, p. A23; and Fred Hiatt, "Pentagon Again to Seek Funding for Nerve Gas," *Washington Post,* January 18, 1984, p. A15. The *Washington Post* articles are reprinted by permission of the publisher.

Thanking you in advance for your cooperation in responding to this request, I am

Sincerely yours,

Chairman

CJZ:isj

attachment

APPENDIX I

APPENDIX I

Attachment

Questions for Analysis Based on Existing Information

Topic 1. Deterrence.

1. What are the different ways to achieve deterrence against use of chemical weapons and which way has the U.S. chosen to pursue it?

Topic 2. Soviet Capability

(2) What is the nature, extent, and condition of the Soviet stockpile?

(3) To what extent do the Soviets have chemical weapons production/research facilities?

(4) What chemical weapons delivery systems do the Soviets have?

(5) What is the Soviet CW defensive capability?

Topic 3. U.S. Offensive Capability

(6) What is the current U.S. chemical warfare doctrine?

(7) How has the needed U.S. stockpile size been determined?

(8) Are munitions in our current stockpile compatible with delivery systems introduced or being introduced in Europe?

(9) What other options, besides the binary, exist for modernizing our chemical warfare capability?

Topic 4. Binary Chemical Weapons

(10) Will the binary program affect the U.S. ability to achieve both a CW denial and punishment capability?

(11) How would deployment of binary munitions affect military operational flexibility?

(12) How do binary and unitary munitions compare in toxicity?

(13) How do unitary and binary weapons compare in safety?

(14) To what extent will binaries increase the risk of proliferation?

Topic 5. Disarmament

(15) What are the verification problems with regard to a chemical weapons ban?

Chemical Warfare: Many Unanswered Questions
U.S. General Accounting Office, Comptroller General

CHAPTER 1

INTRODUCTION

Claiming Soviet superiority in all aspects of chemical warfare as well as the failure of years of bilateral negotiations aimed at banning chemical weapons, the U.S. Department of Defense (DOD) requested a fiscal year 1983 appropriation of $705 million from the Congress for its chemical warfare program. Although this figure is up sharply from the 1978 chemical warfare budget of $111 million and the 1981 budget of $259 million, it does not tell the whole story of the effort to overhaul the U.S. chemical warfare program. DOD has a 5-year plan for increasing the U.S. chemical warfare capability from 1983 to 1987, and its estimate of the total price tag is $6 billion to $7 billion. Other estimates run up to $14 billion for the next decade. With billions of dollars at stake in an area where emotions run high, controversy naturally has been acute. As a result, expectations about the proposed plan range from spending billions of dollars unnecessarily or even harmfully to endangering the security of the United States and its European allies if the money is not spent.

We were asked by the House Committee on Foreign Affairs to look into some of the issues that underlie the current debate on the need to increase the U.S. chemical warfare capability. In this report, therefore, we assess and synthesize the information that is available for addressing four issues of particular concern to the Committee:

--the different ways of deterring chemical warfare,

--the comparability of the United States and the Soviet Union in chemical warfare capability,

--the options for modernizing the present U.S. chemical warfare system, and

--the likely effects of modernization on the prospects for disarmament.

We describe the nature and extent of the information that is available on each topic, determine the best sources for addressing each topic, and discuss the general level of confidence we have in the findings. We also identify gaps and inadequacies in our knowledge and raise questions that remain to be addressed. Given the considerable number of unknowns that continue to exist in this area, refining and pinpointing the precise nature of these questions was a major effort.

REVIEWING THE CHEMICAL WARFARE DEBATE

Chemical warfare uses weapons that disperse incendiary mixtures, smoke, or irritating, burning, or asphyxiating gas.

Chemicals have been used in warfare throughout history, but the participants of World War I witnessed the first and last large-scale use of chemicals on the battlefield. During that encounter, the Allied forces, in an effort to build up world opinion against Germany, embarked on a campaign against chemicals, calling their use "barbarous" and "inhumane." The campaign contributed to a public objection to chemical warfare that still exists today.

The moral revulsion to chemical warfare that arose in World War I led to the Geneva Protocol of 1925, which prohibits the use of asphyxiating, poisonous, and other gases in war. The Protocol also banned biological (or bacteriological) warfare, even though biological weapons had not been used in any significant sense. Most signatories of the Protocol added a provision that they would not be bound by it if an enemy used gas or biological agents against them first. Many gases are stockpiled today, even though the stockpiling of biological weapons was banned by international agreement in the 1972 biological warfare convention.

While there have been numerous allegations that chemicals have been used in international conflicts over the past 6 decades, few have been substantiated. In all the substantiated cases, lethal chemical weapons were used against an enemy known to be deficient in antigas protective equipment or retaliatory capability.

The United States maintains the ability to retaliate in kind should an enemy use chemical weapons first. However, partly because of an open-air test accident that killed more than 6,000 sheep, and partly because of public concern about the effect on the environment of transporting and disposing of chemical weapons, legislation was enacted in 1968 that restricted the movement of chemical munitions and agents in peacetime and the development of new weapons where open-air testing is required. At about the same time, there was also a wave of adverse public opinion over the use of riot control agents (tear gas) and herbicides during the Vietnamese War, contributing further to the deemphasis of U.S. chemical warfare capabilities. The United States has produced no chemical weapons of any kind since 1969 and has been restrained from testing its stockpile since 1968. Many believe that the U.S. chemical warfare capability has become inadequate over this rather lengthy period of time.

Meanwhile, the Soviet Union has been under no similar restrictions. Also, some have charged that the Soviets have violated the international agreement not to develop, produce, or stock biological weapons and that they have encouraged and abetted the use of chemicals in Southeast Asia and Afghanistan.

It is against this background that the need to increase the U.S. chemical warfare capability is being debated. We have not

been silent on the subject, having produced six reports since 1977 on lethal chemical warfare. In 1977, we looked at the condition of the U.S. stockpile of lethal chemical munitions and agents (GAO, 1977c), and in 1981 we reviewed the status of DOD's implementation of our recommendations concerning the stockpile (GAO, 1981).[1] Also in 1977, we examined the U.S. lethal chemical munitions policy in terms of issues facing the Congress (GAO, 1977b), and in 1979 we updated that report with a fresh look at the status of issues facing the Congress (GAO, 1979). Again in 1977, we reviewed U.S. chemical warfare defense, looking at both readiness and costs (GAO, 1977a), and in 1982 we again investigated the readiness of U.S. forces, equipment, and facilities to survive and recover from a chemical attack (GAO, 1982). In the present report, we draw upon our earlier reports, especially our 1982 readiness review, but with considerably different objectives, scope, and methodology.

OBJECTIVES, SCOPE, AND METHODOLOGY

The House Committee on Foreign Affairs specifically asked us to synthesize and assess existing information on questions related to (1) deterrence against the use of chemical weapons, (2) Soviet and U.S. chemical warfare capabilities, (3) U.S. chemical warfare modernization, and (4) the likely effect of modernization on the prospects for disarmament. Debates about chemical warfare usually discuss one or more of these topics. We analyzed and synthesized information on chemical warfare to determine what is known about it, the confidence we can have in this information, and the gaps and inadequacies that remain. Thus, our objective is to assess and synthesize the rapidly accumulating information on chemical warfare relevant to these topics.

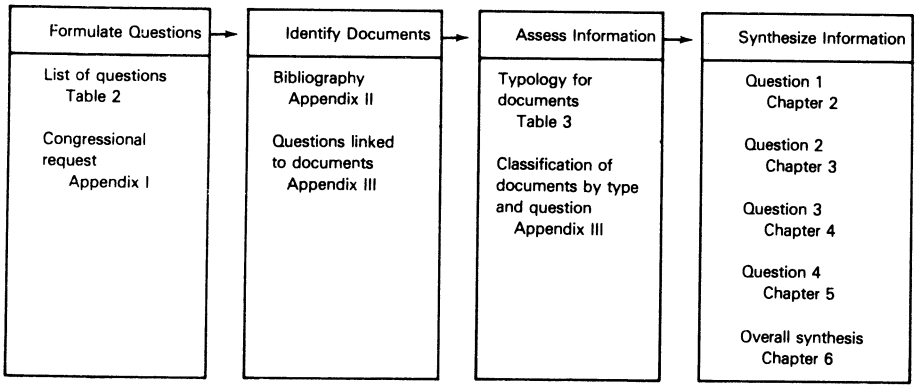
Our method with regard to documents has had four steps. First, we developed study questions on chemical warfare, basing them on the Committee's request and organizing them in a logical sequence. Second, we identified and collected our information sources (a term that we use interchangeably with the word "document"). Third, we assessed the information, classifying each source according to the study questions it addresses and the type of information it presents. When it was appropriate, we also reviewed the overall quality of the information. Fourth, in the synthesis, we determined which information is best for addressing each question, indicated the general degree of confidence that can be attributed to the findings, and identified remaining information gaps or inadequacies. In table 1 on the next page, we present an overview of our methodology and link it to the report's contents.

---

[1]Interlinear bibliographic citations are given in full in appendix II. The names of authors that are agencies are abbreviated, as here.

Table 1

An Overview of the Methodology and a Map of This Report



Along with this effort regarding documentation, we undertook several supplementary and complementary activities. We conducted interviews with a wide range of experts. We attended briefings and congressional hearings on issues related to chemical warfare. We performed these activities throughout the duration of the project. We used the results of these efforts to inform each step of our review. The review was performed in accordance with generally accepted government audit standards.

Formulating the study questions

Developing the questions of interest to the Congress on chemical warfare, we began with the four basic questions in the chemical warfare debate: (1) How is deterrence against the use of chemical weapons achieved? (2) How do the United States and the Soviet Union compare in their chemical warfare capabilities? (3) How can the United States modernize its present chemical warfare system? (4) What are the likely effects of modernization on the prospects for disarmament? As we show in table 2, we divided each question into several others. While the list is not exhaustive, each question is undeniably important to a comprehensive analysis of the chemical warfare debate. In the table, we have marked the specific questions the Committee asked with an asterisk. The Committee's letter is reprinted in appendix I.

Identifying the information sources

The controversy surrounding chemical warfare is reflected in the tremendous amount of popular and other literature that

Table 2
Chemical Warfare Questions and Subquestions [a]

QUESTION	SUBQUESTION
1.0 How is chemical warfare deterred?	1.1 What is a credible deterrence capability? 1.2 What are the different ways of deterring chemical warfare?* 1.3 How has the United States chosen to pursue deterrence?*
2.0 How do the United States and the Soviet Union compare in chemical warfare capability?	2.1 What are the U.S. and Soviet doctrines governing the use of chemical weapons?* 2.2 How does the U.S. chemical stockpile compare with the Soviet Union's and how is stockpile need determined?* 2.3 How do the U.S. and Soviet chemical warfare delivery systems compare?* 2.4 How do the United States and the Soviet Union compare in defensive equipment and personnel?* 2.5 How and to what extent have the United States and the Soviet Union prepared for implementation?*
3.0 How can the United States modernize its chemical warfare system?	3.1 What factors are necessary for modernization? 3.2 What are the alternatives to binaries?[b]* 3.3 Do binaries have substantial advantages over unitaries?*
4.0 How does modernization affect the prospects for disarmament?	4.1 How successful have chemical warfare disarmament efforts been? 4.2 What are the verification problems in banning chemical weapons?* 4.3 What implications does modernization have for disarmament?*

[a] Questions marked with an asterisk (*) were specifically raised for review by the House Committee on Foreign Affairs.

[b] Instead of containing actual nerve gas, binary weapons contain two relatively nontoxic chemicals in separate canisters that are allowed to mix and react only when the munition is being delivered to its target (or being readied for delivery), the chemical combination being a nerve gas.

has been written on it. There are literally hundreds, if not thousands, of newspaper items and editorials, popular magazine articles, technical journal articles, books, studies, and reports on chemical warfare. It was clear at the outset that our review of the literature could not be exhaustive, but it was less clear whether we wanted to be comprehensive or representative in our readings, how we would know whether we had been comprehensive or representative, and whether we would vary our approach for the different types of information.

Given our study approach and our purpose of separating fact from fiction, we focused on the information sources that would be the most likely to contain either original data or original arguments about chemical warfare. Therefore, sources such as newspaper items and popular magazine articles are underrepresented in our sample. We concentrated on articles in military and technical journals and on research studies and reports. While we looked at testimony in congressional hearings on chemical warfare, we were more interested in reviewing the sources on which the testimony had been based. We examined classified literature in addition to open literature. Our use of intelligence data in assessing Soviet capability is described in chapter 3.

To identify the relevant literature, we used chemical warfare bibliographies and reference lists as we encountered them, searched the literature, and conducted interviews. We reviewed the chemical warfare files of the Congressional Research Service and asked the Defense Technical Information Center, the Defense Logistics Studies Information Exchange, and SCORPIO to search the literature. We interviewed representatives of the U.S. Army's nuclear and chemical directorate and representatives of the Office of the Secretary of Defense and the Arms Control Disarmament Agency.

Following these procedures, we identified a large number of technical reports and articles on chemical warfare. The Defense Technical Information Center search, for example, provided a list of about 250 unclassified technical reports on chemical warfare, although we did not review them all. If a report concentrated on an area that was not a focus of one of our questions, such as demilitarization, we did not review it. If we had several recent references on a topic, we did not review all the older references. When we followed up on reference lists, we concentrated on items that were cited frequently and on items that appeared to focus on study questions for which we had limited information. Thus, we attempted to be comprehensive in our search of the literature and selective in our review and analysis. We completed our selection of documents in May 1982.

We relied on expert opinion to confirm that the final list of references that we reviewed does in fact represent the literature available for addressing the study questions. Toward this end, we asked five experts to review a draft of our bibliography

and indicate additional sources that contain factual information or arguments not accounted for in it. The experts, who take different positions in the debate on chemical warfare modernization, were Niles Fulwyler (then head of the U.S. Army's nuclear and chemical directorate), Amoretta Hoeber (Principal Deputy Assistant Secretary of Research and Development for the U.S. Army), Matthew Meselson (professor at Harvard University), John Erickson (professor at the University of Edinburgh), and J. Perry Robinson (professor at the University of Sussex). In general, these experts confirmed that our bibliography is representative, and we added references suggested by their reviews.

The bibliography of documentary sources we used to address the study questions is in appendix II. We have arranged the references in the following categories: reports by congressional agencies and organizations, military and technical journal articles, other military publications, publications by other organizations, conference papers and testimony, and books by individuals.

Assessing the information

Once we had identified the sources of information for each question, we classified them by type and by the questions they addressed. Then we made judgments about the quality of the information according to a set of assessment criteria. Later in the synthesis step, these judgments about type and quality helped us determine our confidence in the information. This, in turn, determined whether and how we used each information source.

Classifying information sources by type and by questions addressed

We classified each document we reviewed by type and by the questions it addressed. We defined eight types, which we have listed in table 3 on the next page. We also classified each document by the four study questions and their subquestions listed in table 2. In appendix III, we have displayed this classification of the information sources. Each document is classified by only one type but shares several questions with other documents.

We found that the types of information that are available differ considerably. For example, some reports give accounts supporting a particular stance on a chemical warfare issue and raising major points of controversy. Others merely identify the points of controversy in a neutral way, attempting not to take a stance on any issue. Still others describe complex simulations of scenarios of real-life situations, and yet others report on tests and evaluations. For documents that have mixed characteristics, we selected the predominant characteristic for their classifications.

Table 3

Chemical Warfare Document Types and Their Definitions

Type	Definition
Historical	Provides a historical account of the subject.
Opinion	Presents the beliefs of individuals who have special knowledge about the subject and only one side of an argument.
Issue review	Raises major points of controversy but does not attempt to resolve the controversy and supports no one argument.
Issue analysis	Raises major points of controversy and seeks to resolve the controversy.
Policy study	Evaluates alternatives systematically according to stated criteria and, in some cases, identifies a preferred alternative.
Simulation	Reports on the examination of a problem not by direct experimentation but by structured, frequently computer-based, gaming techniques.
Documentary	Presents expository "eye witness" material, often secondhand.
Test and evaluation	Collects and examines expository material critically by means of various structured procedures such as content analyses, case studies, surveys, field experiments, and intelligence procedures.

Judging the information quality

Next, we made judgments about the quality of the reasoning in each document and the purported facts pertaining to chemical warfare issues. Because so much of the information on chemical warfare is not empirical and, therefore, not subject to the usual questions about the soundness of methodology, we developed an exploratory set of criteria for our assessment of the quality of information. We list these criteria in table 4. Their applicability differs from source to source, and we made no attempt to use each criterion in every case. We made no effort to "score" the information sources on their quality or to verify the consistency of different reviewers in meeting our criteria. In short, we used the criteria as guides to assessing information rather than rigorously rating its quality.

Synthesizing the information

Our last step was to identify and integrate the best sources of information for addressing each question, to determine the overall degree of confidence in the answer to the question, and to identify remaining gaps and inadequacies. All else being equal, we judged test and evaluation information to be superior to other types of information. If we had "good" test and evaluation information, we relied on it and did not necessarily use sources of other types, except in briefly presenting the pertinent arguments. For questions for which we did not

Table 4

Document Assessment Criteria and Their Definitions

Criterion	Definition
Bias	To what extent is the author or source potentially involved in chemical warfare outcomes? Is the source a lobby organization for the military? All else being equal, an independent, uninvolved source is more credible than a potentially biased one.
Values	To what extent does the author make value judgments? How closely do values underlie the argument? To what extent do values rather than logic constitute the argument? The more the document substitutes values for logic, the less credible it is.
Assumptions	Are the assumptions explicit or implicit? Are they reasonable or unreasonable? What support is there for them? A document based on unstated, "shaky," or false assumptions loses credibility.
Logic	To what extent is the logic flawed? The tighter the logic, the more credible the document.
Facts	To what extent are facts the basis for the arguments? To what extent are the sources for the facts cited? A document that is based on facts that have been or can be verified is more credible than one that is not.
Competing alternatives	Does the argument account for competing strategies, hypotheses, or courses of action? Is a case made for rejecting alternatives? An argument for which competing alternatives have been analyzed has more credibility than one for which they have not.
Political and operational feasibility	To what extent does the argument take into account the political and operational feasibility of what is being recommended? Could the recommended course of action be implemented?

have test and evaluation information, we judged simulation information to be superior to other types, all else being equal. We followed the same procedure in relying on policy studies. We made no similar distinctions for relying on the other information types. If we had information consisting of only arguments, we used our assessment criteria to identify any weaknesses in them.

CHAPTER 6

QUESTIONS ON U.S. CHEMICAL WARFARE CAPABILITY, SUMMARY OBSERVATIONS, AND AGENCY COMMENTS AND OUR RESPONSE

The controversial chemical warfare issue has been raised by the present Administration's plan to modernize the nation's chemical warfare capability. In the 5 years 1983-87, the U.S. Department of Defense anticipates spending between $6 billion and $7 billion to upgrade the U.S. retaliatory and defensive chemical warfare capabilities. With this sum of money at stake, the results of the proposed modernization program range from spending billions of dollars unnecessarily, or even harmfully, to endangering U.S. national security and that of its allies if the money is not spent.

The House Committee on Foreign Affairs asked us to synthesize and assess the nature, extent, and quality of information available to answer the following specific questions:

1. How can chemical warfare be deterred?

2. How do U.S. and Soviet capabilities compare?

3. How can the United States modernize its chemical warfare system?

4. How will modernization affect the prospects for disarmament?

e current debate on the need to increase the U.S. chemical rfare capability usually revolves around one or more of these estions.

Our purpose in synthesizing the information on chemical rfare was to determine (1) what is known about chemical warre (the facts and other data and the analyses that are availe to support various positions), (2) the general confidence t can be placed in that information, and (3) the gaps and dequacies in it. Toward this end, we reviewed and assessed ssified and unclassified chemical warfare literature, focus- on military and other technical documents and on empirical dies. Experts representing different positions on the chem- l warfare modernization debate helped us establsh that we had luded all major references in our review, indicating sources n additional factual information or arguments we had not eady identified. Despite the technical and empirical focus our review, we found that the arguments in most references based on belief. Most of the factual information is pported by citations. Few simulations or actual test evaluation studies exist.

We found a multitude of unanswered questions related to chemical warfare modernization. The number of unresolved issues, both broadly and narrowly defined ones, is large. Some questions have been partly and inadequately addressed; others have apparently not even been raised. The general picture is that the chemical weapon system is not perceived as a credible deterrent, little is known about its functioning or its usefulness, and a large amount of money is being sought for it. We are particularly concerned that so many questions remain unanswered since the United States has maintained chemical weapons for so many years and since we have issued a long series of reports identifying deficiencies in U.S. chemical warfare retaliatory and defensive readiness.

HOW CAN CHEMICAL WARFARE BE DETERRED?

The concept of deterrence is generally premised on dissuading hostile actions through the perception of the will and the ability to inflict unacceptable consequences on a potential adversary. Deterring chemical warfare is premised on the same concept, except that analysts differ, according to their individual perspectives on tactical warfare and their views of the utility of chemical weapons, on what specifically is most likely to be able to inflict, and to be perceived as able to inflict, unacceptable consequences. Chief among the views are that the threat of tactical nuclear attack is a credible chemical warfare deterrent and that a chemical retaliatory capability is necessary for deterrence.

The literature also presents the essential elements of retaliatory, or offensive, and defensive chemical warfare capabilities. These elements include (1) having a well-developed doctrine, (2) maintaining a sufficient stockpile of weapons, (3) having delivery systems for the weapons, (4) having adequate and appropriate defensive equipment and personnel, and (5) being able to implement the system. The fifth element includes training, production facilities, and deployment logistics.

Empirical evidence of the significance of these elements in establishing a credible chemical warfare deterrent is scant. The literature suggests that lack of chemical warfare assimilation by the military, legal and moral proscription, and fear of retaliation played important parts in forestalling an extensive use of chemicals in World War II. Historical analyses of alleged uses of chemical weapons suggest that both the ability to defend against an enemy using chemical weapons and the ability to launch a retaliatory attack on the enemy (although not necessarily with chemicals) are important components of deterrence.

The literature identifies three broad policy options for chemical warfare deterrence. Emphasizing different elements of capability, these are policies on arms control, weapons, and defense. Policies emphasizing weapons and defense

call for some offensive or retaliatory capability, whether nuclear or chemical, yet all three require a strong protective posture. The emphasis on weapons differs from the emphasis on defense by calling for a major conventional, nuclear, or chemical warfighting capability; the emphasis on defense includes a limited chemical retaliatory capability, sufficient only to force the enemy into chemical protection.

The issues that are prominent in discussions of these three policy options are (1) the extent to which the use of chemical weapons could be rendered ineffective if protective shelter, clothing, and equipment were adequate to defend against them, (2) the extent to which protective clothing and equipment severely degrade military efficiency on both sides, and (3) the likelihood, necessity for, and utility of a verifiable ban on chemical weapons. Those who argue that strong defensive measures or the threat of tactical nuclear retaliation deter the initiation of chemical warfare generally look favorably on arms control as a way of achieving a chemical weapons ban. Those who disagree with this view and argue for the importance of imposing an equal degradation of performance on an enemy often favor retaliation-in-kind as a chemical warfare policy.

The literature shows that the United States has consistently declared the policy of retaliation-in-kind. Given the existence of the U.S. chemical weapons arsenal and current proposals to upgrade both its retaliatory and its defensive capabilities, the United States can be seen as having adopted either a policy of weapons emphasis or a policy of defense emphasis with limited retaliatory potential. Some argue, however, that U.S. policy should be characterized as emphasizing arms control, since they believe that the United States has been unilaterally disarming.

HOW DO U.S. AND SOVIET CAPABILITIES COMPARE? HOW CAN THE UNITED STATES MODERNIZE ITS CHEMICAL WARFARE SYSTEM?

Whether emphasizing defense with limited retaliatory capability, weapons, or arms control, U.S. chemical warfare deterrence policy requires both chemical retaliatory and defensive, or protective, capabilities. Retaliatory and defensive capabilities consist of many elements, the basic ones listed in the literature being doctrine, stockpile size and composition, delivery systems, defensive equipment and personnel, and implementation. We reviewed the literature to determine U.S. and Soviet status on these elements of capability and investigated DOD's modernization program in light of the current U.S. status.

The literature generally agrees that the United States lacks a credible chemical warfare deterrent in terms of the capability elements. That is, perceptions and data agree that the

United States does not have the means or the ability to respond effectively to a chemical attack. In contrast, the literature generally reflects the perception that the Soviets are highly able to wage chemical warfare. However, open sources and classified reports contain only limited information to support the various assertions about specific levels of Soviet capability.

As for defensive capability, we found a body of facts and supporting evidence that the Soviets have built a strong ability to defend against nuclear, biological, and chemical warfare. We found U.S. inadequacies well-documented with respect to the ability to retaliate and defend in a chemical warfare environment. The most favorable comparison for the United States is in individual protection, but even here the literature describes unresolved problems with the U.S. protective suit and mask.

The question that is implicit in DOD's modernization plan is whether or not modernizing the U.S. chemical warfare capability will improve deterrence. Modernizing a chemical warfare system requires (1) adequate information on the several alternative ways of modernizing, (2) a strong rationale, based on reliable data, for selecting one alternative rather than another, and (3) comprehensive and integrated plans to coordinate the improvement of capability in a variety of elements--among them doctrine, stockpile, delivery systems, defensive equipment, and implementation. In our review of existing information on DOD's modernization program, we did not find convincing evidence that these three requirements have been adequately met.

Doctrine

The following statements are supported by credible information:

--The Soviets are perceived as having a well-developed and clearly articulated offensive chemical warfare doctrine.

--The United States is attempting to develop chemical warfare doctrine.

--There are many combat scenarios in which chemical weapons could be used against U.S. forces and there is no comprehensive U.S. doctrine for sustaining combat operations in many such situations.

Information on the following issues is sparse or inadequate and we are unable to draw conclusions about them with a minimum level of confidence:

--whether the Soviets do have a well-developed and clearly articulated offensive chemical warfare doctrine;

--whether the major obstacles to the development of U.S. chemical warfare doctrine have been identified and whether they can be overcome;

--whether procuring binary weapons will complicate efforts to develop retaliatory doctrine;

--whether U.S. retaliatory doctrine can adequately address the following: the effects of combining chemical weapons and improved conventional munitions in warfare, the likelihood of inflicting casualties on well-protected Soviet troops, the likelihood that area-denial tactics can be pursued given Soviet collective protection capabilities, and the likelihood that U.S. forces can acquire targets most susceptible to chemical attack without causing unacceptable civilian casualties;

--whether in the immediate future U.S. defensive doctrine should be made to reflect the lack of adequate collective protection in combat vehicles and stationary shelters, vehicle and equipment decontamination facilities, and remote-area sensing and alarms.

Stockpile

Regarding the stockpiles of munitions held by the United States and the Soviet Union, our review finds substantial evidence of the following:

--The United States maintains chemical stockpiles in arsenals within the United States, in a depot on Johnston Island in the Pacific, and in Europe.

--Most U.S. munitions are short-range artillery projectiles; the arsenal contains some chemical-filled bombs

--The stockpile in Europe contains

--The total size of the U.S. chemical stockpile and its condition are not precisely known; estimates range consistently from agent tons to agent tons.

--There are approximately agent tons of lethal chemicals in bulk storage in the U.S. stockpile; in addition, there are between agent tons of serviceable or repairable munitions.

--The size, mixture, and deployment of the Soviet stockpile is ; guesses about its size range from agent tons to agent tons, indicating the of knowledge in this area.

The information that is available is inadequate to support conclusions on the following chemical stockpile questions:

--whether comprehensive logistics plans exist for timely deployment of chemical weapons to NATO;

--whether the chemical weapons in Europe are enough to degrade Soviet forces to the same level NATO forces can expect to be degraded;

--what tonnage need in chemical munitions has been estimated for theaters other than NATO's central region;

--the extent of preventative and rehabilitative measures being taken to preserve the existing chemical weapons stockpile;

--whether there is a sound basis for determining a stockpile of munitions that effectively meets the Soviet threat and takes advantage of any of its vulnerabilities.

Delivery systems

Analysis of the literature shows that evidence supports the following assessments of chemical warfare delivery systems:

--

--

--The Army is not following recommendations to produce binary bombs first, rather than artillery projectiles, in order to acquire a long-range capability.

--The Soviet chemical warfare delivery means are virtually unknown, even though many sources cite them as consisting of missiles, rockets, bombs, aerial spray tanks, and artillery.

We found limited information or none on the following delivery issues:

--U.S. progress in developing a long-range surface-to-surface chemical warfare delivery capability;

--U.S. progress in developing short-range chemical warfare delivery means

--whether air-delivered chemical munitions are practicable in the face of Soviet anti-aircraft capabilities;

Defensive equipment

The information on defensive systems supports the following assessments:

--Tests have shown that the U.S. protective suit causes less heat stress than Soviet suits.

--U.S. suits are flammable, cannot be laundered, and must be disposed of when they are saturated.

--U.S. protective masks need a flexible lens and external filters that are easy to change.

--The United States lacks an adequate chemical sensing and alarm capability.

--The United States has limited collective protection capabilities for vehicles; the Soviets have seriously pursued collective protection.

--The United States lacks efficient equipment for the large-scale decontamination of troops, weapons, and vehicles; Soviet forces appear to have a substantial decontamination capability.

--The United States planned to have 7,400 chemical defense specialists by fiscal year 1982; the Soviets have been estimated as having between 50,000 and 100,000 troops dedicated to nuclear, biological, and chemical defense.

Our knowledge is less certain, or nonexistent, on the following points:

--plans for and progress in fitting various existing U.S. combat vehicles for collective protection;

--the operability of Soviet collective protection systems in combat vehicles, as planned, under combat conditions of high mobility and repeated weapon firings.

Implementation

In examining implementation capabilities, we found credible evidence supporting the following statements:

--

the United States does not have plans for deploying binary munitions in Europe.

We identified very little information on implementation issues such as whether the operational characteristics of binaries (such as their mixing time) require special training or doctrinal considerations.

In essence, the findings of the literature on the five elements of doctrine, stockpile, delivery systems, defense equipment, and implementation can be summarized as follows:

1. The United States does not have a chemical warfare doctrine, yet DOD is preparing to modernize the chemical weapons arsenal. There is evidence that the Soviets have developed a defensive doctrine for integrated conventional, nuclear, and chemical warfare scenarios; little is known about Soviet offensive doctrine.

2. The precise size and condition of the U.S. stockpile are not known, but it is known that and no long-range surface-to-surface capability at all. Little is known about the size and mixture of Soviet chemical munitions.

3. There appears to be no U.S. plan for developing a long-range surface-to-surface chemical weapons delivery capability. The Soviets are assumed to have every conceivable means of delivering chemical warfare agents, but

4. The United States has put into the field relatively good protective suits but needs to improve decontamination capability, remote area detection, collective protection in vehicles, and stationary shelters, with remote sensing and alarm capability being seen as presenting an especially critical deficiency. The Soviets have made extensive chemical warfare defensive preparations in all areas--decontamination, detection, individual and collective protection.

5. The United States has not pursued initiatives with NATO allies that would allow the forward deployment of binary weapons,

Binary alternatives

Alternatives to the procurement of binary weapons are identified and discussed in the literature. Most commonly it is

argued that the United States has a stockpile of chemical weapons that is sufficient for any likely retaliation-in-kind requirement. The DOD position is that the present stockpile is deficient in both size and mixture of weapons and that only producing binaries will rectify this situation. We find that present knowledge is not adequate either to refute or to support the claims and counterclaims in this debate.

We searched for evidence that indicates that the new binary weapons will give DOD substantial advantages it does not have with the unitary weapons. We found that the following statements are well supported by the available evidence:

--Design characteristics give binary weapons safety features that facilitate their handling, storage, and transportation in peacetime.

--"Arming" the binary weapons diminishes these safety features.

--Open-air testing has been banned since 1969 and as a result no field data have been collected on the performance characteristics of binary weapons.

--Binary weapons require more space for storage and transportation than unitary weapons do. For the 155-mm projectiles, for example, nearly four times as much space is required.

We found little or no information regarding the following issues and, therefore, cannot make conclusions about them with an acceptable level of confidence:

--the extent to which the noise and odor associated with the binary weapons detract from their utility in achieving military objectives;

--the extent to which the technical aspects of binary weapons, including mixing and arming them, place unacceptable constraints on the weapons' tactical utility;

--the extent to which data from simulants are useful in predicting the performance of binary weapons and, therefore, their utility in meeting military objectives;

--whether binary weapons offer significant advantages over unitary weapons on a wide range of operational and technical factors such as dispersion patterns and toxicity levels;

--whether binary chemicals are safe to produce;

--whether procuring binary weapons will significantly improve the U.S. chemical retaliatory capability.

We found that the evidence is generally insufficient for conclusions on the performance advantages of binary weapons compared with unitary weapons. There is support for the assertions about the peacetime safety features of binary weapons, and there are also unexplained indications that these peacetime advantages may have related wartime costs.

HOW DOES MODERNIZATION AFFECT THE PROSPECTS FOR DISARMAMENT?

Having reviewed DOD's plans for chemical weapons modernization, we examined information on the effect modernization is likely to have on the prospects for the ultimate deterrent--a chemical weapons ban. We found a history of slow progress in treaty negotiations, which have been substantially hampered by a lack of agreement on the issues of verification. Although the United States and the Soviet Union have agreed that the verification of a chemical weapons treaty should be based on a combination of national and international measures, the Soviets have consistently rejected requests for on-site verification of treaty provisions. A draft paper delivered in 1982 to the United Nations by the Soviet Union may offer some hope of flexibility in the Soviet position, but the Arms Control and Disarmament Agency is taking a "wait and see" attitude toward the draft paper. The verification issues are complex, and in many areas information potentially useful in resolving them is lacking.

For example, we found no objective evaluations of whether using several nonintrusive verification techniques at one time would bolster the likelihood of detecting activities related to chemical weapons. In addition, we found that a number of pertinent questions have not been addressed:

--Have technological advances in the last decade made long-range sensing devices (such as remote sensors in air or on space platforms) likely verification tools?

--Is computer-based verification realistic and not overly intrusive?

--What techniques or combination of techniques give the greatest probability of detecting treaty violations?

As to whether U.S. chemical warfare modernization plans would result in a negotiations breakthrough or breakdown, we found advocates for both positions but little data. The arguments depend on beliefs about how a U.S. chemical weapons build-up would be perceived. We inquired whether procuring binary chemical weapons would mean a proliferation of chemical weapons and a further complication of disarmament negotiations. Arguments on these issues depend on how easily binary weapons can be produced and the way in which binary weapons would further complicate the already complex verification issue. Resolution of the arguments will require answers to these questions:

(1) How easily can binaries actually be produced? (2) What nations have the ability to produce binaries? (3) How would producing binaries affect the value of existing verification procedures? We find that these questions are rarely enunciated and even more seldom analyzed.

SUMMARY OBSERVATIONS

The general impression left by the literature is that there is little empirical data in areas pertaining to the functioning and usefulness of chemical weapons. Conjecture plays a major role in the formulation of theories of chemical warfare deterrence and in the analysis of Soviet threats and U.S. responses. We offer the following seven observations on primary information needs.

Observation 1

The literature agrees that more reliable information is needed on Soviet offensive capabilities. The evidence is strong that the Soviets have been building their nuclear, biological, and chemical defensive capabilities, but this does not necessarily imply, as is sometimes assumed, that U.S. retaliatory chemical warfare capabilities require strengthening.

Observation 2

It is argued reasonably in the literature that some retaliatory chemical capability is necessary in order to degrade enemy performance and remove the potential advantage of an enemy's using chemical weapons, but the literature shows no analysis of the proportion of chemical to nonchemical munitions that would be required to achieve this objective. No analysis identifies the implications for the U.S. stockpile when degradation is the major military objective.

Observation 3

The literature does not conclude that chemicals are tactically more advantageous than other weapons in achieving military objectives other than the degradation of an enemy's performance. There seems to be no information on the comparative ability of chemical and other weapons, alone and in combination, to cause casualties in attacking specific battlefield targets. If analysis is to be conducted, it should assume a well-protected enemy, given what is known about Soviet defensive capabilities.

Observation 4

Comparative analyses of the effectiveness of the various chemical delivery systems have not been made. The literature is confined to concern about reliance on the Bigeye bomb for long-range capability.

Observation 5

Despite the fact that a simulation sponsored by the Joint Chiefs of Staff indicates that as much as

, there is no evidence that steps are being taken to protect civilian populations in the event of a chemical war.

Observation 6

The literature shows that historically chemicals have been used in warfare in only limited ways because chemical warfare has never been assimilated into armed forces procedures, preparing everyone on the battlefield with respect to chemical weapons so that they know what to do, how to do it, when to do it, and what will happen if it is done. The literature shows that it has still not been assimilated.

. However, the simulation study sponsored by the Joint Chiefs of Staff indicates that, in a European conflict,

. The question of a chemical versus a tactical nuclear response, and the associated costs, deserves further analysis.

Observation 7

Given the implications for national security and dollar expense in DOD's proposal to modernize U.S. chemical warfare capability by producing binary weapons, the literature contains surprisingly little analysis of the advantages and disadvantages of these weapons compared with the unitary weapons they would replace. What is known about the ability of other countries to produce nerve agent and munitions should be brought up to date in a way that considers their binary capabilities and identifies the implications for the issue of the verification of a weapons ban.

AGENCY COMMENTS AND OUR RESPONSE

Draft copies of this report were submitted to DOD for comment on December 9, 1982, and we granted a request for additional time beyond the customary 30 days for review, extending DOD's comment period to January 21, 1983. On January 24, 1983, we met with DOD officials at the Pentagon. Our representatives were advised that written comments would not be available and that the purpose of the meeting was to provide us with official oral comments on the draft report. These official oral comments were presented by Dr. Theodore Gold, the Deputy Assistant to the Secretary of Defense for Chemical Matters. Dr. Gold began his comments by acknowledging a need for good analyses on chemical

warfare. We concurred with this view and indicated that we were aware that his office was proposing to sponsor analyses, through the Institute for Defense Analyses, on chemical warfare joint test and evaluation. We also indicated our familiarity with previous IDA analyses on chemical warfare. After this preliminary, Dr. Gold presented four points as the official DOD comments on this report.

DOD point 1

A literature review is not an adequate method for addressing issues in this area because some relevant information is not in documented form. Moreover, the draft report does not cover some documents that are pertinent to the issues. Giving an example of the limitation of a literature review as a basis for addressing issues in this area, Dr. Gold cited our discussion in the report of the size and condition of the U.S. chemical stockpile. He contended that quoting figures from various documents written over a period of several years does not constitute an adequate basis for judging stockpile size or condition. He noted that DOD had recently attempted to assess the chemical weapons stockpile.

Our response

We informed Dr. Gold that we used several techniques in preparing the report. We reviewed the literature but we also made use of a panel of experts, who assisted us in determining which documents to include in our review. We assessed the value of each document in terms of how well it supported its conclusions and the degree to which its findings were reinforced by similar conclusions in other studies. We incorporated information from interviews we held with officials of DOD, including the armed services, and with notable experts and independent researchers. In the course of collecting data, we attended briefings and congressional hearings on chemical warfare issues. The information we gained in these activities supplemented the information we gathered from the literature and helped us identify the major issues in the subject of chemical warfare. (In chapter 1, we present full details of our methodology).

With regard to the stockpile example Dr. Gold raised, we informed him that we used two recent documents sponsored by DOD to address stockpile issues in our report--the 1981 Defense Science Board study and DOD's 1982 report to the Congress on chemical warfare. When we asked Dr. Gold for documentation on the more recent DOD efforts to assess the stockpile size and condition, he did not provide any additional sources.

DOD point 2

The report does not provide a balanced and complete picture of the important issues in chemical warfare. Giving an example,

Dr. Gold stated that we had not reviewed primary intelligence data regarding an enemy's threat of using chemical weapons.

Our response

We discussed with Dr. Gold and the DOD officials how we used intelligence information, and we agreed to clarify the report to show that we did not use primary intelligence data, did not challenge any intelligence data, and accepted at face value and used intelligence information that is cited in DOD documents. We also pointed out that the Central Intelligence Agency reviewed a draft of the report and did not challenge the way we have referred to intelligence information.

DOD point 3

The report contains many factual errors and errors of omission, and there is additional documentation that would have been of assistance in the preparation of the report.

Our response

We requested Dr. Gold to support his statement that the report contains many factual errors. However, he offered us no examples of error in the report, responding only that DOD did not make a line-by-line review. When we asked for the titles and sources of the additional documentation that Dr. Gold had referred to, none were given.

DOD's point 4

GAO did not work through Dr. Gold's office and did not talk to responsible officials in DOD or the individual services.

Our response

Regarding Dr. Gold's concern that we did not work with his office and did not talk with responsible officials, we pointed out that we had conducted the interview and data collection phase of our work before he arrived at DOD and that we will make this clearer in the report. We also presented him with a list of individuals in DOD and the services whom we made contact with during our audit. The list includes Major General Niles Fulwyler and members of his staff (his office served as the Army's focal point for chemical warfare matters during the period of our review), Colonel John Tengler of the Joint Chiefs of Staff, Victor Utgoff and Colonel Horace Russell of the National Security Council, Robert Mikulak of the Department of State, and Professor John Deutch of the Massachusetts Institute of Technology (during a briefing on chemical warfare that he presented at the MITRE Corporation). We added that we had attended and obtained testimony presented to the Senate Appropriations Committee in May 1982 by Dr. Richard L. Wagner, the Assistant to

the Secretary of Defense for Atomic Energy, by Dr. Theodore Gold, the Deputy Assistant to the Secretary of Defense for Chemical Matters, and by the Honorable James F. Leonard, former Ambassador and senior official in the Arms Control and Disarmament Agency on chemical and biological warfare issues. Dr. Gold indicated that Amoretta Hoeber, the Principal Deputy Assistant Secretary of the Army for Research, Development, and Acquisition, has no records indicating that she received, reviewed, and commented on the list of sources we compiled for this report. We replied to Dr. Gold that we can provide documentation that verifies that she did review a draft version of our bibliography (printed as appendix II in this report).

We have revised the report so that it includes a discussion of how we treated intelligence information, which we hope clarifies the concern that DOD raised. The other official comments were so general that, without more specific reference, we were unable to make any revision that could be based on them.

We received a written response from DOD well past the established time for the submission of agency comments. However, since it documents the oral presentation we have discussed above, we have included it in the final report in appendix IV. The letter of response we sent to DOD is also printed in appendix IV.

The Response from the Department of Defense

THE UNDER SECRETARY OF DEFENSE

WASHINGTON, D.C 20301

RESEARCH AND
ENGINEERING

4 FEB 1993

Ms Eleanor Chelimsky
Director, Institute for Program Evaluation
U.S. General Accounting Office
Washington, D. C. 20548

Dear Ms Chelimsky:

This is the Department of Defense response to your draft report entitled "Will the Billions of Dollars for the Chemical Warfare Modernization Program Accomplish Its Stated Objectives?", Code 973544 (OSD Case 6152). Fulfillment of this report's intent (as stated on page 1) could have provided valuable assistance to elevate and inform the current national debate on how best to eliminate the threat of chemical warfare (CW). However, as currently written, the report does not provide a complete, accurate, or balanced review of the questions (as was the stated purpose of the effort), or offer any recommendations for action to those responsible for administering the program. As a result, the report does not provide useful views and data that will raise the level of debate, or enhance the knowledge or understanding of either responsible proponents or critics of the CW Modernization Program.

As acknowledged in the report, Soviet CW capabilities, US arms control efforts, and the DOD program to deter chemical warfare are addressed and assessed using as a basis only a literature review. The auditors did not review intelligence data, did not talk to responsible officials, did not read Congressional testimony, did not visit facilities and installations, did not review pertinent arms control verification documents, and did not review applicable service manuals and plans. In short, critically pertinent information and sources necessary to an informed judgment were omitted from the review.

The report indicates that there are a "multitude of unanswered questions." Many of the questions appear unanswered, because the proper source was not contacted and pertinent questions were not raised during the audit. For example, DOD has an office--Office of the Deputy Assistant to the Secretary of Defense (Chemical Matters)--that is the focal point for all chemical warfare matters, but that office was not contacted during the course of the audit. An example of the limitations of the report's literature search approach is found on page 6-9, where the authors state that "The total size of the US chemical stockpile and its condition are not precisely known; our review

consistently found estimates ranging from * agent tons to * agent tons." These estimates were apparently extracted from a variety of documents written over a period of years. These sources do not constitute an adequate basis to judge what DOD believes is the pertinent question. That is, does the current custodian of the chemical stockpile know its size and composition? As far as we can determine, the auditors made no attempt to evaluate DOD's current state of knowledge, or to evaluate its recent effort to assess stockpile conditions. This type omission is evident throughout the report, rendering it unreliable as a guide to understanding the issues, even if the audit had not been based entirely on an incomplete literature review.

The study and identification of the true points of contention in the important and emotionally-charged issues surrounding the CW Modernization Program would be a valuable asset to a national debate. Alternatively, a comprehensive discussion of the substantive positions of both proponents and critics of modernization of our CW deterrent capability would be of great value. Although review of this draft report shows it will contribute to neither objective, DOD will continue to cooperate in any effort to illuminate the key issues involved in the central objective of eliminating the threat of chemical warfare.

Sincerely,

Jim Wade

James P. Wade, Jr.
Principal Deputy Under Secretary of
Defense [illegible]

* Numbers are classified.

UNITED STATES GENERAL ACCOUNTING OFFICE
WASHINGTON, D.C. 20548

INSTITUTE FOR PROGRAM
EVALUATION

February 22, 1982

Mr. James P. Wade, Jr.
Principal Deputy Under Secretary of
Defense for Research and Engineering
Department of Defense

Dear Mr. Wade:

Thank you for your letter of February 4 giving me the written position of the Department of Defense (DoD) on our Chemical Warfare paper. As you know, your letter was delayed beyond the time which GAO allocates for agency comments (DoD had the full 30 days, plus a 10-day extension requested by your staff and granted by GAO). However, since your letter contains no new information and reiterates some of the points already made to us by your staff in the official "verbal comments" session of January 24, you may be sure that we have carefully considered all of your points and that we will be responding generally to the DoD comments in our report.

One thing you may want to note: I think we are in presence of a misunderstanding about the nature of our report methodology: it is neither a "literature review" nor an audit. It is an information synthesis which does indeed begin with a literature review but goes very much further, analyzing the quality of each piece of information (in terms of the evidence supporting it) with an end-product of refined information about the state of knowledge in a particular area at a particular time.

The purposes of such an effort are: (1) to try to make sense out of conflicting information that exists on a given topic (conflicts cannot always be easily resolved, of course, but sometimes they can be when it turns out, for example, that one study has been soundly designed, implemented, and reported, whereas another is based solely on the author's opinion or on anecdotal evidence); (2) to develop an agenda showing clearly where the gaps in needed information are that call for new agency research; and (3) to lay the groundwork for further GAO evaluation or audit work in the area.

In using the information synthesis approach, we do not expect to propose any agency action, other than the filling of important knowledge gaps our work has revealed. Therefore we make no recommendations, contrary to the procedure we would use in a methodology featuring original data collection, such as an effectiveness evaluation or an economy and efficiency audit. However, we do make conclusions and observations about the information we have found and to

do this naturally entails the prior elaboration of a synthesis framework laying out the questions and subquestions to be answered, the scope, nature, and time-frame of the initial literature review, and the criteria for assessing the quality of the information. If you look at our report, you will see that we have documented this important front-end work in considerable detail.

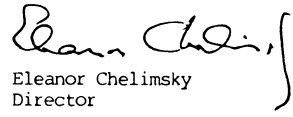
A potential problem in such an approach might be the question of the "universe": that is, how can we be sure we've got all the major studies? In this case, although it was an especially arduous task to accomplish--given the breadth, international character, and classification of the topic, and the obscurity of some of the work--we now feel assured that we have covered all the major studies done as of May 1982 (end-date for our data collection effort). One of the methods we use in the synthesis approach to reach this assurance is through the combined knowledge of a panel of experts. (In this case, we included DoD's General Niles Fulwyler and Dr. Amoretta Hoeber. The OSD focal point position was not filled at that time, as you know.) We were further confirmed in our confidence by peer reviews of our work (including the CIA) and our January 24 session with your staff in which no title, document, or source was produced that GAO had not already reviewed and analyzed.

With regard to the potential benefits of the synthesis approach, we feel they are enormous. First, the ability to draw on a large number of soundly designed and executed studies adds great strength to the knowledge base when findings are consistent across different studies by different scholars using different methods. No single study, no matter how good, can have this kind of power. Second, when studies are not well designed and executed, the knowledge that there exists no firm basis for action is also an important benefit: the size of the risk is clarified, necessary caution is introduced into the debate, and over the long term, the number of failed shots in the dark is likely to be diminished.

I hope this letter will better explain what we are trying to do and how it differs from an audit or literature review. A GAO staff paper describing the synthesis methodology may be of additional help. Please let me know if you would like to see it.

With kind regards,

Sincerely yours,



Eleanor Chelimsky
Director

Congressional Record—House
June 15, 1983

AMENDMENT OFFERED BY MR. ZABLOCKI

Mr. Zablocki: Mr. Chairman, I offer an amendment.

The Clerk read as follows:

Amendment offered by Mr. Zablocki: Page 2, line 15, strike out "$2,272,500,000" and insert in lieu thereof "$2,157,900,000".

(Mr. Zablocki asked and was given permission to revise and extend his remarks.)

Mr. Zablocki: Mr. Chairman, for 3 years now the Congress has been engaged in a vigorous debate with the Department of Defense over the question of breaking this country's longstanding moratorium on the production of lethal nerve gas.

Last July, in an historic vote, the House of Representatives overwhelmingly rejected—by a margin of 251 to 159—the Pentagon's request for new nerve gas when it adopted an amendment offered by myself and the gentleman from Arkansas (Mr. Bethune), my Republican colleague, to delete funds for chemical weapons production.

Last year, when the Congress decisively rejected any binary production money, the House at the same time expressed a clear desire to maintain our existing chemical retaliatory capability, while pursuing serious chemical arms control negotiations.

I recall that in that debate, we pointed out, that rather than produce the binary, we should spend money on defense against chemical warfare.

I believe that the American people support continuation of the policy which seeks to eliminate the horrible threat posed by chemical warfare.

It is for this reason that I am once again introducing this amendment, along with the gentleman from Arkansas (Mr. Bethune), to delete $114.6 million requested for binary chemical weapons production. We do not delete any money from research and development.

I would like to point out to my colleagues that this amendment enjoys broad based support from Members on both sides of the aisle.

To name just a few, Mr. Bethune and I are joined today by Mr. Edwards of Oklahoma, Mrs. Roukema, Mr. AuCoin, Mr. Leach, Mr. Bonior, Mr. Green, Mr. Gejdenson, and Mr. Kastenmeier.

I want to make it perfectly clear again to our colleagues, Mr. Chairman, precisely what the effect of my amendment is. The amendment leaves intact nearly 90 percent of the funds requested for the U.S. chemical warfare program. That amount, that 90 percent is over three quarters of a billion dollars.

This includes almost $70 million for badly needed chemical defensive programs, $76 million for demilitarizing the small portion of the current stockpile that is unusable, and about $40 million for research and development and maintenance of the existing stockpile of lethal chemical weapons.

Now, why are we striking this $114.6 million? It is because in my opinion, Mr. Chairman, we are not ready to go into production of binary munitions. They have not been adequately tested. We have problems with the Big Eye bomb, in particular, which has been blowing up in laboratory tests.

Simply put, my amendment would only delete $114.6 million for the production of binary weapons, which are not needed, are not fully tested, and which would result in billions of dollars of unnecessary Government spending.

Mr. Chairman, I am convinced that the Congress wisely deleted the funds last year and that the Congress is equally justified in deleting the funds for this costly and unnecessary program again this year.

Two recent developments which I will now discuss reinforce this point. Just a month ago, the Pentagon asked the House Armed Services Committee to defer $43 million it had sought for the production of the Big Eye binary bombs, because the bomb could explode on its own and spew deadly nerve gas while being carried by American fighter aircraft.

According to Dr. Ted Gold, the Deputy Assistant Secretary of Defense for Chemical Matters, the problems with the Big Eye binary bomb were not discovered until late last year. This particularly disturbing in light of the fact that the Big Eye bomb had been under development since the late 1950's.

The Chairman pro tempore: The time of the gentleman from Wisconsin has expired.

(By unanimous consent, Mr. Zablocki was allowed to proceed for an additional 5 minutes.)

Mr. Zablocki: And also, because the Pentagon had sought congressional approval of funds for Big Eye bomb production last year before these serious technical problems were discovered.

I believe there are two important lessons to be learned from the Big Eye bomb "bomb." First, it underscores the importance of not proceeding with the production of weapons systems that have not been adequately and fully tested, which the binaries have not been; and second, it calls into question the viability of the binary concept which was thought to be a so-called proven technology.

Mr. Chairman, the second development which I referred to earlier was the release by the Committee on Foreign Affairs of a report prepared by the General Accounting Office, entitled "Chemical Warfare: Many Unanswered Questions." I would hope our colleagues, Mr. Chairman, would carefully review this document.

This new report which was begun nearly 2 years ago, concludes that spending billions of dollars on the production of binary nerve gas will not provide the United States with a "credible chemical warfare deterrent" capability.

The GAO report reaffirms what many of use have argued for the past several years; that is, that the first priority of our chemical program should be the correction of long standing deficiencies in U.S. chemical warfare doctrine, training, and defensive equipment. Production of new binary nerve gas weapons will not improve the U.S. chemical warfare deterrent posture in the fact of these serious deficiencies.

Further, Mr. Chairman, the GAO report made the following findings:

First, contrary to DOD assertions that the Soviet Union possesses a massive chemical warfare capability, GAO found that:

> Little is known about the size and mixture of the Soviet stockpile of chemical munitions" and, that
>
> The Soviet chemical warfare delivery means are virtually unknown.

Second, despite DOD claims that the U.S. stockpile of chemical weapons is obsolete and deteriorating, GAO found that:

> The precise size and condition of the U.S. stockpile are not known:

Third, although DOD had claimed that binaries are more advantageous than the existing unitary chemical weapons, GAO disagrees. The report says:

> Available data does not sustain the argument that binaries offer substantial technical and operational advantages over existing weapons.
>
> Binaries place greater space requirements on storage, transportation, and deployment.
>
> The mixing requirements of binaries may diminish their operations effectiveness, and
>
> Peacetime advantages (of binaries) may have related wartime costs (such as mixing time and more complex logistics).

Fourth, despite Reagan administration claims that binary chemical weapons production may have a positive effect on prospects for chemical weapons arms control, GAO found that:

> Binary production might complicate verification procedures.

Finally, although the United States has possessed lethal chemical weapons for decades, the GAO found that:

> The United States does not have a chemical warfare doctrine.
>
> The United States, unlike the Soviet Union, has not built a strong ability to defend against chemical warfare, and
>
> The United States does not have realistic or adequate training and exercises for chemical warfare.

Mr. Chairman, if there are any Members here that have some concerns or questions regarding the U.S. retaliatory capability, I do not have a secret document which outlines our stockpile. I shall not quote from it, but it is available. I am sure the Armed Services Committee and the gentleman from

Arkansas (Mr. Bethune) have the same information. I do not think we need to go into a secret session, but the data which supports my claim as to the adequacy of our existing chemical weapons stockpile is available.

Finally, as to DOD claims that binaries are safer and better than existing unitary munitions, the GAO disagrees, noting that binaries have significant technical and operational disadvantages over existing weapons, and therefore, more investigation is needed and more research and development is needed.

That is what many of us in Congress are saying. Research and development, yes. Production of untested and unreliable binary munitions, no.

This is basically the bottom line of this amendment.

I am one who has continuously supported an adequate defense for our country. I would not be offering this amendment if I thought I would be putting my country, our country, at a disadvantage.

Mr. Chairman, the Committee on Foreign Affairs has had an active interest in U.S. chemical warfare policy for the past decade. During the many hearings we have held over the years on this issue, we have always sought to answer one key question: "Does resumed nerve gas production contribute to U.S. national security interests?"

Mr. Chairman, the answer has and continues to be "No."

The United States already has an adequate stockpile of usable nerve gas munitions which have been fully tested and can be relied upon to provide a sufficient retaliatory capability in a crisis situation.

Yes, yes, there are a few old munitions that must be destroyed. But to produce new binary artillery shells, which are 18 times the cost of upgrading one existing unitary shell, is simply a waste of the taxpayers' money.

The Chairman pro tempore: The time of the gentleman from Wisconsin (Mr. Zablocki) has again expired.

(By unanimous consent Mr. Zablocki was allowed to proceed for 3 additional minutes.)

Mr. Zablocki: To move forward with the production of binary munitions at this time, Mr. Chairman, would be detrimental to our relations with our friends and allies in Europe who oppose chemical weapons and could seriously jeopardize NATO's ability to go forward with intermediate nuclear force deployments in the event that there is no progress in the INF talks by December of this year.

My colleagues, I just got back from a NATO assembly over the weekend. In talking with and questioning our counterparts, parliamentarians from our NATO ally countries, I found that they continue to be concerned about chemical warfare. Where would we place these weapons? What theater are they to be used in? Is it not Europe?

But our allies will not allow them to be deployed in their countries. Only one country at the present time has chemical weapons on its soil. If we do not have a place to keep them, the question in my mind is why are we spending so much money on the binary program and not spending money where we really need it, which is on chemical defensive programs.

Resumed nerve gas production affords the Soviet Union yet another propaganda advantage vis-à-vis the world community. Yes, you and I know they have been charged with using chemicals in Afghanistan, in Kampuchea. But if we go into production after some 14 years of a moratorium, the Soviets would say the United States intends to use chemical warfare.

I believe that it is clear that binary production undermines rather than enhances U.S. security interests.

Mr. Chairman, I urge my colleagues to support this bipartisan effort to delete the funds for production of binary weapons.

Mr. Bethune: Mr. Chairman, will the gentleman yield?

Mr. Zablocki: I will be delighted to yield to the gentleman from Arkansas.

Mr. Bethune: I would like to compliment the gentleman on his amendment and express my appreciation for the way in which he has worked with me over the last year. I think the gentleman has brought out some very excellent points and certainly in the space of the short time in which he was permitted to address the House he has provoked thought among Members about the salient issues which we must deal with before we can make any decisions on this binary weapons programs.

The Chairman pro tempore: The time of the gentleman from Wisconsin (Mr. Zablocki) has again expired.

(On request of Mr. Bethune and by unanimous consent Mr. Zablocki was allowed to proceed for 2 additional minutes.)

Mr. Bethune: The fact of the matter is since 1969 this country has neither produced nor used chemical weapons.

This is in stark contrast to what the Soviet Union has done. I think the whole world knows certainly by perception if not in reality that the Soviet Union is producing and is using chemical weapons. The stories of Sverdlovski, Afghanistan, Southeast Asia, are enough evidence, I think, on that point.

I hope as Members listen to the debate they will not only consider the very particular arguments that you and I and others will make about the particular weapons under consideration, the nature and extent of the stockpiles and so forth, but that they will pay close attention to the point the gentleman made about how our allies feel about the binary weapons and about the environment that exists at the Committee on Disarmament presently and in the world, which this country could take advantage of in its quest for a treaty, which would be the ultimate cure to the problem, and that is a ban on chemical weapons. I hope that Members will listen attentively because I am satisfied that those who wish to commence production will try very hard to offer very interesting amendments or substitutes that would lead Members to believe that we are not really making a decision here today to produce, we are only making a decision to sort of get started.

I think it will become clearer as the debate wears on that you cannot have it both ways and that the best procedure for this country today is to hold to the existing policy until such time as we are satisfied that we need to do this,

which we do not need to do right now, and until we are satisfied that the weapons that they are offering to this House for approval work, and the ones that they are offering now do not work.

So I think those are the two critical issues that will have to be developed here in the course of this debate, but always being sensitive to the question of arms control and the feelings of our friends and our allies in nonalined countries around the world.

Mr. Zablocki: I thank the gentleman for his contribution. We should not give the Soviet Union an opportunity to transfer the criticism of the world, that is now on their back, on us.

Amendment Offered by Mr. Leath of Texas to the Amendment Offered by Mr. Zablocki

Mr. Leath of Texas: Mr. Chairman, I offer an amendment to the amendment.

The Clerk read as follows:

> Amendment offered by Mr. Leath of Texas to the amendment offered by Mr. Zablocki: Strike out the amount proposed to be inserted by the amendment and insert in lieu thereof "$2,272,400,000".
>
> At the end of the amendment add the following:
>
> "At the end of title I (page 10, after line 12) add the following new section:
>
> "LIMITATION ON THE PRODUCTION OF BINARY CHEMICAL MUNITIONS
>
> "Sec. 109. Funds appropriated pursuant to authorizations of appropriations in section 101 for ammunition for the Army for binary chemical munitions may be used for the establishment of a production base for binary chemical munitions and for the procurement of components for 155-millimeter binary chemical artillery projectiles, but may not be used for actual production of binary chemical munitions before October 1, 1985. For purposes of this section, production of binary chemical munitions means the final assembly of weapon componets and the filling or loading of components with binary chemicals."

(Mr. Leath of Texas asked and was given permission to revise and extend his remarks.)

(By unanimous consent Mr. Leath of Texas was allowed to proceed for an additional 10 minutes.)

Mr. Leath of Texas: Mr. Chairman, those of you that have been here since I came to Congress in 1979 know that I have subscribed to a proverb that my father had that you can learn a great deal more by listening than you can by talking. I have seldom come to this well in those 4½ years, but today I am compelled to do so because I sincerely believe there has never been a single issue in those years more misunderstood than this one we are debating here today.

This is indeed an issue that can be made so very, very emotional because there is no question about the fact that chemical weapons are horrible. It can

be drowned in rhetoric, and it has been, because it is a very complex issue involving a great deal of personal opinion and a great many assumptions on the part of all of us.

Real arms control is the issue. But if you want a strong conventional force structure, if you want to keep the chance of us using nuclear weapons much less than it would be under our current circumstances, then, ladies and gentlemen, I think you will support this amendment and this concept.

On an issue as vital as this issue is, one that is so very vital to our conventional forces' threat, Mr. Chairman, it is imperative that we keep our focus on the facts. It is imperative that we keep our focus on historical reality. It is imperative that we keep our focus on knowledgeable assessment from those who know what constitutes relevance on a subject so complex as this one.

Because the program was unnecessarily delayed last year by this same opinion, our committee went into great detail this year to make sure that we were on solid ground by once again redefining all of the credible opinion that we could find in our Department of Defense on this program, and the timing of this program as it relates to our arms control efforts and cooperation with our negotiators.

Mr. Chairman, let me go through the concerns that have been expressed—legitimate concerns that have been expressed here—as best I can and then present to the Members of the House the logic that I believe will very effectively address those concerns.

Finally, Mr. Chairman, I want to show how my amendment can satisfy those concerns and still give us that little-bitty degree of insurance that we need just in case things do not work as we all hope they will.

First of all, Mr. Chairman, the strongest points made by Chairman Zablocki and Mr. Bethune last year in the debate, and this year, are purely that; they are points based on opinion, points based on their opinions. I happen to believe that both fact and history invalidate the conclusions that these dear colleagues have drawn on this subject. So let us examine their opinions in the light of existing facts and in relation to historic perspective and in relation to what this country needs to have a credible defense threat.

The first point—and interestingly enough neither gentleman has denied the fact that we need a chemical warfare capability, but the first point is, What is the current status of our chemical weapons threat? Both gentlemen contended that we already have a sufficient amount of chemical weapons in our stockpile both in quantity and quality to provide a credible deterrent.

Mr. Chairman, that contention is absolutely not true and it is not borne out by anybody in our services with any degree of credibility that says that it is. The size and condition of our stockpile is precisely known and it is completely inadequate, comprising only a small portion of what the JCS requirement is. Yes, we have tons and tons of highly toxic, deadly, dangerous chemicals, many of them 50 years old, 40 years old, 35 years old. The truth is 72 percent of that stockpile is either in unserviceable rounds and rockets that

we no longer have launchers to use or they are in bulk. My friend likes to stand up and talk about the beautiful safety of these chemicals. Yes, they are like a rattlesnake. Out in the western part of my district, a rattlesnake will not bother you as long as you have him in his lair, when you step on him, it is deadly.

The fact is if we had to move this tremendous amount of chemical agents that we have stored we indeed would be in trouble, not only from the troops that would have to use it but from the population centers and in all of the districts where we would have to take it.

Of the 22 percent of our chemical short-range capability only 6 percent of it, only 6 percent of that is useable. And the amount that we have in deep-strike capability, Mr. Chairman, is so minute that it is not even worth talking about.

Mr. Chairman, the very best military opinion that we have in this country strongly states that our stockpile is totally inadequate to pose any semblance of a counterthreat to deter the Soviets from using chemical weapons in the event of a conventional confrontation. We have letters which we can put into the RECORD and let the Members read, from General Vessey, from everybody in the Department of Defense, who has addressed this subject.

The truth is, Mr. Chairman, that we can deliver a chemical strike only to the first echelon of a conventional force. That is approximately 16 kilometers. If you envision a conventional battlefield where you have at least four echelons, that attacking echelon, the second in reserve, then the third and then the fourth which would constitute the airfield, ships, and so forth.

We have the capability in our arsenal to go the first echelon, 16 kilometers. Now, the Soviets by contrast can strike all four echelons. It has been estimated that they could wage a chemical war for at least 30 days on a 500 kilometer range and focus on all four echelons of our forces on that same battlefield, which means they could get our ships, they could get our airfields, our supply depots, our reserve areas up to 500 kilometers. They outnumber us at least 12 to 1 in chemical warfare personnel.

They have 85,000 troops especially trained in chemical warfare, in every division they have. They outnumber us over 5 to 1 in weapons, over 4 to 1 in stockpile, even as inadequate as ours is. Get this, they have 14 current producing chemical weapons plants. My colleagues, we have none, we have zero. They outnumber us at least 25 to 1 in decontamination equipment. These figures have to convince anyone that has an open mind that there is no comparable threat, when we compare our aged, deteriorating arsenal to theirs.

Now, why is comparable threat absolutely necessary doctrine in chemical weapons? Let us talk about that a little bit because that is where the focus of this debate should be. The focus of this should be on that doctrine. History, Mr. Chairman, that is why, history graphically verifies that doctrine.

It is probably the most valid military doctrine that we have. We can argue forever in this Chamber about the validity of a doctrine that says we have to

have absolute parity in nuclear weapons threat and I am not sure who is right on that argument.

I think I could make a credible argument either way. None of us really knows because, thank God, we have never had to go through a nuclear exchange, so it is all opinion. But there can be no argument concerning the validity of that doctrine in relation to chemical warfare, my colleagues, because 65 years of history, 65 years of history since World War I, prove that conclusion to be completely factual.

The quality of that stockpile has direct relation to the comparability of threat concerning conventional forces. Now, my friend if the two most heinous enemies in the history of modern warfare, Nazi Germany and Japan, did not use these weapons, there was only one reason why they did not. We had the ability retaliate in spades and they knew it.

Our intelligence confirms conclusively that they both considered this option and they rejected it and they rejected it for that one reason, that the threat was comparable.

If the same disparity that exists today had existed in 1945, Mr. Chairman, history would indeed have been rewritten and we would not be here today talking about this issue.

If the first foundation of the Zablocki-Bethune approach assumes a comparable threat, and it does, Mr. Chairman, it is indeed a foundation that is built on sand. It is a totally invalid and undefensible, by either fact or history. You can shoot at it with rhetoric, you can shoot at it with assumptions and you can shoot at it with opinions, but you cannot subtitute it for fact.

The second basic foundation of the Zablocki-Bethune initiative assumes that we are just on the verge of reaching agreement with the Soviets to ban chemical weapons and for us to do something to protect our interests after 14 years would harm those efforts. Here again, my friends, they make some very dangerous assumptions, some very loose assumptions that in my judgment cannot be validated in fact, either historical or judgmental.

The both made this same contention so many times over the years that I think in fact they destroy their own case.

Again, my colleagues, this is pure opinion. Let me tell you what the facts are and then you draw your own opinions, you make up you own minds, based on history and fact. This scenario began 14 years ago in in 1969 when President Nixon, hoping that we could entice the Soviet Union into agreeing on banning the manufacture, stockpiling and use of chemical weapons, unilaterally, on our own, halted our efforts.

So, for 14 years now we have done nothing but talk. During that period of time the Soviets have talked and they have manufactured and they have stockpiled and they have used. After completion of the Biological and Toxic Weapons Convention in 1971, an international discussion of complete chemical weapons ban began in earnest. This work was carried out in the conference on the Committee on Disarmament, which is the forerunner of the current Committee on Disarmament.

The United States participated very actively in these discussions. The United States and the Soviet Union then in 1976 ensued bilateral discussions pursuant to an understanding that was reached in 1974 at the Vladivostok summit. These technical and exploratory discussions laid in mid-1977 to the opening of intensive bilateral negotiations on a comprehensive chemical weapons ban. Althogether 12 rounds of talks were held. The most recent being in July 1980.

While the United States and the Soviets reached agreement in principle on the scope of the ban, little progress was made in resolving a number of key verification-related issues.

The United States pressed for mandatory international onsite inspection in a number of situations. For example, the mothballing and eventual destruction of all declared chemical weapons production facilities. The Soviets obviously rejected the U.S. approach, advocating instead reliance on photographic satellites and on self-inspection by each nation. The United States made clear, and quite accurately, that such an approach could not provide competence in compliance.

Now remember, my friends, that we are talking about a scenario that started in 1969. Because of the stalemate in the bilateral negotiations, the Committee on Disarmament established a working group in chemical weapons in 1980. The purpose was to create a forum for intensive, multilateral consideration of a chemical weapons ban.

The United States actively supported this very important step in view of the failure to make satisfactory progress behind closed doors with the Soviet Union.

After reviewing this situation, the Reagan administration concluded that the prospects for progress would be best if the Committee on Disarmament remained the principal forum. U.S. representatives were directed to participate vigorously and focus on resolving verification issues.

Consequently the United States supported the gradual expansion of the working group's mandate, which now calls for elaboration of a convention. The majority of the members support effective verification which includes onsite inspections. The Soviets, while claiming to have made concessions in this area, failed to clarify its actual position.

Unfortunately, even in the multilateral forum, progress has been much slower than we had hoped.

With a view toward accelerating the committee's work, Vice President Bush addressed the committee on February 4. He emphasized U.S. support for a complete and effective chemical weapons ban. He called on the members of the Committee on Disarmament to join the United States in serious negotiations toward that end and announced that the United States would table the document contained detailed views of the content of a convention.

The U.S. initiative has been well received. However, the Committee on Disarmament failed to resume serious work on a chemical weapons ban until late

in the spring session. And why? Why was it late in the spring session? Due to a series of procedural roadblocks thrown up by the Soviets and their surrogates. At the Committee on Disarmament's summer session, which begins in mid-June, the U.S. representatives will continue to press for acceleration of the Committee on Disarmament's work on a chemical weapons ban.

The Chairman pro tempore: The time of the gentleman from Texas (Mr. Leath) has expired.

(By unanimous consent, Mr. Leath of Texas was allowed to proceed for an additional 5 minutes.)

Mr. Leath of Texas: The record is perfectly clear, Mr. Chairman. For the better part of 14 years now we have talked and we have talked, while they have stalled, and built, and used chemical weapons.

Both the gentleman from Wisconsin and the gentleman from Arkansas stood on this floor last year about this same time and assured us that we were on the verge of agreement. Now they come back 12 months later with that same reasoning. All we need is a little bit more time.

As Ambassador Lewis Fields clearly states in a letter to Senator Warner in April of this year:

> Unfortunately, during the first eight weeks of our 12 week spring session serious discussions were obstructed by procedural maneuvering of the Eastern Bloc. The result of this deplorable development has been a blockage of all substantive work in the committee.

Ambassador Fields continues:

> The Soviets have not yet responded in any substantive way to our detailed views. While I wish that I were in a position to give the Senate an optimistic view, my judgment is that we are years away.

He goes on to say that even if we do reach a moratorium after many more years of negotiating, it is going to take even more years in order to get that treaty where we can use it. To codify the detailed arrangements.

And in the meantime our 6 percent capability is going to obviously reach zero. Our threat will virtually disappear.

Are we close to an agreement as these gentlemen contend? Not according to our negotiator. Is the reason we are not our fault? Hardly our fault. Ambassador Fields concluded by stating, "that the public debate on the binary issue has the effect of reducing such leverage as we may have," and is damaging negotiations.

The gentleman from Arkansas claims that the United States has the high moral ground since we have not produced any chemical weapons since 1969. He states that this example of restraint and world opinion will force the Soviets to negotiate seriously. He claims that should we begin production of chemical munitions, we will no longer have world opinion on our side.

Mr. Chairman, I do not believe this argument is accurate. Fourteen years inaccurate. The United States has not produced any munitions since 1969.

This restraint has not caused the Soviets to negotiate seriously. It has enabled them to build a truly impressive arsenal which flies in the face of unilateral U.S. disarmament.

The gentleman's own statements on the subject are somewhat contradictory. He quotes:

> The world has not focused much on the fact that the Soviets are producing and using chemical weapons, and I think the reason that they have not is that they have not really noticed yet the stark contract between our policy and practice and that of the Soviet Union.

Then in the same CONGRESSIONAL RECORD of May 5, he stated:

> Here I think is advice that rings true to every person of commonsense, and it comes from former Ambassador James Leonard. He said that the United States has gotten much credit from refraining from producing and building binaries.

On the one hand we have not received any credit, on the other hand we received much credit. The fact is that after 14 years, credit or no credit, nothing constructive has happened.

Ambassador Fields accurately assesses the situation when he says:

> At some point, If the Soviet decide to negotiate seriously, they will pay little attention to world opinion. They will be influenced only by their perception.

To me, Mr. Chairman, the second basic foundation then of this approach, of this initiative, is just like the first. It is good rhetoric, it is sincere opinion, but it is poor policy and it is a policy of Russian roulette in our approach to our security interests.

On the other hand, my amendment gives a totally rational, sensible approach that should satisfy all among us with the exception of those who simply want to totally unilaterally disarm in this area. It exhibits faith, but it does not exhibit blind faith. It stresses first and foremost the desire for a ban. Yet it buys us some insurance just in case.

The amendment guarantees us 2½ more years on top of 14 long years. Obviously we will have to make annual authorizations also.

My amendment just merely guarantees that we have got a minimum of five negotiating sessions, a minimum of 2½ years before we can build a chemical munition.

The amendment clearly focuses on a ban treaty as our primary goal, but just as importantly, Mr. Chairman, it buys us some insurance.

This amendment gives us a path that leads us toward the objectives that we want, I think every Member in this House wants, which is a verifiable ban on chemical weapons and pending that ban, a credible deterrent to chemical warfare. It buys us a little insurance. And I ask my colleagues to join with me in this approach that I think will accomplish all of the things that all of us want to accomplish.

* * *

Mr. Anthony: Mr. Chairman, I thank the gentleman for yielding.

Inasmuch as the gentleman has quoted from the GAO report, if he would allow me, I would also like to quote some sections of that report that I think counteract some of the parts that he has taken out, and I do quote from the same report from which the gentleman spoke.

It notes that not only do the Soviets have a wide range of ways of deliver chemical warfare munitions but also this array gives them the ability to strike anywhere within NATO.

Then the report adds that there is a serious deficiency in the U.S. ability to threaten Soviet and Warsaw Pact targets in rear echelons.

That was the main purpose the DOD was asking for the Big Eye to be developed, so that we could go past what the gentleman from Texas (Mr. Leath) said was the first echelon and be able to go to the rear echelon.

So I think it should be noted that we can go through a report that may be critical in some aspects, but if we take a close look at that report, we can also find paragraphs and sentences that indicate the authors of that report are not so sure themselves.

So I would indicate my support for the Leath amendment.

The Chairman pro tempore: The time of the gentleman from Michigan (Mr. Bonior) has expired.

(On request of Mr. Anthony, and by ananimous consent, Mr. Bonior of Michigan was allowed to proceed for an 1 additional minute.)

Mr. Anthony: Mr. Chairman, if the gentleman will yield further, I would repeat that one can go through a particular governmental report and pick it apart on the words that are spoken, but I think that one can also go through a particular report like that and start tearing it apart for lack of quality, inadequate methodology, and false and unsupported facts and data and conclusions.

So I think that once we start citing reports, we have to take a hard, critical look at those reports.

Mr. Chairman, I thank the gentleman for yielding.

Mr. Bonior of Michigan: Mr. Chairman, I would just respond by saying that no one denies the Soviets have the capabilities to deliver their message, but I think the basic premise of the GAO report was very, very critical of the assumptions on which this Government is operating to go ahead with the Big Eye or indeed with the binary program. It is an extremely critical report, and I think in no way can it be asserted that it supports the policies which the committee has brought forward to this House floor.

Mr. Anthony: Mr. Chairman, will the gentleman yield for an additional comment?

Mr. Bonior of Michigan: I yield to the gentleman from Arkansas.

Mr. Anthony: Mr. Chairman, as I recall, the GAO report itself says that there is little known about the Soviets' chemical war capability. Would the gentleman in the well not agree that our military intelligence knows a great deal about their capabilities?

Mr. Bonior of Michigan: Well, we know a great deal, but we do not, unfortunately, have a definitive answer to what has happened in Laos, Kampuchea, and Afghanistan.

Mr. Stratton: Mr. Chairman, I move to strike the requisite number of words, and in support of the Leath amendment.

Mr. Chairman, one of the interesting issues of our debate, since we resumed consideration of our bill on yesterday, is that we are facing in this particular debate the same question that faced us when we were dealing with the amendment against the Asat.

Why is it all right for the Soviet Union to have some particular military capbaility, but when the United States proposes to duplicate the capability, somehow it is immoral, somehow it is bad, and somehow it is going to contribute to the destruction of the entire world? This is a kind of a double standard that some of us are using, and I find it hard to understand how anybody can defend it.

The fundamental purpose of establishing an up-to-date, modern chemical capability is to deter the other superpower from using his very considerable capability.

The gentleman from Wisconsin (Mr. Zablocki) indicated at the beginning of his remarks that he was attending the North Atlantic Assembly last week, and I was honored to be a small part of the distinguished delegation led by our great chairman. On our way back we stopped off in London. I got a copy of the Sunday Telegraph, and was reading some of the memoirs of the great wartime premier, Winston Churchill. It had this to say:

> Churchill was determined to be in a position to retaliate. On September 28, 1940 he pressed Lord Ismay to insure that Britain's stocks of gas were being maintained. "We should never begin," Churchill noted, "but we must be able to reply. Speed is vital here. The highest priority must be given." And he added, "I regard the danger as very great."

Mr. Chairman, I think we have something of this danger today, because we have no retaliatory capability, speed is vital, and the danger is very great.

Mr. Zablocki: Mr Chairman, will the gentleman yield?

Mr. Stratton: I will yield when I get through with my presentation, but my time is short.

Mr. Chairman, what we are doing as a result of this failure to provide a credible deterrent is that we are failing to provide for our infantry soldiers the kind of protection that a genuine retaliatory capability would give them. I would like to demonstate to the Members of the House exactly what this means as an extra burden on our infantry troops.

Mr. Zablocki: Mr. Chairman, will the gentleman yield?

Mr. Stratton: Mr. Chairman, if the gentleman would watch my presentation, I think he will be enlightened.

I find that it is against the rules of the House to actually clothe a Member of Congress with the equipment of one of our fighting troops, but the gentle-

man from California (Mr. Hunter) can properly demonstrate that capability. We will just show the Members what Mr. Hunter, if he were an infantryman, would have to undergo because we have no chemical retaliatory capability. He would have to put this strange mask on his face. It is all rubber, it is very hot and uncomfortable, and it has to go over his head. Mr. Hunter would have to breathe through this mouth filter and if he wanted to drink there is just a little hole where he can insert a tube from his canteen down here. That is not a very appetizing kind of thing to use.

Then the infantry soldier needs a very heavy coat to go on—I think this is just about your size, Duncan—which is designed to be impermeable to the current type of gas munitions or chemical munitions that the Soviets have—but I daresay we can assume the Soviets are already trying to make their munitions more able to penetrate heavy clothing.

These are the pants. This is the belt that includes the water, if he can drink it through that tube, and also where he carries his ammunition.

Then finally, of course, he needs some rubber boots, and he needs his metal helmet on top of that uncomfortable face mask.

Finally, for his hands he would first have to put on these white gloves because the heavy rubber outside gloves get very sticky and the rubber ones would not be able to get off.

This is what our infantry soldiers are having to put up with, and as a result of wearing this cumbersome uniform, the performance of the individual soldier is reduced by 50 percent.

Not only that, but it is in fact impossible to attend to the calls of nature. Therefore, the soldiers' span of daily combat is extremely limited.

Mr. Chairman, this is the burden which we in this House are inflicting on our U.S. troops all because we refuse to provide the credible retaliatory capability in chemical weaponry that we ought to have and that Churchill wanted so desperately in the dark days when Great Britain faced the threatened Nazi invasion.

The Chairman pro tempore: The time of the gentleman from New York (Mr. Stratton) has expired.

Mr. Zablocki: Mr. Chairman, I ask unanimous consent that the gentleman from New York (Mr. Stratton) be permitted to proceed for 2 additional minutes.

The Chairman pro tempore: The gentleman from Missouri (Mr. Skelton) has been seeking recognition, and the Chair recognizes the gentleman.

* * *

Mr. Montgomery: Mr. Chairman, I would like to move on to another subject pertaining to the General Accounting Office report, which is named, "Chemical Warfare," and it says, "Many Unanswered Questions."

I have had the privilege to look at that General Accounting report and it seems to me the report has many unanswered questions.

It also seems to this gentleman in the well that the report was written at the General Accounting Office and it did not really move out into the field and find out about this chemical situation. They did not contact the person or the agency of the Department of Defense, which is the Office of the Assistant to the Secretary of Defense for Atomic Energy, who has been heading up the DOD chemical focal point since 1980.

Under the General Accounting Office report, they admitted they did not even contact the person in the DOD that heads up the chemical production warfare.

Let me say that under the General Accounting report, they say, "Little is known about Soviet offensive chemical warfare capabilities."

But did they check any primary intelligence documents?

The General Accounting Office did not follow up on this. They did not check on anybody.

My last point is that the second item pointed out in the General Accounting report is that the size and condition of the U.S. chemical stockpile, the GAO said that the military did not know what they had in their stockpile.

The Chairman pro tempore: The time of the gentleman form Misissippi has expired.

(At the request of Mr. Leath of Texas and by unanimous consent, Mr. Montgomery was allowed to proceed for 2 additional minutes.)

Mr. Montgomery: Let me make this point. The General Accounting Office did not run a survey or run a stockpile audit, nor did they do a review of the inventory of what the Army had in their chemical stocks: so the General Accounting Office report on chemical stockpiles and chemical warfare, as far as I am concerned, is shot full of holes and it was written downtown at the General Accounting Office and they did not do much research on the report.

Mr. Leath of Texas: Mr. Chairman, will the gentleman yield?

Mr. Montgomery: I yield to the gentleman from Texas.

Mr. Leath of Texas: I think the gentleman makes an interesting point. I think the Members really need to perk up and listen to this.

This GAO report was made on what they call an evaluation synthesis methodology.

Now, when queried as to what that was, they said, "We sat down and read a bunch of reports and arrived at our own conclusions."

Now, I find it passing strange that if you are going to investigate something, you do not even go to the people in charge of it or to the source.

So the Department of Defense questioned this and they wrote GAO and said, "What is going on here?"

Let me give you the response. It goes, to Mr. James Wade, Principal Deputy Under Secretary of Defense for Research and Engineering:

"Here are three copies of the evaluation synthesis methodology as requested by your staff. One caution"—I want you all to listen to this, because you are talking about bureaucratese—

> the methodology differs in certain key respects from the information sysnthesis work we performed in the chemical warfare paper. For example, one dissimilarity is the prospective, rather than the retrospective, nature of the studies included in the information synthesis. This, of course, necessitates the development of independent assessment criteria, a problem one doesn't have in the evaluation of synthesis because of the availability of evaluation paradigm.

Now, if that is not a pretty good indication of how valid this GAO report is.

Mr. Montgomery: This is what we have been concerned about for the last couple years; they are just pumping out reports over there and they are really not doing the proper research, especially on such an important subject.

The Chairman pro tempore: The time of the gentleman from Mississippi has again expired.

(At the request of Mr. Anthony, and by unanimous consent, Mr. Montgomery was allowed to proceed for 1 additional minute.)

Mr. Anthony: Mr. Chairman, will the gentleman yield?

Mr. Montgomery: I yield to the gentleman.

Mr. Anthony: Mr. Chairman, I thank my colleague, the gentleman from Texas, for pointing out that the GAO report was based on other documents that the authors had an opportunity to read. Let me quote from that GAO report:

> The literature agrees in general that the United States lacks a credible chemical warfare deterrent. Inadequacies in the U.S. ability to retaliate and defend are all well documented.

Another section of that same report goes on to say:

> In contrast, the literature generally reflects the perception that the Soviet Union is highly capable of waging chemical warfare.

Case closed. The jury ought to bring in a verdict for the amendment of the gentleman from Texas (Mr. Leath).

* * *

Mr. Bateman: Mr. Chairman, I move to strike the requisite number of words and rise in support of the Leath amendment.

Mr. Chairman, I have listened with some interest this afternoon and I have read the proceedings of this body last year when I was not a Member. This issue is one of such significance that it has concerned me sufficiently that even though not a member of the committee I wanted to share a perspective of this Member.

We are dealing here, I think, with one of the great paradoxes that we have spoken to on this floor for many hours in the course of the last 60 to 90 days, a great paradox. That paradox is that deterrence to nuclear war, deterrence to chemical warfare, lies in our ability to retaliate against the use of such weapons by our adversaries.

We have heard a great deal of discussion about whether or not the Big Eye bomb, as it is referred to, is presently a perfected, totally confident weapon. We have heard discussions of many things.

We have heard references frequently to the General Accounting Office report which I have bothered to read. I applaud the distinguished gentleman for Arkansas (Mr. Anthony) for his having pointed out that in that very report the deterrent capability of the United States in chemical warfare is definitely regarded as deficient.

Yet, again to the paradox. If we do not have a credible deterrent how can we expect to avoid the risk of our forces being subjected to nuclear warefare agents?

I would say that we heard references earlier to conversations between the distinguished gentleman from Arkansas (Mr. Bethune) and the Soviet Ambassador at the chemical warfare negotiations in Geneva. I applaud that dialog and certainly we want to continue that dialog.

He told us what he told the Ambassador of the Russians at Geneva, and it is what I would have said to him had I been there. It is what any one of us would have said to him had we been there.

But what he did not tell us was what was the Russian Ambassador's response. What was his response then to our Ambassador Fields who has been there?

I am more concerned about what response they give to us than I am about a recitation of what all of us would obviously say to them, because there is no Member of this body who wants to use chemical warfare. We want to maintain a capability to improve that capability for the very reason that we do not want to use it.

It is a fact, not a theory, that the Soviet Union has a superior capability in chemical warfare, a capability that is increasing and will continue to increase through 1985 and 1986 if we do nothing. Doing nothing after 14 years is simply going to give them an opportunity to increase their capability, which goes unresponded to.

High moral ground? I do not believe there is any higher moral rectitude to saying disarm ourselves or let us not be prepared or maintain a deterrent in the face of another capability which is superior.

I think indeed that the morality of the issue lies on the part of the proponents who would say, and who do say through the Leath substitue that we will produce none of these weapons but since we have had 14 years without interruption of our not producing anything we will return to and produce them after this priod of time unless you have reached an agreement with us.

I think this is the prod which is indeed necessary to produce meaningful negotiations and a successful negotiation in order that we produce that paradox of deterrence. It means having a capability; having the capability is the incentive to the negotiations that we all want to succeed.

* * *

Mr. Zablocki: Mr. Chairman, Members of the Committee, I do not intend to take the entire 5 minutes. But in this debate we have heard accusations and allegations about who is right and who is wrong. And of course the Comptroller General's report to the Committee on Foreign Affairs entitled "Chemical Warfare, Many Unanswered Questions," was quoted by both sides and misquoted and taken out of context on many occasions. But I must call to the attention of the Committee and the Members of the House that this lobbying document that DOD has prepared and distributed to the Members, is full of distortions. Let me just read one of the first charges. I am not going to try to answer every one of their allegations or misstatements.

Mr. Chairman, here is what it says in the very first sentence, very first page: "The report was prepared for the House Foreign Affairs Committee and reflects the views of the committee staff."

Now who among you can possibly believe that a Government agency, an independent body that serves us, would be listening to and writing a report dictated by any staff?

Now does this not exemplify how little value must be given to the rest of the comments and charges in this DOD lobbying document?

I said I was going to be very brief. Let me just go to the bottom of the page. I could comment on every one of the 12 pages, but I will not. On the bottom of the page the charge is that the GAO has stated that binary production might complicate arms control. And DOD responds: Key precursor chemicals are identical for binary and unitary and verification is equally difficult. GAO report is not a useful guide to the issues involved and reflects methodolgy and poor staff work.

Well, I would say this is the kettle calling the pot black.

Let me just quote Dr. Fred Ikle, our current Under Secretary of Defense on this question of the effect of binary production on chemical arms control efforts. Several years ago he argued that production of binaries would undermine rather than enhance the prospects for achieving a chemical warfare ban. That is your own Under Secretary of Defense. I will quote him accurately. He stated:

> If we start on a new type of production program it becomes even harder to envisage construction arms control agreements limiting competition in chemcial weapons.

Mr. Chairman, the Committee on Foreign Affairs has just received this letter from the Comptroller General of the General Accounting Office, Mr. Charles Bowsher. This letter, which I request be made a part of the record, states quite clearly that "the chemical warfare report was prepared with the same objectivity and independence that GAO applies to all its products."

Defending Nerve Gas

Colman McCarthy

Before discussing the killing and choking of chemical warfare, Theodore Gold, the Reagan administration's chief promoter and explainer of nerve gas, offers cookies to a visitor. In Gold's Pentagon office, which is secured by a combination-lock door, the work of defending America against aggressors also includes cookies as a deterrrent force against sneak attacks of the mid-afternoon hungries. Wherever the war, the Pentagon mounts a united front.

Gold, who is an athletically trim 43 and an engineer who has worked in the research and development of nuclear weapons at a Livermore, Calif., laboratory, has been the deputy assistant to the secretary of defense for "chemical matters" for 16 months. He has been busy of late, with Congress about to decide whether to lift the 14-year freeze on production of nerve gas.

As the human face and breathing body behind the Pentagon's managerial tests that speak of chemical warfare policy, Gold appears at first to be miscast. He has a companionable, I'm-just-a-regular-guy manner, and the talks of his "special abhorrence" of chemical weapons. He has been consistent. In May 1982 he said that, "if ranking weapons on their immorality, nerve gas would be at the top of the list."

These are the required protests, as standardized in Pentagon lingo as rifle salutes at a general's funeral. After these proper anti-barbaric references, it was Gold's time for fun: cool-headed distinction-making. Imperturbable, he is good at distancing himself from what he calls "the smoke and noise" of the debate.

He welcomes the challenge of confronting opponents of nerve gas like Sens. David Pryor (D-Ark). and Mark Hatfield (R-Ore.) who argue on moral, economic and tactical grounds against removing the nerve-gas ban. The Pentagon, says Gold, should pass muster on the obligation to be clear-headed in making its case: "We're now saying that we want to resume production after a freeze or whatever you want to call it, and the burden should be on us to say what has changed and why we want to produce it."

The changes on Gold's mind are those of the 1970s, when, with the United States out of the nerve gas race, the Soviet Union roared ahead. Now, according to Pentagon claims, its chemical arsenal is large and its troops well-trained in using it. The only purpose in getting back to even, says Gold, "is to deter the other side. The only way we know how to deter is to have a strong protective posture and also the abilty to retaliate if he uses them first. That would put him into a protective posture and then he would see no advantage in using them."

Gold has been trying to persuade Congress that the current stockpile is not adequate. He is not shy about telling his dovish opponents that they, not he, are the graver threats to peace. Gold told a House subcommittee in April that

if our current stockpile is inadquate "then failure to redress this situation makes war more likely, makes escalation to more terrible forms of war more likely, and makes arms control less likely."

Gold gives shorter shrift to an April 1983 report from the General Accounting Office that criticized the Defense Department's poor case for pushing ahead with chemical war preparations. The claims made for modernizing nerve gas weaponry, said the GAO, "are not supported by empirical evidence and must be recognized as possibly inaccurate."

Gold dismissed the report as shallow, worthy of an F if he were a teacher grading it. For many in Congress, the GAO's investigation earned an A-plus. It persuaded them that the Pentagon should be denied money for nerve gas. The House had a 14-vote margin against production and the Senate tied 49–49, with vice President George Bush, pro-nerve gas, breaking the tie.

In the end, Gold, though quick of mind and a relisher of debate, couldn't do much more than rely on the deadweight phrases in the Pentagon's promo sheets for its other weapons: the Soviets have superiority; we need deterrrence; we need a bargaining chip; we must send our enemies the right signal.

Gold's thinking was on the mind of Pryor, the leading critic of nerve gas, during the Senate debate. Pryor said: "on weapon after weapon, cause after cause, this seems to be the mentality of this city, of this town, of this administration, of this Congress, of all of us—let us build more, let us produce more, so that ultimately we can have less."

Gold didn't have a high regard for Pryor as an intellectual opponent. Apparently, the Arkansas senator lets his emotions enter the debate. He can't distance himself. He's unmanagerial.

Editors' Note: The chemical warfare issue did not die after the conclusion of the congressional action, as the following article from the *Washington Post* explains.

Pentagon Again to Seek Funding for Nerve Gas

Fred Hiatt

The Defense Department, renewing the only major battle it lost in Congress last year, again will seek approval to begin producing chemical weapons, Pentagon officials said yesterday.

The Pentagon's fiscal 1985 budget request, to be made public early next month, is expected to allocate slightly more money for nerve-gas production than the $114.6 million sought in fiscal 1984, they said.

After a long, seesaw battle last year that included two votes by Vice President Bush to break ties in the Senate, Congress refused to appropriate the funds.

The administration's decision to try again is certain to spark another emotional fight on Capitol Hill, where even many hawkish legislators consider nerve gas too ghastly a weapon. Administration officials do not disagree that chemical weapons are abhorrent but argue that only a modernized U.S. stockpile will deter the Soviet Union from using such weapons in a conflict.

"It's an issue that just won't go away," said a congressional aide who fought the funding last year. "It has come up before after it's been given up for dead."

The United States has not produced chemical weapons since 1969, and nerve-gas opponents in Congress say the moratorium should be continued. Opponents contend that existing U.S. stockpiles are sufficient. These probably total about 35,000 to 40,000 tons of chemicals in mines, artillery rounds, bombs, mortar rounds and rockets.

Sen. David H. Pryor (D-Ark.) and Rep. Ed Bethune (R-Ark), two of the leading congressional opponents, represent a state where the nerve gas would be manufactured, at Pine Bluff Arsenal. Bethune argued last year that the United States should not cede the moral "high ground" it has staked out by not producing the weapons.

"His feelings on the subject have not changed at all," a spokesman for Bethune said yesterday.

Administration officials said the stockpile is obsolete. Many such munitions are in danger of leaking lethal gas, they said, and others are intended for use with weapons that the Army no longer uses. The administration would like to produce two new munitions, Bigeye chemical bombs to be dropped from airplanes and 155mm artillery shells.

Both would be "binary weapons," whose lethal punch is formed by the mixture of two relatively harmless chemicals shortly before use of the weapons. Proponents said binary weapons would be safer to handle, transport and eventually dispose of.

The administration also maintains that the Soviet Union has used chemical weapons, known as "yellow rain," during fighting in Afghanistan and Southeast Asia. Those claims, based on refugee reports and analyses of contaminated samples from those regions, have been disputed by experts outside the government.

The Soviet army also is engaged in large-scale training for offensive and defensive use of chemical weapons, administration officials said.

They said that a modern U.S. stockpile might deter Soviet use of the weapons and that absence of useful chemical weapons could force the United States to use nuclear arms in response to a chemical attack.

Some Pentagon officials also said that, while they would never use chemical weapons first, the munitions can be useful for immobilizing large groups of troops. They cause nausea, vomiting, loss of vision, difficulty in breathing and death.

9

Pussycats, Weasels, or Percherons?

Current Prospects for Social Science Under the Reagan Regime

Peter H. Rossi

The purpose of this paper is to make an assessment of where the social sciences stand today in Washington, D.C. and what this standing means for the kinds of things social scientists will be able to do that depend on the existence and strength of a flow of funds from federal agencies. I emphasize the social sciences because, the diversity among social scientists notwithstanding, we are all being lumped together. The interdisciplinary variance is great, as we all have experienced, and the gulf is also large between applied social science and basic social science. But we are all being hit by the same solid waste matter nonetheless.

Clearly, these are not the best of times, but neither are they the worst of times. Although liberal social scientists no longer sit close to the high tables in Washington, there are some spiritual compensations; seated at those high tables are characters with whom one might not want to associate too closely in any event. So there is some good and some bad in the present situation, as well as some encouraging signs, as I will point out.

On the down side, the high levels of financial support to which we have become accustomed in almost two decades of free-flowing federal monies are being cut to very modest levels and, in some cases, threaten to disappear entirely. The social programs to which much of the support for applied social science work was attached are also being cut back, sometimes to the bone and beyond. The welfare-social science "industrial complex" is clearly suffering, and social science departments in universities are suffering as well.

From Peter H. Rossi, "Pussycats, Weasels, or Percherons? Current Prospects for Social Science Under the Reagan Regime," *Evaluation News*, 1983 4(1), 12–27.

On the up side, we stand to gain some self-esteem in these times. After all, one is known not just by one's friends but also by one's enemies. Hostile conservatives provide liberal social scientists with honor when they attack us. If we are dangerous enough to be denounced and singled out for special torment, we are thereby honored.

Back in the McCarthy era, we noted (when it was all over) that there were some surprising benefits to having been hauled before the Senator's Committee in Washington or its counterparts in the hinterlands. Although those occasions were surely traumatic personally and sometimes quite damaging to one's career, they also provided opportunities for demonstrating heroism, a feature of the experience that loomed larger as the McCarthy era slipped further into the past. Indeed, the fact of being summoned (and especially of having been summoned) was visible demonstration of one's importance and professional stature. That was so much the case that some of my colleagues at Chicago invented a new clinical entity to cover the feelings of jealousy that afflicted those who had *not* been called before a legislative committee and grilled about leftist leanings. The clinical entity was called "subpoena envy," manifested by lowered self-esteem, caused by the invidious comparisons drawn between those who had been worthy enough to come under attack and those who had not been important enough to have been issued a subpoena.

A necessary word of caution: I do not mean to suggest that there are any close parallels between the McCarthy period and the present. There are vast differences, and more would be lost in clarity and understanding than would be gained in any attempt to draw parallels. In contrast to the McCarthy period, the current assault on the social sciences is not directed at individuals but aimed at us collectively. Furthermore, the entire science R&D establishment is under attack to some degree, a source of small comfort of the misery-loves-company variety, but also provides the opportunity for more support from a wider base. In addition, much of the attack on the support for social science is a fallout effect of an attack on the social programs put in place over the last few decades, not a fallout effect of the Cold War.

There are three aspects of the current situation to which I will address my remarks. First of all, I want to review the kinds of attacks on the social sciences that have been made by the current administration. Second, I will try to interpret these attacks, dwelling on the views of social science that they appear to manifest. Finally, I will try to assess the

future prospects for the social sciences, net, I hope, of the general prospects for science support.

WHAT DID THEY TRY TO DO TO US?

Attacks on the social sciences are nothing new. Sociology, psychology, political science, and anthropology—and the corresponding engineering branches, social work, planning, mental health, and so on—have been under attacks for many years. It seems there have always been members of the Flat Earth Society in Congress and among staff members in almost every administration. For example, Representative Ashbrook of Ohio over the past few years would annually introduce amendments to the NSF authorization or appropriations bills abolishing the behavioral sciences basic research program in NSF. Among other things, Ashbrook claimed that the research funded by the NSF sociology program was both too theoretical and too dangerous. Theoretically oriented research was too abstract to be useful. Dangerous research consisted of providing legitimacy to deviance by studying it: Thus, a study of homosexuals, according to Ashbrook, implicitly endorsed homosexuality and studies of divorce increased the divorce rates. Ashbrook's remarks were as cogent as the statement alleged to have been made by a spokesman of the tobacco industry that cancer causes cigarettes or that fire engines cause household fires because they are always to be found at the scene of such events.

In the Senate, Proxmire has long been scanning the social science grant announcements to find those with the appropriate twist of interpretation that could be held up to ridicule. Indeed, there are plenty that require just a touch of distortion to make headlines in the *National Enquirer* or to be the subject of a press conference for Proxmire in which another Golden Fleece prize would be awarded.

Up to two years ago, the Ashbrooks and the Proxmires annoyed us with their sallies but rarely were able to inflict fatal wounds. But the advent of the Reagan administration brought about much more serious attacks on the social sciences. Of course, we had some intimation earlier than the inauguration that the new administration would scarcely be sympathetic to social science. We learned before the election that a new set of house intellectuals would be welcome at the White House and

called on by the administration. Irving Kristol and the Free Enterprise Institute were asked to prepare position papers on the social sciences. The Hoover Institution at Stanford showed some promise of replacing the Brookings Institution in Washington as the source of policy papers on a variety of social issues. All these moves were early signs that we would be in trouble, since these groups either held normal social science in some contempt or advocated a social science perspective that was scarcely mainstream liberal.

But the major shock came with the unveiling of the administration's 1982 budget. Funds for the NSF Division of Social and Behavioral Sciences were to be halved for that year and ravaged by similar proportions in subsequent years until the budget was reduced to zero. This proposal essentially torpedoed the grants for basic research in sociology, social psychology, and economics that were funded through NSF. In addition, grants for social research funded through NIMH were to be completely abolished—NIMH was instructed not to fund any additional social research. The National Institute of Drug and Alcohol Abuse was to be stripped almost to nonexistence, and responsibility for alcohol and drug abuse programs were to be shifted to other agencies—drug abuse programs were to be handed over to the Drug Enforcement Administration, for example.

Similar cuts were proposed in the social research programs of other agencies. The Department of Labor's Employment and Training Administration research program heavily devoted to evaluation was brought to a virtual halt. In the Department of Agriculture, evaluation studies of programs that were scheduled to be axed or substantially reduced suffered severe losses. The applied social research industry almost immediately fell on hard times. The larger firms suffered as well as the smaller ones. Everyone was in trouble across the variety of disciplines within the social sciences as well as both the applied and basic branches of the associated research activities.

There were some surprising exceptions to the general slaughter: Research on crime that was supported through the National Institute of Justice went unscathed. Indeed, a budget cut scheduled under the Carter regime was restored and there were hints that research support might be increased in later budget years. Apparently criminology was not a social science activity in the same sense as NSF's social science program. Similarly, social research supported through the Department of De-

fense's ARPA program was also not touched—except to receive additional funds.

Of course, the 1982 administration budget proposed cuts in most civilian agencies and hence there was a lot of company for social science's miseries. But we also were scheduled to suffer more than our proportional share. Apochryphal stories began to spread through the social science communities that the social sciences were being targeted for cuts on ideological grounds. For example, it was said that David Stockman and/or members of his staff had stated, off the record, that an objective was "to get those pinkoes out of the federal trough."

Now it has been several decades since anyone high in a presidential administration has called any of us pinkoes. Veterans of the McCarthy era searched through stockpiles of ironic quips, gallows humor, and sick jokes to find basic formats that with appropriate modifications could be responsive to the current situation. About the best we could come up with was this riddle. Question: What do the following have in common: Increased support for the missile stockpile and cuts in the social sciences? The answer: This is all part of an overall strategy to restore geopolitical parity between the Soviets and the U.S.A. The missile program is being given funds so that we can catch up with the Russians, who are way ahead of us in stockpiling missiles. In the same way, the social sciences are being cut back in order to let the Russians catch up with us and thus restore the geopolitical balance in social research.

We all know now that the attack on the social sciences did not succeed entirely. The NSF behavioral and social sciences budget was restored, and the program has done well in the recent budget process. Although NIMH still is prohibited from funding social research, semantic adjustments have been made in programs and research grant application language that promised to evade some of the impacts of this prohibition. On several occasions, representatives of OMB have stated that they thought the attack on the social sciences had not been carefully thought through. Congressman Ashbrook died earlier this year, so attacks will not be coming from that quarter in any conventional manner. There are also some recent signs that prominent conservatives in the House and Senate have come to the realization that some kind of social science research is useful to have around.

It is beginning to look as if we are in the same boat with other groups that are suffering from the successes of the general effort to cut back on social programs. It is nice to be in the big boat and not in a separate

sloop with a big hole below the water line, as appeared to be the case in 1981. Note, however, that the big boat is sinking a little bit at a time nevertheless. Social Science Research—pure and applied, small and large scale, of whatever disciplinary persuasion—is in for troubles.

WHY WERE WE ATTACKED?

The Reagan administration's direct attack on the social sciences appears now to be considerably muted, at least for the time being. But there are still residual efforts that have strong ideological overtones. Research agencies such as NIE are being pressured to appoint to review panels persons with credentials that are more appropriate to the fundamentalist Christian ministry than to judging the internal and external validity claims of social research proposals. A recent announcement of a grant program for programs and researches into adolescent pregnancy specifically barred applications from any agency that in the past had offered adolescents advice on abortions.

Social scientists tend to be liberals. Of that there is no doubt. Indeed, as the faculty opinion surveys undertaken by Lipset and Ladd have shown, majorities of social scientists have favored the center and left of center consistently in support for presidential candidates and other offices. Among the faculty surveyed by Lipset and Ladd, the more productive social scientist faculty were even more liberal than the less productive, and researchers were the most liberal of all. Within the social sciences, sociologists were found to be especially liberal in their political outlooks, and economists the least liberal. But there are, of course, differences within each group. There is a strong Marxist wing within sociology, but there are also many who have supported quite conservative views and candidates. Indeed, mainstream social scientists have scarcely been radicals of either left or right: I venture that most of the NSF grantees as a group are likely to be a little to the left of Lyndon Johnson and about as dangerous in their ideologies as the *New York Times* or the *Christian Science Monitor.*

Why did the Reagan administration take after what are essentially pussycat liberals? The answer seems to lie in the undifferentiated view held by the conservatives of what the social sciences were all about.

There were several sources of this confusion, as follows:

First, there was a confusion between social critics and social scientists. The social scientists they wanted to get were those who produced the dangerous ideas that were undermining society—the welfare state advocates, the advocates of open marriages and easy divorce, not to mention free abortions for welfare mothers, and those who were pushing for equality for women and homosexuals, who were for mollycoddling criminals and deviants, and so on, through the lists of things that represent the most "progressive," permissive, and supportive styles that some believe our society should adopt.

Some of the advocates of these viewpoints are social scientists, but many are not. Some of the research being supported by the federal agencies was either on such topics or illuminated by a concern that the directions of social change be bent toward those ways. The most direct target, the social and behavioral sciences program at NSF, had rarely supported anything but mainstream projects, usually quite neutral or even conservative (by sociology's standards). And the program included economics at NSF, a stronghold more of free market ideas than of free love ideas. Somehow the administration was misled by its advisers into redlining any social science that showed its head. Type I errors dominated.

Indeed, it is almost as if the views of the administration were adopted from those of the worst sector of the mass media. As you all know, the easiest way for a social scientist to get into the mass media is to have some kinky or counterintuitive statement to make. Indeed, one of the main reasons there are stereotyped views of welfare mothers portrayed in the mass media is because the media gatekeepers do not find it interesting if you were to find that many persons who are on welfare are nice middle-class ladies who, after divorce, were left high and dry with their children. That is not a newsworthy item. It is a newsworthy item, however, if you were to find one or two welfare families who could trace their descent from a line of welfare mothers and grandmothers and great grandmothers. If you study homosexual marriages, reporters will hang around your computer terminal to get a line on your latest results; if you study the retirement adjustment problems of heterosexual couples, no one from the *National Enquirer* will be pressing you for your results.

One of the major consequences of this view of the social sciences held by the administration was that it was not aware of the essential workhorse roles social scientists play in providing essential intelligence

on how the whole society is doing and on how public agencies are functioning, let alone whether specific programs were effective. Indeed, as you all know, much of the public policy and applied work of social scientists is done quietly, without the glare of publicity, work that rarely reaches the notice of the *National Enquirer* or of Senator Proxmire.

The initial attack of the administration on the social sciences was driven by a series of massive misunderstandings, as follows:

(1) that the social critics occupying left of center were social scientists;

(2) that most social scientists were considerably to the left of center;

(3) that social scientists and their research were undermining the moral fiber of American society: Deviance and disobedience were being celebrated—drugs, premarital, extramarital, and kinky sex were being pushed by the social scientists (read leftist critics) along with divorce, crime, and generous handouts to the slothful and indolent.

In short, the administration had social scientists tagged as subversive weasels, generously supported by taxpayers' money, while sneakily gnawing away at all that was good and moral in American society.

"NORMAL" SOCIAL SCIENCE: THE WORK OF PERCHERONS

Members of the Evaluation Research Society and the Evaluation Network may not need to be reminded individually of their own work and how different that work often is from the stereotypes of social science held by the right-wingers of our society. But we may need to be reminded of the full scope of such work and the extent to which it has become so deeply embedded in our society and its central political institutions that many of us may not fully appreciate the extent of such incorporation. Certainly the administration was not aware of the scope of social science operations within the government and within our society generally.

I have found it useful to classify social science research—either basic, discipline-oriented work or applied social science—as falling roughly into one or another of three basic classes: descriptive research, model construction, and evaluation research. Of course, no single piece of research is likely one or the other; rather, almost every social science

research project is usually a mixture of the three, varying in the precise proportions of each as well as in size, quality, and other attributes that have been applied to social science work.

Perhaps the largest category of social science research is primarily descriptive in intent. Descriptive research can be identified as research that is concerned with the following issues: How many? How much? How distributed in physical and social space? What are its characteristics? The substantive concerns of such research range widely, from simple body counting, as in the census, to the more subtle charting of attitudinal trends. The location of such research is also wide-ranging: The largest descriptive operation of all likely is in the Bureau of the Census, with mandates to run the decennial census of population and housing and also the massive sample/surveys that provide monthly labor force measurements, annual housing inventories, and a variety of periodic and episodic operations. Whatever its political origins in the struggle between large and small states for "proper" representation in the Congress, since 1850 the Census Bureau has become more and more a social science operation to which a variety of social science disciplines have contributed.

Outside the Bureau of the Census are literally scores of social science census-like and sample/survey operations designed to monitor the ways in which our society and its parts are working. Marketing surveys sponsored by businesses chart the trends of consumer preferences and intentions and explore the potential markets for new products or the receptivity potential to new marketing strategies. Government agencies explore the extent to which the general public will accept new strategies in public policies.

Much of the social science research that is primarily descriptive has become so routinized and incorporated into normal operations of public and private agencies that often it is not recognized as essentially social science in character. I am sure that few persons would designate the Census Bureau when asked which is the biggest social science research operation in the country.

Although the administration did take a few swipes at the Bureau of the Census and at other ongoing descriptive social science operations, it was not because the Census Bureau was potentially subversive or frivolous, but because in the absence of political experience, administration members did not recognize the extent to which the running of the country was dependent on good data on a variety of parameters. Nor

was the administration apparently aware of the extent to which the nonfederal public sector and private sector were dependent on such information, ranging from the mayors of municipalities who disputed the distribution of block grants and needed up-to-date poverty counts on their towns, to the manufacturers of dishwashers and air conditioners whose marketing departments needed to know how deeply into the potential markets their efforts have so far succeeded in going.

Model construction, the second major category of social science research activity, is concerned with the development of relatively abstract models of how things work and testing those models in the real world. At the margins, it is not easy to distinguish model construction from basic research, on the one hand, and descriptive research, on the other. Social scientists who use the public use Census Bureau tapes to test theories of status attainment are hard to distinguish from policy analysts who use the *Uniform Crime Reports* data on arrests to test deterrence models: Both are using primarily descriptive data to test relatively abstract models, neither of which may be particularly policy-relevant.

Social science models permeate current social policy and that of the immediate past. Models of the economy lie behind the collection of the money supply indices that the Federal Reserve board monitors so carefully and tries even more desperately to control. Implicit microeconomic theories of how to motivate individuals to enter or leave the labor force lie behind the incentive systems that are at the core of welfare programs.

When social science model construction includes policy variables, variables that are related to what parameters of a system could be influenced by the range of acceptable alternative policies, model construction can also be an applied research activity, as in the case of the TARP experiments on released prisoners or experiments with structured work.

Model construction and testing that are discipline oriented are hardest to justify to the Senator Proxmires and to members of the administration (in fact, any administration) because the variables are seemingly unconnected with "real" life. This is the source of the often voiced complaint of unrelatedness. But in the long run (whenever that is), we will be best off if model construction can proceed whether or not directly policy-related. At the least, we need to know which variables to control in policy-oriented research. At best, better models may

provide promising leads to policy changes that will improve the functioning of our society and its institutions.

The basic social science models are the potential sources of much of the policy-oriented research we may want to carry out. We need to understand what we should be counting in descriptive studies. Thus, understanding more about crime can lead to better crime indicators, or understanding more about unemployment and employment may lead to better monitoring of the problem in our society. We also need basic theories and knowledge in order to build the policy oriented models, as indicated earlier.

The third variety of social science is at the core of the concerns of the Evaluation Research Society and the Evaluation Network. Evaluation research, of course, did not start with the founding of these two professional associations, nor would it end with their disbanding, should that occur. Humans are evaluating animals, continually assessing the worth of their own actions and of occurrences around them. What is particular to evaluation research is that it is a self-conscious application of social science ideas and methods to the study of collective (or public) actions. Although its development in recent years expresses the maturity of the social science research methods, it also expresses the historical peculiarities of this period—namely, that our society and particularly its decision makers are no longer sure about what they are doing and are skeptical that theirs is an obviously correct and effective set of ways to accomplish a given social end. The liberal and conservative philosophies that have dominated federal administrations and congresses over the past 20 years have not been fully confident that they knew exactly what they were doing and how to accomplish their ends. A tentative and skeptical commitment accompanied the enactment of the social welfare program of the 1960s and 1970s, an orientation that predisposed decision makers both to commission evaluations and to pay some attention to their findings. Indeed, one may venture the self-critical judgment that our potential audiences were even extraordinarily charitable, retaining their faith in evaluation research despite our frequent fumblings and failures.

If this analysis is at least partly correct, then the question quickly arises whether or not the current administration and its ideologues are as skeptical about their policies and programs as previous administrations. Certainly they are skeptical to the point of outright hostility to the programs and policies that were put in place by the last few administra-

tions. Whether they would be as skeptical about their own programs is not clear. Indeed, evaluators more easily may find employment as hatchet persons, given the mission to clear out the underbrush of leftover programs from liberal regimes. But there may be little work for the future, at least if the administration can have its way entirely. Fortunately, Congress apparently retains its skepticism, as the recently expressed congressional support for evaluation research may portend.

In any event, the short-run fate of evaluation rests not so much on its social science character as on whether or not the climate of skepticism concerning all public programs changes under the influence of a new administration adhering to ideological strains that are supported by much strong faith.

Normal social science activities are integrally incorporated into the working of our complicated federal, state, and local system. Social science intelligence monitoring social trends and the workings of public agencies is indispensable. Least likely to disappear of all social science activities, and more likely to enjoy at least the level of support they have in the past, the descriptive studies funded at all levels of government are likely to continue, simply because their outcome is necessary for the functioning of any regime.

Model construction—or social science theory—is more precariously positioned. Although basic social science is clearly the foundation for applied work from variable specification through measurement to analysis, the foundations may be laid so long before the building is erected that connections may not be obvious. Furthermore, some of the theory building is vulnerable to ridicule, a style of commentary that is often devastating.

Evaluation research occupies a middle ground as far as security is concerned. It is clearly needed to provide information on policies and programs about which the society is skeptical but unwelcome in arenas where true believers rule.

In short, the workhorses or percherons of social science are welcome when they do not appear to be social scientists—as in the case of the Census Bureau—but not welcome where there is no work—as in the case where policies and programs are accepted or denied on the basis of their doctrinal purity—and uneasily accepted when they are playful and speculative—as in the case of theory building.

WHAT OF THE FUTURE?

What does all of this signify about the roles social scientists may play in the policy arena in the near future? Some of the future has already been hinted at broadly in the discussion in the last few pages, but it is worthwhile to draw out and clarify a few things.

First of all, as indicated earlier, the initial (and somewhat rabid) attack on the social sciences has abated. Those of us who were looking for martyrdom will probably not achieve that state of grace, and those of us who wanted martyrdom and would not have been granted it are likely spared the agonies of subpoena envy or its 1980s counterpart. The current budget contains few direct attacks on the social sciences, and the mutterings about getting "those pinkoes" appear to have died down.

True, there is less support for basic social science research and for applied activities of all sorts, as we all have noted, but this is not the result of a singling out of the social sciences for especially hard times; rather, it is a result of the general hard times for nonmilitary parts of the budget all told.

Does all this signify a change of heart on the part of the administration and of right-wingers? I don't believe it does. Rather, it appears to be an increase in knowledge about what constitutes social science. The changes were of two kinds. A differentiated view of social science replaced the unitary view of social scientists as decadent liberals and dirty-minded advocates of morally corrupt social change. The differentiated view recognizes the legitimacy of social science activities that are concerned with providing internal intelligence on the functioning of our society and its central public institutions (except the military). The view of social scientists as decadent pussycats and sneaky weasels has been replaced with an understanding that there are brave and loyal percherons who do a lot of useful work, whatever the regime may be.

Of course, it helped change the image that the social sciences managed to show some degree of solidarity one discipline with the others and ability in reaching members of Congress. The organization on short notice of COSSA, The Council of Social Science Associations, was impressive. Social scientists throughout the country contacted their representatives and senators and explained to them the impact of proposed cuts on our ability to deal with pressing social problems. It also helped that the other disciplines came to our support, as in the declarations of the AAAS and National Academy of Sciences. The

image of social science as a disorganized horde of politically weak pussycats was replaced by one in which we were seen to have some support.

But the most important thing to keep in mind about the future of social science research is that whether or not the conservatives like us, they need our percheron qualities to provide workhorse types of research on how the system is going. Furthermore, they will have to rely on liberal social scientists to provide that information because there simply are not enough conservative social scientists to go around. The Bob Jones universities have not hired social scientists who have the skills to do what is needed.

Based on the considerations I have laid out in this paper, I foresee the following trends for social science research over the next few years.

First, we will be in some trouble, with leaner times ahead, but not more than other nonmilitary sectors. Of course, social science work for the military will grow, and some of our efforts will be thrown into such substantive problems as plague the military. I venture, however, that there is not enough military social science to keep us in the style (and funds) to which we have become accustomed.

Second, basic research (i.e., research not clearly directed at specific applied problems) will still be regarded suspiciously and suspected of being subtly subversive. Social scientists will not be trusted to define their own problems because what we define is too closely tied, they believe, to our liberal ideology. This is the meaning of the changes we are seeing in the composition of peer review groups, the most notorious being those of the National Institute of Education. Academic redoubts with conservative ideological climates will be scoured for social scientists with curricula vitarum that are at least a shade above being ridiculous. Grant-giving agencies will be pressured to appoint some representatives from among this group of ideologically anointed.

Of course, some of you in the audience will ask whether this is so very different from what happened when there were liberal swings in political climates in the past. There is some substance to this viewpoint. The liberal elite universities lost a great many of their faculty to Washington when Kennedy succeeded Eisenhower. But there is a big difference: However muddle-headed were the liberals and however easily and copiously their hearts bled, they were competent, percheron-like workers. Our census is one of the best in the world, courtesy of the

liberal Ph.D.s who built it during the depression and who maintained control over personnel quality all through the Eisenhower years.

Third, applied social research will gradually come back to somewhere near its former strength as the administration and Congress realizes that applied social research is a necessary function to be fulfilled whatever the ideological character of the regime. It may well be, as I speculated earlier, that applied social research will be forbidden to touch some sacred areas and narrow its focus in other ways. A prime example of the trend I have in mind is the announcement earlier this year of a program of action and research on adolescent pregnancy. The announcement of the action grants contained a prohibition on any applicant organization that had ever engaged in delivering contraceptive services to young people or providing advice on abortions. The research part of the program emphasized that proposals would be considered only which emphasized research on how to prevent adolescents from engaging in sexual intercourse. Abstinence and accompanying self-control were to replace the liberal emphasis on how to get children to use contraceptives. Note the implicit view of social science in this research program announcement: Social scientists may be needed for this task, but they are not to be trusted to be given their own head.

In short, I believe that there will be support for social science research. But there will be restraints along the following lines: The goals of researches will be changed to be more in line with conservative views. There will be less freedom to define problems as we will.

To return to the title of this talk: It has become clearer that we are not the weak, incompetent, and superfluous pussycats the conservatives thought we were and, to some extent, *we* thought we were. There developed a surprising amount of strength in the social science community and among Congress. Whatever self-images we had as subversives—whether as weasels (as the conservatives saw us) or as heroes and heroines of progressive social change (as some of the social science community would like to view us)—it has become more clear that the bulk of social science is at best only mildly subversive. There are some weasels (or heroic figures) among us, but they can be differentiated from the bulk of social science research and social scientists who are more like percherons—strong, competent, able to take on a big load and do a good job.

It does not look as if we will be given, as social scientists, a chance at glorious martyrdom, a prospect that is comforting to those who do not

like crowns of thorns or bonfires, but disquieting to those whose identities and self-respect is forged out of persecutions. It does look as if we will not suffer more than the rest of the liberal establishment: We are in for leaner times, but social science research will not die out altogether, as our percheron roles appear to be essential and acceptable.

III

SELECTED APPROACHES TO EVALUATION

Our review of the literature revealed a strong trend toward increasing the repertoire of methodological approaches to evaluation. The general impetus for this trend is improving the effectiveness and utility of evaluation practice. A variety of approaches, representative of this trend, has been selected for inclusion in this part of the volume. An attempt was made to include articles that review the defining characteristics, history, methodology, merits, and limitations of a particular approach.

In the first paper in this section, Weiss reviews stakeholder-oriented evaluation. "Stakeholders" are persons likely to be affected by a program or persons who make decisions regarding the future of a program. Ideally, the stakeholder approach considers these persons' interests and priorities, and it emphasizes information feedback and dissemination to them, both while planning and while conducting the evaluation. The aims of this approach are to increase utilization of evaluation results and to improve the fairness of the evaluation process. As discussed by Weiss, efforts to implement stakeholder evaluations have thus far fallen short of these ideals, partially because some of the assumptions underlying this approach do not always hold true. Weiss provides an excellent critical analysis of these assumptions and, in doing so, presents a fair and realistic appraisal of the future of the stakeholder approach.

Multisite/multimethod evaluations employ a variety of qualitative and quantitative methods of data collection while focusing on sites, rather than individuals, as the unit of analysis. Louis's article, which is the introductory article of an entire issue devoted to this topic (Smith & Louis, 1982), reviews the key challenges faced by evaluators who want to carry out multisite/multimethod studies. These include collecting data across sites too geographically distant to be visited by the principal investigator, transforming and integrating qualitative and quantitative data, and conducting cross-site analyses. A broader issue addressed in this paper is the conditions under which the multisite/multimethod approach should be employed.

The third approach is meta-analysis. Briefly, meta-analysis provides a quantitative review of a research domain for the purpose of integrating the results of multiple studies. Meta-analysis may also summarize other study characteristics, including methodological, procedural, and theoretical variables. Strube and Hartmann provide a comprehensive review of the rationale, applications, and problems of this approach. They consider meta-analysis to be a means of taking stock of what we know, or do not know, in a given area.

Unlike the three approaches described above, the three remaining approaches reviewed in this part of the volume are well established in other areas of research (e.g., anthropology, marketing research) but have received relatively little attention by the evaluation community. First among these approaches is archival data research. Evaluators tend to avoid using secondary data sources because such data frequently prove too inappropriate, inaccurate, or inaccessible to meet the needs of the evaluation. Luckey, Broughton, and Sorenson foresee a future in which fiscal support for primary data collection for evaluation purposes is likely to decrease and, thus, where more evaluators and policy analysts may be compelled to use existing data sets. To facilitate evaluators' use of this approach, these authors specify its limitations and, more important, present a series of strategies for minimizing difficulties associated with archival research.

Cost-benefit and cost-effectiveness analyses are complementary methods of determining whether or not programs are efficient in their use of economic resources. As defined by Wortman, in cost-benefit analysis all benefits and costs are valued in monetary terms; in cost-effectiveness analysis, program benefits are valued in other units—for example, improved morale. Wortman believes that, to the detriment of their profession, most evaluators are unfamiliar with many of the issues concerning these analyses. His paper addresses the need of this audience to learn the general principles of analysis for cost-effectiveness and cost-benefit methodology.

Social impact assessment (SIA) focuses on the impact of policies, programs, or building projects on individuals, families, and communities. Morash describes SIA as anticipatory research that involves the prediction and comparison of a variety of potential impacts resulting from two or more program options. In addition to providing an overview of SIA methodology, Morash's paper demonstrates the application of SIA to the study of criminal and juvenile justice programs.

REFERENCE

Smith, A. G., & Louis, K. S. (Eds.). (1982). Multimethod policy research: Issues and applications. American Behavioral Scientist, **26**(1), entire issue.

10

Toward the Future of Stakeholder Approaches in Evaluation

Carol H. Weiss

Given the very special situation in the Cities-in-Schools (CIS) and Push/Excel (Excel) programs, the stakeholder approach to evaluation has hardly received a fair test. In fact, some people argue that it was implemented with such a minimalist interpretation of its scope that its potential benefits inevitably went unrealized. So many other difficulties beset the evaluation—primarily as a result of the attempt to apply formal quantitative assessment to shifting (and, in the case of Push/Excel, inchoate) programs—that the stakeholder approach did not have much chance to affect the course of events. On the positive side, perhaps it engaged the attention of actors who might otherwise have ignored it entirely, particularly people at the local sites. On the negative side, it probably diverted a fair amount of evaluators' time from strictly evaluative functions. But the turbulent nature of the programs and the mismatch with standard outcome evaluation procedures were probably the critical elements in both cases.

The inability to attain the expected benefits in these cases may have been the result of extraneous factors (incomplete implementation of the stakeholder concept, inappropriate evaluation strategies, fluidity of programs, and

From Carol H. Weiss, "Toward the Future of Stakeholder Approaches in Evaluation," pp. 83–96 in *Stakeholder-Based Evaluation* (New Directions for Program Evaluation, no. 17) Copyright © 1983 by Jossey-Bass, Inc. Reprinted by permission.

so on). Is it possible that stakeholder-oriented evaluation would work in other, more congenial settings? Or are there basic flaws in its underlying assumptions that inevitably limit its capacity to deliver what it promises?

Conversations with colleagues on this project—Robert Stake, Ernest House, Eleanor Farrar, Anthony Bryk, and David Cohen—have encouraged me to see the intentions of stakeholder-oriented evaluation as fundamentally threefold: first, to increase the use of evaluation results in decision making; second, to empower a wider assortment of groups to determine evaluation priorities; and third, to shift governance of evaluation from sole control by NIE to shared control, thereby reducing NIE's responsibility. Partisans of the approach may have had other expectations, such as providing greater legitimacy for evaluations in general, for NIE's evaluations in particular, and for NIE as the evaluation agency. But the three aims listed here appear to represent the nub of the stakeholder argument.

I interpret the term *stakeholders* to mean either the members of groups that are palpably affected by the program and who therefore will conceivably be affected by evaluative conclusions about the program or the members of groups that make decisions about the future of the program, such as decisions to continue or discontinue funding or to alter modes of program operation. These are quite distinct categories of people, although there is some overlap. I include them both as stakeholders, because that is my reading of NIE's intent.

Perhaps it would be useful to maintain and elaborate the distinctions. Analytically, stakeholders can be divided into four categories, depending on the kinds of information that are likely to be valuable to them (Figure 1).

Assumptions

With these prefatory remarks, let us try to disentangle the expectations inherent in the stakeholder notion. As developed at NIE, the approach makes a series of assumptions about the involvement of stakeholders in the evaluation process.

1. Stakeholder groups can be identified in advance of the start of evaluation.
   a. The sponsor, the evaluator, or both can figure out whose interests are at stake.
   b. The sponsor and/or the evaluator will select a representative set of groups to participate in the evaluation.
2. Stakeholders want an evaluation of the program with which they are associated.
   a. They want to have evaluative information available about the program.
   b. They are willing to participate in the evaluation process.

Figure 1. Categories of Stakeholders

Category	*Types of Decisions to be Made*	*Types of Evaluation Results That Are Relevant*
Policy maker (the Congress, the secretary of the Department of Education, local philanthropists, school board members)	Shall we continue to fund the program? Is it achieving the desired results? Shall we expand it or reduce it?	Outcomes of program for participants, causally linked to the intervention.
Program manager (national program staff, program directors in cities, program designers)	How can we improve the program? Should we recruit different staff, serve different kids, use different techniques?	Differential outcomes for different types of students, by types of service received, by type of staff, and so forth. Qualitative information on what is going well and poorly during implementation.
Practitioner	What shall I do to help Joan and Pedro? How can I get Elsie to try harder?	Usually not much, except perhaps for some overview of how the whole project is going. Practitioner's own knowledge and experience are more relevant and salient.
Clients and citizen organizations (students, parents, community groups)	Shall we keep attending the program (assuming we have the choice) and supporting it?	Not much. Outcomes of the program for previous participants should be relevant, but often the evaluation has not gone on long enough to provide such data. In any event, clients' own experiences are more salient.

3. They want specific kinds of information to help them make plans and choices.
   a. They can identify their information needs in advance.
   b. The kinds of information they want are the kinds that evaluation studies produce.
   c. The kinds of information that different stakeholder groups want can be reconciled with one another.
4. Evaluators will respond to stakeholder requests for information.
   a. They have the requisite time, resources, interest, and commitment to the process.

   b. They have the interpersonal skills to solicit realistic information requests from groups, even from those for whom evaluative information is not salient.
   c. They have the political skills to negotiate accommodations in priorities among competing stakeholder groups.
   d. They have the technical proficiency to design and conduct a study that produces valid data to satisfy diverse information requests.
   e. They will report back promptly, responsively, using forms of presentation that are appropriate to various audiences.
5. Stakeholders who have participated in an evaluation will develop pride of ownership in the conclusions.
   a. They will accept them as true.
   b. They will take them seriously.
6. Stakeholders who have decisions to make (mainly federal policy makers, school-district administrators, outside philanthropists, and program directors) will use evaluation results as a basis for decision making.
   a. Information in and of itself is a decisive component in decision making.
   b. The stakeholder approach makes a wide assortment of information available.
   c. The information is relevant to the situation that exists when decisions are being made. If circumstances have changed since the study was planned, the information collected remains appropriate to changed conditions and sufficient to answer current questions.
   d. Stakeholders who do not have program-wide decisions to make (principals, teachers, students, parents, community organizations) know that at least their criteria and concerns were taken into account in the evaluation and that information of importance to them was considered in decision making. They will therefore, perhaps, accord the decisions greater legitimacy.
7. (A revisionist assumption) Even if evaluation results do not sway specific decisions, they will enrich discussions about future programming and illuminate undertakings of program actors.

Analyzing the Assumptions

A number of these assumptions look perfectly reasonable—at least under reasonable conditions. Of course, if we push any one of them too far, we can find situations under which it will break down. Let us see which as-

sumptions seem generally viable and which depend on images of orderliness and rationality that rarely prevail in the program world.

The first assumption, that stakeholder groups can be identified in advance, looks feasible. The CIS and Excel evaluations seemed to have encountered little difficulty on this score. We could ask whether the groups that AIR identified and assembled were truly representative of all important stakeholder interests. For example, how actively represented were teacher, student, and parent concerns? How long did representatives of these groups continue to participate? There is also a perennial question about the representation of potential users of program information. Groups that are not actively associated with the program now can have a real stake in the information that evaluators will produce, such as school districts that would want to adopt the program if it proved successful. No procedural mechanisms appear capable of identifying, let alone representing, the entire set of potential users of evaluation results or the questions that they will raise. But in the normal course of events, adequate representation of stakeholders seems feasible.

The second assumption, that people want evaluations and that they will participate in them, probably holds good for some groups some of the time. Given a choice, however, it seems likely that many groups would forgo evaluation entirely. These groups have learned over past decades that evaluations are more likely to be the bearers of bad tidings than good. When results are circulated, they often pose a threat to the program rather than support and guidance. Information is a minor benefit compared to the questions and criticisms that it can provoke. Only when federal beneficence is contingent upon evaluation do many groups accept it as inevitable and come on board.

The third assumption, that people can specify their information needs in advance, has the same "maybe/maybe not" quality, although it lists toward "maybe not." As cognitive psychologists have demonstrated and as decision theorists have learned to their regret, people do not always know in advance what they will need to know in order to make a decision (Slovic, Fischhoff, and Lichtenstein, 1977; Slovic and Lichtenstein, 1971). Unless situations are routine and repetitive, the human cognitive apparatus is not always up to the task of foreseeing which information will be critical. Moreover, the assumption that evaluation requests can be defined early in the study relies on a vision of an orderly and predictable environment. It assumes that organizations can schedule their choices and calculate their information needs with confidence that things will go as planned. In fact, neither the political environment nor the organizational milieu is stable. Program decision making is beset by unexpected occurrences from inside and outside the organization. Long experience with the development of management information systems and with managers' inability to specify their needs correctly is instructive here (AAACMIS, 1974; Ackoff, 1967; Grudnitski, 1981).

The capacity to define information needs far ahead of time is limited by individual, organizational, and political constraints. Many people will make an effort to tell what they need to know, but much of what they say is a learned, stereotypical response. People in schools, for example, almost always say that they want to know test scores. Whether or not test scores are relevant to the program or useful in decision making, people have been indoctrinated to the notion that achievement is the central mission of schools and that neglect of test scores would therefore be unprofessional conduct.

Not uncommonly, people do not actually want to know anything. If you define these people as stakeholders and ask them to describe their information needs, they will generally give an answer. In today's information society, saying that they do not need data is tantamount to branding themselves as illiterates. So, lacking any clear need, they can take the opportunity to ask for information that they know will cast the program in a good light, such as data on the number of hours of service provided or on parents' satisfaction with the program. Another strategy is to ask for something, without regard for exactly what it is, because information is a scarce resource and therefore worth fighting for. If stakeholders are competitive groups in a competitive environment and if information is the counter in the game, then information is what groups play for—almost regardless of its content.

Thus, the assumption that stakeholders are reliable sources of information priorities is not a very good one if specification is required far in advance, as in most pre-post designs. It is much more plausible if the evaluation is a qualitative, illuminative investigation of program operation. In most qualitative evaluations, evaluators have ample opportunity to shift direction and to follow new questions as they emerge. They are not locked in to a set of measures that can prove to be irrelevant when the "post" time rolls around.

The fourth assumption, that evaluators will respond to stakeholder requests for information, needs to confront the fact that it takes a variety of skills and considerable dedication to be responsive. Under some circumstances, responsiveness might prove to be impossible. If stakeholders press demands for a great deal of information or if their demands are incompatible, the evaluators may be battered in the effort to satisfy all parties. Later or sooner, they may give up the effort to be responsive and assume unilateral control. The CIS and Excel evaluations demonstrate that being a responsive evaluator is an arduous task. Both evaluations also suggest, I think, that the task can be managed under favorable conditions.

The fifth assumption, that stakeholders who participate in a stakeholder-oriented evaluation develop a sense of ownership in the study, is open to considerable question. There is a good deal of experience with this particular strategy, since involvement of potential users in evaluations has long been a staple prescription (Eidell and Kitchel, 1968; Flanagan, 1961; Havelock,

1969; Joly, 1967). Many efforts have been made in the past to conduct evaluations in this style. Some have been successful in giving prospective users a stake in the findings, particularly when users are few in number and when there is relative agreement on most significant issues (Conway and others, 1976; Rothman, 1980). But even under these favorable conditions, many well-intentioned researchers have been unable to secure acceptance of the validity and usability of study results (Berg and others, 1978; Lazarsfeld and others, 1967; Rich, 1977).

It is when disagreement is rife that user involvement is expected to be especially important for winning the allegiance of discordant groups. If each group believes that it has had a say in designing the evaluation, and if each group believes that it has gotten the information that it wants, then presumably all groups will have common commitment to consideration of findings. But when disagreement is rife, the evaluators are caught in a bind. They have to resolve discrepant requests and conflicting advice, and it seems inevitable that they will disappoint one or another of the parties. Groups whose requests are disregarded will lose interest in the study and its findings; on occasion they may become overtly hostile.

Another concomitant of involving users in the planning and conduct of studies is that it gives users an inside look at how the study is done. A close view can engender disenchantment as well as commitment. Insiders know the weaknesses as well as the strengths of the research—the shortcuts, unreliabilities, missing data, contradictions in sources. Some develop considerable skepticism about the worth of the final report (Berg and others, 1978), and they have less allegiance to it than outsiders who were not privy to the compromises in data and method that were made.

Some stakeholders are likely to be happy with evaluation results and to feel a sense of pride, but their happiness can derive more from the support that the evaluation gives to their stake than from the part they played in the study. Groups that find their positions threatened by evaluation results can revoke their support if they see their crucial interests endangered, even if they did participate in the evaluation process.

In sum, experience suggests that participation in a study can increase support for the study, but only if certain conditions are met: One's advice has to be given due attention, one has to see the study as being appropriately and reputably conducted, and results must not threaten significant personal or organizational interests.

The sixth assumption, that people who take part in an evaluation will use the results as a basis for decisions, is constrained by the fact, noted earlier, that not all stakeholders have decisions to make—at least, decisions of the kind for which evaluation has much evidence to offer. But for those people who do make decisions, is it reasonable to expect that those who participate in the

evaluation process will be more likely to base their decisions on evaluation results? On the positive side, it is safe to say that these people know more about the study than they would have if they had not taken part. In that sense, there is a better chance that they will absorb the information and use it. Not to use it takes a conscious decision. They can hardly remain oblivious.

The stakeholder approach also assumes that the results are relevant. Several factors already noted limit the generality of this assumption—for example, the common inability to predict information requests accurately, calls from different groups for inconsistent evaluation designs and information items, the possibility that one group's advice will be disregarded or overruled. Still, the notion that participation will improve relevance remains plausible. For example, when local groups take an active part in the evaluation, the study is much more likely to address the concerns that exercise them.

The rub comes at the point of applying results to a decision. The usual expectation is that decision makers will use evaluation results to choose between alternative A and alternative B. Unless A and B are minor matters, evaluation evidence is not likely to be the decisive element. Decision making about issues of import, such as whether to continue funding a project, is basically a political process. In making such a decision, people have to consider a wide range of factors—who supports the program and how much clout the supporters have, what alternatives there are that can serve clients if the program is terminated, whether alternative programs are likely to be more successful, whose jobs are in jeopardy if the program ends, how clients will feel and how they will fare without the program, what community reaction will be, what costs will be involved, and so on. Evaluation results provide evidence on only a small number of relevant issues. Thus, even if the evaluation is conducted with the broadly inclusive sweep anticipated for the stakeholder strategy, it never addresses all the issues that have to be considered. Nor does it settle the issues that it does address in a conclusive way. Therefore, evaluative evidence about program operations and outcomes goes into the hopper together with an array of other concerns, information, allegiances, ideological proclivities, and interests. Decision makers have to reach an accommodation that satisfies many people on many dimensions. While evidence of program effectiveness is important, it probably never will be the sole determinant—or even the most powerful determinant—of political choice.

If NIE or anyone else expects the stakeholder approach or any other reform in evaluation practice to make research information the major basis for decision making, they are destined for disappointment. Too many other factors must be considered, too many other conditions must be accommodated, for information to play such a stellar role (Lindbloom, 1968; March, 1982; Weiss, 1973).

The seventh, revisionist, assumption is that the stakeholder orientation can increase the use of evaluation for purposes of enlightenment (Caplan, 1977; Pelz, 1978; Weiss, 1977). Responsive, relevant, well-circulated evaluation results can provide information that keeps people well informed about a range of programmatic issues. Evaluation results can provide evidence about what works well and what does not, about the kinds of problems that arise, and about the reactions of staff and students. They can challenge prevailing assumptions about a program and the theories of behavior that underlie it. They can suggest reinterpretations of past experience and help to make retrospective sense about what the program has been doing. Without dictating specific decisions, they can permeate people's understanding of program potentials and limits. Over time, such understanding can have significant influence on the aims that people set, the alternatives that they consider, and the directions that they take in future programming (Weiss, 1980). Use in this sense seems to be a realistic goal for stakeholder-oriented evaluation.

In its early presentations, the stakeholder approach resembled many of the educational and social programs of the past generation. Its high-minded intentions were yoked to untested practices, and it promised too much. Its advocates expected a relatively minor reform to accomplish grand objectives. As evaluation of social programs have demonstrated time and again, changing behavior is not a simple task. More temperate expectations for stakeholder evaluation would put the idea in perspective.

A Tentative Balance Sheet

Our review of assumptions inherent in the stakeholder approach suggests that none of them is open-and-shut. There is leakage at every step along the way. The chances that any one step will be fully realized are less than one—often considerably less—and the cumulative chances of achieving expected benefits decline multiplicatively. Prospects for significant gains in evaluation utility do not seem especially bright unless collateral changes are made in the substance of evaluations and in the structure of the programs. Simply pasting the stakeholder process alongside current practice involves acceptance of many existing constraints.

Over all, the stakeholder approach seems to hold modest promise for achieving modest aims. It can improve the fairness of the evaluation process. It can probably make marginal improvements in the range of information collected and in the responsiveness of data to participant requests. It can counter the federal tilt of many previous evaluations and give more say to local groups. It can democratize access to evaluative information. If stakeholder groups take an active role, it can make them more knowledgeable about evaluation results

and equalize whatever power knowledge provides. When many groups know the results of a study, the likelihood increases that the information will be absorbed and drawn upon in later deliberations.

However, the stakeholder approach will not ensure that appropriate information is collected. Stakeholders will not usually be able to specify the kinds of data that matter to them with much accuracy, and even when they can, program conditions and outside events will probably change before the data become available. By the time that stakeholders confront decisions, the evaluation will be able to provide evidence on only a fraction of the questions at current issue.

The stakeholder approach will probably not visibly increase the use of results in the making of specific decisions. For example, a philanthropist who sees a report of no success for a program that he supports may find that his participation in the evaluation process makes little difference to his decision about whether to continue support. He still has to think about the implications of his position on many dimensions. Nor will stakeholder evaluation bring harmony to contentious program arenas. It can elicit diverse views, but it cannot contain them. In fact, if differences are wide, the opportunity to stake out turf during the evaluation process can make people more aware of the conflicts that exist. Even if they can work out accommodations over evaluation priorities, accommodations over program issues will be no easier to arrange.

If the stakeholder approach has potential for improving evaluations, it also makes new demands. It increases the burden on evaluators, and it demands time and attention from groups associated with the program as policy makers, managers, planners, practitioners, and clients. Some of these groups—including, perhaps, evaluators—will find the experience illuminating and worthwhile, but it is likely that others will not. The approach will trade some people's heightened satisfaction for others' annoyance or frustration.

Questions for the Future

Some conditions of stakeholder participation can profit from further thinking. I nominate three issues for consideration: the definition of participation, the competing claims of a single study and several independent studies, and the mode of study design.

Participation. Which groups should be involved? Does it make sense to limit participation to groups that face decisions and care about information, such as funders, managers, and planners? Other groups have interests in the program that deserve consideration, but it is the program and its future that concern them, not information about the program. They want a voice in what happens, not in what data shall be collected. Evaluation planning is not necessarily the best forum to engage them. Participation with a more specifically

programmatic focus could effectively attract their participation and profit from their perspectives.

The inclusion of multiple groups in the evaluation process is an attempt to redress the inequitable distribution of influence and authority. But evaluation planning is a strange avenue for such redress. The stakeholder approach could be construed as a way of deflecting stakeholder attention from decisions that more directly affect them. Indeed, it almost appears to be a substitute for involving stakeholders in the making of policy. A Machiavellian mind could conceive of the stakeholder approach as a way to mire stakeholder groups (particularly powerless groups) in the details of criteria definition and item wording, while the powers that be go blithely on with decisions as usual.

Of course, no such demonic scheme is at work. The reasons for involving stakeholders in evaluation is that NIE has control of evaluation, whereas it has little voice in program decision making. NIE is taking advantage of the opportunity to broaden representation in the one domain over which it has authority. The intent is high-minded. But the actuality is that participation takes place at some remove from the center circle of program decision making. Whether a reduction of inequities in the evaluation process results in net gains for all stakeholders is a matter that deserves attention.

One Study or Several. In the first two stakeholder-oriented evaluations a single contract was let. Placing the responsibility for an entire evaluation on a single team of evaluators lays a heavy burden on its members, particularly when they have to cope with all the extra demands that the stakeholder approach entails. It makes them the arbiter of the only game in town. It gives them the responsibility for adjudicating among rival interests (including their own interests) and for deciding the direction that the study shall follow.

There is nothing intrinsic to the stakeholder approach that requires the funding of a single study to accommodate the interests of all parties to the program. The single blockbuster study appears to be an unthinking carryover from previous evaluation practice. For some time, it was assumed that one large study was better than several smaller studies, because the large study would have larger sample sizes, use more consistent measures, and therefore produce more precise estimates of effect. The stakeholder approach was tacked on to existing contracting practice.

As recent critics have noted, the blockbuster study suffers from severe limitations. It provides only one set of readings on one set of indicators, and the results depend on the particular operationalization employed. Cronbach and his associates (1980) have advocated "a fleet of studies" using different methods and different measures, done by different teams of investigators. If separate studies converge on results, the pattern of evidence is much more convincing than the results of a single study.

For the stakeholder approach, does it make more sense to fund several small studies? Each study could examine the program from the perspective of one set of stakeholders. The separate studies would be able to use the criteria of the separate groups and follow the issues that mattered to them. From the series of separate studies, a multidimensional view of reality would be more likely to emerge. The various pieces of evidence would illuminate the varied viewpoints.

It remains to be seen whether multiple studies would enrich understanding of the program, or whether they would create more conflict as each group pressed the evidence that supported its own case. It seems possible that multiple studies could do both. But they might enable interested groups of stakeholders to focus on issues that they defined as important without overloading traffic in a single study.

A sequence of studies could also explore diverse facets of programming. As new issues arose, new studies could pursue them. Since no one can foresee all contingencies in advance, sequential evaluations would be more likely to keep pace with shifting conditions. They could follow the variety of issues that a program encounters over the course of its life. Of course, there might be problems in maintaining continuity. A government funding agency like NIE would have to maintain its commitment to the exploration of a program's implementation and outcomes over a period of time. If early results proved disappointing, would the agency be under pressure to divert evaluation resources elsewhere, or could it continue to support study of the program, its problems, and its achievements?

Qualitative or Quantitative. What kinds of evaluation designs are compatible with the stakeholder approach? Does it fit best with qualitative, illuminative, ethnographic, process-oriented evaluation? The two case studies included in this volume seem to suggest so. Is that an idiosyncrasy of the particular programs, or is it inherent in the stakeholder idea? Can the approach ever be linked successfully with quantitative before-and-after evaluation? Could it work if the program under study had stabilized and settled down?

Are there ways that a stakeholder-oriented evaluation can serve both formative and summative purposes? Past experience suggests that studies that attempt the dual task tend either to scant one function or the other, or else they are swamped by floods of data, much of which usually goes unanalyzed. Can modifications in design overcome these problems, or should formative-qualitative and summative-quantitative studies routinely be separate undertakings?

As an attempt to cope with recognized shortcomings in evaluation practice, stakeholder-oriented evaluation retains modest promise. It has been tested with two particularly difficult programs, where its achievements were limited. Clearly, it cannot right all past wrongs or attain the nirvana that its advocates hoped for. At this point, I think it deserves further testing. As expe-

rience accumulates and if we conscientiously learn from that experience, we should be able to specify the conditions under which the stakeholder approach is likely to prove useful and to probe the realistic limits of its potential.

References

Ackoff, R. L. "Management Misinformation Systems." *Management Science,* 1967, B147-B156.

American Accounting Association Committee on Management Information Systems (AAACMIS). "Current Accounting Issues in the Area of Management Information Systems." *The Accounting Review,* supplement 1974.

Berg, M. R., and others. *Factors Affecting Utilization of Technology Assessment Studies in Policy Making.* Ann Arbor, Mich.: Center for Research on Utilization of Scientific Knowledge, 1978.

Caplan, N. "A Minimal Set of Conditions Necessary for the Utilization of Social Science Knowledge in Policy Formulation at the National Level." In C. Weiss (Ed.), *Using Social Research in Public Policy Making.* Lexington, Mass.: Lexington Books, 1977.

Conway, R., and others. "Promoting Knowledge Utilization Through Clinically Oriented Research: The Benchmark Program." *Policy Studies Journal,* 1976, *4* (3), 264-269.

Cronbach, L. J., and Associates. *Toward Reform of Program Evaluation.* San Francisco: Jossey-Bass, 1980.

Eidell, T. L., and Kitchell, J. M. (Eds.). *Knowledge Production and Utilization in Educational Administration.* Eugene, Ore.: Center for the Advanced Study of Educational Administration, University of Oregon, 1968.

Flanagan, J. C. "Case Studies on the Utilization of Behavioral Science Research." In *Case Studies in Bringing Behavioral Science into Use. Studies in the Utilization of Behavioral Science.* Vol. 1. Stanford, Calif.: Institute for Communication Research, 1961.

Grudnitski, G. "A Methodology for a Listening Information Relevant to Decision Makers." In C. A. Ross (Ed.), *Proceedings of the Second International Conference on Information Systems.* Cambridge, Mass.: Second International Conference on Information Systems, 1981.

Havelock, R. G. *Planning for Innovation Through Dissemination and Utilization of Knowledge.* Ann Arbor, Mich.: Center for Research on Utilization of Scientific Knowledge, 1969.

Joly, J. M. "Research and Innovation: Two Solitudes?" *Canadian Education and Research Digest,* 1967, *2,* 184-194.

Lazarsfeld, P. F., Sewell, W. H., and Wilensky, H. L. (Eds.). *The Uses of Sociology.* New York: Basic Books, 1967.

Lindbloom, C. E. *The Policy-Making Process.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

March, J. G. "Theories of Choice and Making Decisions." *Society,* 1982, *20* (1), 29-39.

Pelz, D. C. "Some Expanded Perspectives on Use of Social Science in Public Policy." In J. M. Yinger and S. J. Cutler (Eds.), *Major Social Issues.* New York: Free Press, 1978.

Rich, R. F. "Uses of Social Science Information by Federal Bureaucrats: Knowledge for Action Versus Knowledge for Understanding." In C. Weiss (Ed.), *Using Social Research in Public Policy Making.* Lexington, Mass.: Lexington Books, 1977.

Rothman, J. *Social R & D: Research and Development in the Human Services.* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Slovic, P., Fischhoff, B., and Lichtenstein, S. "Behavioral Decision Theory." *Annual Review of Psychology,* 1977, *28,* 1-39.

Slovic, P., and Lichtenstein, S. "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment." *Organizational Behavior and Human Performance,* 1971, *6,* 649-744.

Weiss, C. H. "Where Politics and Evaluation Meet." *Evaluation,* 1973, *1* (3), 37-45.
Weiss, C. H. (Ed.). *Using Social Research in Public Policy Making.* Lexington, Mass.: Lexington Books, 1977.
Weiss, C. H., with Bucuvalas, M. J. *Social Science Research and Decision Making.* New York: Columbia University Press, 1980.

11

Multisite/Multimethod Studies
An Introduction

Karen Seashore Louis

There has been a tremendous growth over the past decade in the funding of applied policy and evaluation research. Much of this research has followed relatively traditional research models, heavily influenced by the disciplinary training and methodological preferences of the principal investigator. However, as a consequence of growing dissatisfaction with the ability of these models to meet some of the demands of applied research, there have also been many efforts to develop new methodological approaches to answer questions about social programs and their impacts. Some of these involve applying nonresearch models or perspectives to the task of evaluation, (Smith, 1981a and 1981b). Others have been attempts to develop new techniques of inquiry drawn from classical social science research methods but formulated and combined in new ways.

In this issue we explore some of the "innovations" in studies that have tried a mix of qualitative and quantitative methods. We begin with a framework for understanding the methodologies that have emerged concurrently in independent studies conducted to answer a range of very different questions.

TWO PARADIGMS FOR RESEARCH: TRADITIONAL HOSTILITIES AND SIGNS OF CHANGE

For the past thirty years it has been common to refer to two distinct "paradigms" governing the methods of social science inquiry. The first

---

Author's Note: *The preparation of this article has been supported by Abt Associates Inc., the Center for Survey Research, University of Massachusetts, Boston, and by the Far West Laboratory. The comments of my colleagues, Allen Smith, Catherine Baltzell, Bob Herriott, and Bob Dentler, are greatly appreciated.*

From Karen Seashore Louis, "Multisite/Multimethod Studies: An Introduction," *American Behavioral Scientist*, 1982, 26(1), 6-22.

stresses the need to apply research design and analysis principles derived from the hard sciences, and emphasizes the desirability of experimental or quasi-experimental design and statistical analysis of multiple cases. The second paradigm argues that social phenomena are essentially different from those observed by the hard sciences and that, in order to understand them adequately, we must understand how they occur and what they mean to members of the social structure. A holistic understanding of human social structures and behaviors requires a qualitative, observationally based study of individual cases rather than experimental manipulation and analysis of selected variables.

Scholars disagree about the degree to which attempts are being made to build bridges between the qualitative and the quantitative paradigms. Twenty years ago Zetterberg (1962) advanced the hypothesis that qualitative and quantitative sociology each had an appropriate and essential role in the advancement of the discipline: Qualitative field work can help illuminate the theoretical and measurement issues in new areas of study, while quantitative methods are appropriate for testing the hypotheses derived from rigorous field methods. As recently as 1977, on the other hand, another observer of these two camps commented that the gulf between them was so great that it was unrealistic to expect any "grand synthesis" in the foreseeable future, since any steps toward synthesis were on the fringes of paradigms (Rist, 1977; see also McDonag and Schwirian, 1981). However, there are a number of other indications of a perceived need for something more than a simple détente between camps. Some experimental methodologists, for example, have recently taken tentative steps toward acknowledging the existence of an alternative paradigm and its suitability for studying phenomena that have typically been approached through quantitative methods (Campbell, 1974; Cook and Cook, 1977). Similarly, advocates of qualitative methods call for greater attention to standardization of analysis procedures (Sieber, n.d.; Yin, 1981).

In this issue we are concerned largely with design decisions that involve the use of multiple methods within the same study: The combination of ethnographies and surveys, traditional survey methods coupled with repeated measure analysis of a small number of sites, or the combination of intensive field work in a small number of sites with less intensive qualitative data collection in a larger number. Of equal concern is the way in which current studies choose to transform data from one type to another—for example, coding qualitative data on many sites to allow statistical analysis.

Before we turn to alternative methods, a discussion of terms—quantitative, qualitative, structured, and unstructured—is in order.

SOME CLARIFICATION OF TERMS

QUALITATIVE AND QUANTITATIVE

Sieber's (1973) deliberate use of the specific terms "field" and "survey" methods rather than the more general terms qualitative and quantitative raises an extremely important issue. In our view, the terms qualitative and quantitative are so imprecise when applied to the range of social science methodologies that they often detract from researchers' ability to discuss what they are doing. For some social scientists qualitative refers to the kind of data collected (e.g., unstructured interviews or observations); for others, qualitative refers to an analysis strategy. In fact, most definitions of qualitative and quantitative, particularly from a paradigmatic approach, assume a congruence between data collection and analysis techniques. Although often poorly defined, such terms clearly are part of the language of social research. Some diversity of usage occurs in this issue, but in general each author has attempted to confine the use of the term qualitative to data collection methods that involve nonnumeric data and to analysis that does not use statistical methods.

STRUCTURED AND UNSTRUCTURED

Some of the existing confusions about methodology can be avoided through frequent substitution of the concept of structure. Methods may be more or less structured, where "structured" is used to mean formally and systematically organized, while "unstructured" refers to the lack of systematic organization, also "loose, open, free" (Webster's New World Dictionary, 1976). The notion of more or less captures the complexity of methodological practice: There is no simple dichotomy between methods that are qualitative and those that are quantitative, but rather, there is a continuum on which a variety of different methods may be located, most of which range between the far ends of the scale.

TRANSFORMATION

To complicate matters further, large-scale studies in particular may be more or less structured at different points in their conduct, depending

upon the emphasis on data transformation as a component of the methodology. In reviewing the cases covered in this volume (and many others) we find that, in an increasing number of cases, data collected through very unstructured techniques may be transformed through coding into quantitative data bases, which are then analyzed using descriptive or inferential statistical techniques. Conversely, data may be collected through open-ended survey methods and analyzed "holistically" by site. In fact, to understand the variety of methods currently being employed, we must examine the nature of the design and practice at three points in the study: during data collection, during data-base formulation, and during the actual data analysis.

Perhaps the simplest way of presenting the range of options that may be available to the investigator at the current time is graphically, as suggested in Figure 1. Here we see the three dimensions arrayed next to each other and along each of the dimensions are some of the most commonly used techniques. For example, the data collection technique may vary from the most unstructured ethnographic approach to the most structured "census" approach. Similarly, the resulting data base may vary from unstructured, "naturalistic" field notes to one composed exclusively of ratio-scale data (as many economic data bases are). Finally, the data analysis techniques may vary from unstructured "journalistic" techniques, which may be called informal pattern recognition, to the most sophisticated inferential statistics and causal modeling.

A key issue in data transformation is what kind of data is transformed into what other kind of data, at what point in the research process, and by whom. The more traditional explanations of the research process treat transformations as events that occur (if at all) at a specified point in time (e.g., after the data have been collected, but before the analysis has begun). But, as we begin to examine many recent research projects, we see the elegant simplicity of the classical design overwhelmed by the complex set of potential transformations. These could included prespecified transformations during the data collection period (on-site coding of data or structured reporting by field staff who are not part of the analysis team) or transformations even after the data has been completely prepared and analyzed (visual patterning and coding of data for more succinct presentations of highly complex findings across several sites).

The transformation process lies at the heart of many of the case studies presented in this issue, and is also probably one of their most controversial features. It frequently represents an attempt to gather some of the best features that are attributed to the two paradigms, and to integrate them in a report that mixes more and less structured data and analyses. As we shall see, this often entails certain costs, so why is the current thrust of much research toward the use of novel or mixed methods?

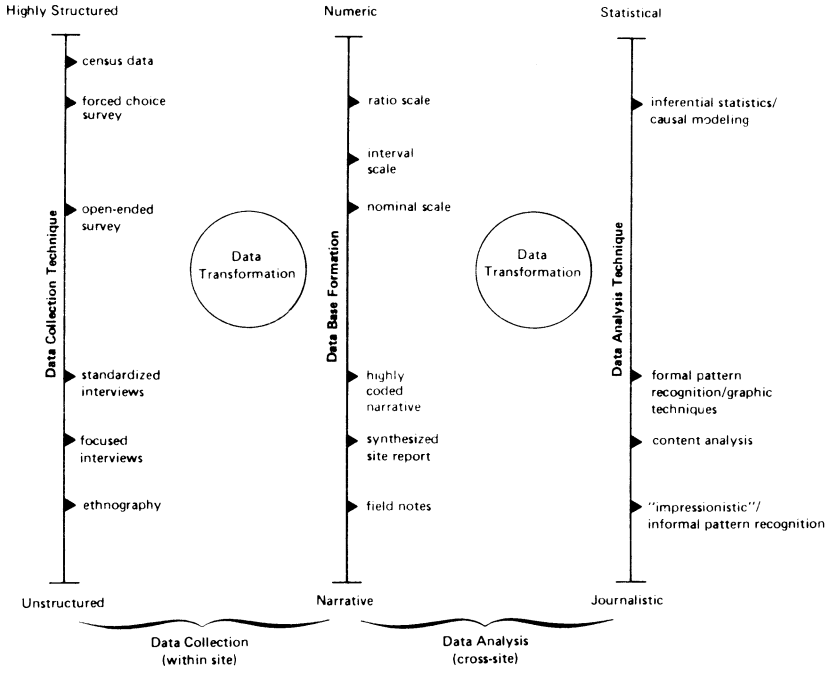


Figure 1: Dimensions of Variation in Multisite/Multimethod Studies

MULTIPLE METHODS AND POLICY RESEARCH

The arena of policy and evaluation research has not, until very recently, been equally hospitable to more and less structured research methods. It is safe to say that in the late 1960s and early 1970s experimental and quasi-experimental designs were viewed as the only logical options (see Weiss, 1972: 74). The relative value that influential policy and evaluation researchers placed upon less structured methods is most significantly revealed in a 1976 survey of 48 well-known experts, who were asked to estimate the importance of certain kinds of methodological training for people entering the field of evaluation and policy research (Anderson and Ball, 1978: 169-177). Of 32 "content areas," the five most highly rated were descriptive statistics, inferential statistics, statistical analysis, quasi-experimental design, and experimental design. At the bottom of the list were content analysis, case-study methodology, and job analysis.

Despite this apparently inhospitable environment, the movement toward integrating qualitative and quantitative methods has been fostered most evidently in social policy research, for several reasons. First, some of the initially high hopes for quantitative social policy research were deflated by an accumulation of null findings, and as "black box" research designs were unable to reveal why apparently massive experimental treatments should produce no measureable effects (Abt and Magidson, 1980). Researchers began to look at qualitative research methods as a means to improve their analysis—to point to interaction effects that should be explored, to allow them to account for otherwise inexplicable findings, and so forth.

A second reason for integrating methods in social policy research has been that the most rigorous and sophisticated of designs has not eliminated doubts about the durability of policy research findings. Rather than settling the controversy over the results of social policy and evaluation research, rigorous designs have simply raised a new dimension to the debate. Any researcher who does not like the results of a major policy study can almost always argue that a variety of methodological or analytic flaws undermine its validity. Not surprisingly, some policy makers have a deep-seated skepticism about supposedly "hard" findings—unless they are supported by qualitative data that make sense in the light of ordinary knowledge and experience (Corbett and Firestone, 1980; Sundquist, 1978).

Third, certain practical considerations stemming from federal policy have promoted the use of qualitative methods. Rather than designing programs as experiments, as advocated by many researchers (e.g., Gilbert et al., 1975), the federal government tends more often to permit wide latitude in program design and implementation at the local level, and is typically reluctant to randomly allocate individuals or organizations to "treatments." In addition, government agencies, while demanding definitive assessments of programs, have also typically imposed one-time ex-post facto designs for a variety of reasons, including the need for quick answers and the enormous costs of mounting longitudinal experimental studies. Also, evaluation or policy analysis is often an afterthought. The researcher called upon in midstream must be inventive with regard to study design (Weiss, 1972: 75).

A fourth factor is forms clearance for standardized data-collection instruments, which can take from four to six months. The federal agency that asks for both qualitative and quantitative data can begin to know something about the topic in question long before a survey or testing program could begin. Thus, particularly in cases where there is only

limited interest in a "bottom line" assessment, qualitative approaches may be perceived as more efficient.[1]

Finally, qualitative designs may also be viewed by government agencies as more flexible than traditional experimental designs in responding to changing policy contexts and questions. The current funding process begins with a government agency's Request for Proposals, which specifies the general tasks to be carried out. The winning proposal becomes the basis for a contract with a specified scope of work (the activities to be performed and the level of effort). Less structured studies often have less detailed scopes, which makes it easier to shift resources without a contractual modification. The appeal of such an arrangement is particularly understandable during periods of political or programmatic instability, when policy priorities may change before the results of a large, structured study are available.

Despite these pressures, one should not infer a change in paradigmatic preferences on the part of policymakers (and at least some researchers) to draw upon the best of both methods. Policymakers typically hope to retain the strengths of quantitative research (generalizability of results, reliability of observations, and the ability to synthesize a large complex study in a brief report), while capitalizing on the advantages of more qualitative methods (holistic description, flexibility, and sensitivity to causal relationship). There has also been an increasingly perceived need to deal with the integration of findings across different methodological approaches in a more formal way.

MULTISITE/MULTIMETHOD RESEARCH AS A SOLUTION

As an approach to solving some of the perceived limitations of earlier, large-scale policy and evaluation research, a number of recent studies have used a multisite/multimethod research design. One major characteristic of this approach is a focus on a unit of analysis that is programmatically based and larger than an individual. This represents a major deviation from the earlier traditions of evaluation and policy research, which tended to emphasize changes in individuals and individual behavior. The emphasis on sites as the unit of analysis reflects the fact that programs as implemented are rarely the same as the intended federal or state design (Pressman and Wildavsky, 1973). To account for a program's operations and impacts, it is therefore necessary to attend to the unit that serves as the administrative basis for the social program and activity.

The multisite/multimethod research approach is further characterized by a relatively large-scale collection of less standardized data, where the number of sites involved is at least greater than 5, and sometimes as many as 50 to 60. It is quite common to use some scattered collection of less standardized data to help focus a study that is predominantly dependent on standardized data and quasi-experimental design (e.g., Fienburg, 1977). What distinguishes the multisite/multimethod approach from either the more traditional single case study or impressionistic data collection at a small number of sites is the emphasis upon gathering less structured data from a sample of sites.

A third attribute of this method is that less structured data collection is the major component of the study, but not the sole component. Unlike the approach used by most serious proponents of highly structured experimental approaches to policy research, the multisite and multimethod approaches that are described in this volume view less structured data as coequal with the more structured data (and sometimes more important or valuable for achieving the overall study objectives).

The final attribute in the emphasis is upon cross-site analysis, rather than solely upon the preparation of independent case reports. Individual case reports constitute a technique frequently used to provide illustrative data, vignettes, or testimonial data to policymakers and legislators who frequently need "real examples." The development of individual cases, no matter how carefully constructed, typically follows the traditional conventions of "holistic" causal analysis for each individual site. Often there is little or no effect devoted to investigating patterns across sites. In the research innovations discussed in this volume, however, preparation of individual case studies was often secondary to the development of a narrative data base for cross-case analysis. In some instances, no individual case material was presented at all.

COLLECTING AND ANALYZING LESS STRUCTURED DATA IN MULTISITE/MULTIMETHOD STUDIES: ISSUES

Designing and carrying out a study with these characteristics provides the researcher with a variety of challenges in regard to major design options, issues of less standardized data-collection activities, and the problems of pulling more and less standardized data together into a

coherent analysis. More specifically, the researcher needs to grapple with these questions:

(1) When are multisite qualitative methods desirable?
(2) What are some of the major design issues to be faced?
(3) What are some of the major analysis issues to be faced?

WHEN TO CHOOSE THE MULTISITE/MULTIMETHOD APPROACH

In the previous pages we implicitly rejected the methodological imperative inherent in the narrow "two paradigms," but we are still left with questions about when the multisite/multimethod approach is appropriate. While there are always researchers who believe that "more data is better," there are inherent tradeoffs that must be carefully considered.

Among the first of these is expense. A study employing a minimum of two methods (one more structured and on less so) will inevitably be more costly than one using a single method. In many cases, the extra cost may not be worth the investment: The program or policy under study may not be sufficiently important or powerful, for example, to justify a public investment, particularly in times of declining resources. A quick survey of the current effectiveness of competing programs in achieving a desired goal may be enough. Using a simple post-facto design with control groups, while not scientifically pleasing, may give policymakers what they need.

The scientific management cost in a multisite/multimethod study must also be considered. In all cases with which we are familiar, these costs are far greater than in studies employing a single (or predominant) methodological approach. There are many problems facing the research team: coordination of data-collection approaches, collaboration by teams of researchers with different methodological preferences and perspectives, and the much-dreaded possibility that the two or more different types of data will produce contradictory results (see Trend, 1978). These problems are unavoidable, time-consuming, and often painful.

There are also costs of innovation, felt largely in the receptivity of audiences to studies that appear to have too many facets, too many different types of data, and too many components. Unconventional designs may distract serious readers from the messages in the findings and may be perceived with hostility by those in both paradigmatic camps.

From a disciplinary, theory-building perspective, the multisite/multimethod approach may be the best in instances where the research setting is ill-matched with more conventional experimental or quasi-experimental designs, yet where a study based on qualitative data alone may not fit the needs of the funding agency.

MAJOR DESIGN CHOICES

Among the controversial design choices facing multisite/multimethod studies are the selection of field personnel, methods to control data quality, and the degree of standardization to be imposed on the data collection. The first design question facing investigators is who will collect the data. Because there are multiple sites, often scattered over the entire continent, it is virtually impossible for the principal investigator(s) to conduct all of the interviewing and field work personally. Thus, the process of personal immersion, and integration and analysis through participation, is not possible. There are several ways to settle the issue: Trained social scientists located near the sites may be hired to work part- or full-time during the data collection phase; staff may be sent out from the research institute or university where the work is being performed; or people who actually live in the site, and are members of the community may be hired. The tradeoffs that must be considered in making this decision are many, including the costs of travel and boarding, the importance of a "shared perspective" on the issues, and the degree to which the data-collection activities are prespecified and well-defined.

Another issue of vital importance to the design is the development of a system for monitoring data quality across sites. Quality control is much more difficult in multisite/multimethod studies because of the lack of standardization in the data. Managerial solutions to the problem involve monitoring the data-collection process regularly, keeping in touch with field workers, reading over their notes, and possibly even conducting some independent investigations.

Data triangulation, which may be both intra- and intersite, is another solution to data quality control. Within a site field workers may be instructed to bolster their assertions with information from several sources (see Denzin, 1978). Intersite triangulation occurs when the principal investigator observes that a site seems to be behaving differently and requests additional information to either confirm or deny the patterns that are apparent. Intrasite investigator triangulation most often occurs when the project uses a pair of field workers, which in some instances may include an indigene and a social scientist and

in others, two specialists in different techniques. (A recent study of social scientist/field worker as part of the team.) Intersite investigator triangulation is more risky, but in some instances there may be deliberate decision to select field workers with different theoretical perspectives and to provide them with relatively little orientation: If similar conclusions are or can be drawn across sites, the robustness of the conclusion is thus validated. The quality of less standardized data at the site may also be controlled by intrasite triangulation with quantitative data.[2] If there are discrepancies, the search for a more complete explanation is begun.

Finally, by far the most common and inexpensive form of quality control is post hoc methodological triangulation between more and less standardized data sources—in other words, determining whether the separate streams of data, separately analyzed, confirm one another. This approach also has obvious risks unless the project is funded with sufficient resources for ironing out any inconsistencies that may be found.

Another major design dilemma for the investigator—perhaps the most controversial—is how to strike a balance between gathering comparable data across sites (to help achieve the unified, cross-site analysis that is desired) and preserving the advantage of flexibility in the less standardized component of the study. Those who are methodologically committed to the qualitative paradigm feel that the value of of the multimethod approach is undermined by the necessary structuring of less standardized data to ensure that they can be analyzed within a finite period. This brings up another design chocie: The tradeoff between having many sites for which some less structured data are available, and having fewer sites studied in greater depth using techniques more akin to the traditional ethnographic approach. The tradeoff between depth and breadth in the less structured data is a clear one and strikes at the heart of the paradigmatic controversy: Is it preferable to preserve the distinctness of the two approaches, or do the methodological innovations discussed in some of the latter chapters represent breakthroughs in the development of social scientific methodologies uniquely suited to the systematic study of "squishy" social phenomena?

MAJOR ANALYSIS ISSUES

The dilemmas facing the investigator during the analysis phase of the project are, if anything, more difficult and novel than those attending the design- and data-collection phases. This is largely a consequence of

the lack of well-articulated analytic strategies for qualitative data. Miles, for example, has indicated that, despite some progress in articulating guidelines for analysis, "The analysis process is more memorable for its moments of sheer despair in the face of the mass of data, alternating with moments of achieved clarity, soon followed by second-guessing skepticism" (1979: 597). Miles goes on to assert that the guidance available for cross-site analysis is even more slim. Yin (1981) has developed a cogent argument for the systematization of cross-case analysis of qualitative data; unfortunately, the particulars available to guide the novice are also lacking in this brief article.

The case studies presented here represent much more tentative steps toward finding and defining a range of alternative analysis strategies. They also emphasize a need to develop new techniques for systematically matching and comparing patterns that are located in individual sites. Some of the studies used visual techniques to summarize data, some have quantified and counted relationships, and others have developed staffing and process procedures to stimulate pattern recognition and to verify patterns. Another major issue is that of transforming raw qualitative data to a form more suitable for pattern recognition. While qualitative analysts have not traditionally been concerned with the development of a formal "data base" aside from field notes, many current studies are struggling with the notion that there is a need for alternatives to traditional coding techniques for translating raw notes into forms more amenable to intersite analysis. The need to transform the data raises other questions, such as who shoult be involved in this activity (the field worker, the ultimate analyst, or others who can assess reliability and validity), and when it should be done (during data collection, during preliminary analysis, or after data collection has been completed).

The press toward more systematic data bases, and toward locating and counting patterns that emerge, raises the question of how to judge the strength of findings within sites and across sites. Projects are typically required to develop rules of thumb, either explicit or implicit, as to whether or not a finding is a real or generalizable one. The generalizability problem is a haunting one for investigators who feel that, because they have examined sizable samples, they should be making generalized assertions, even though they are concerned and uncertain as to how the idiosyncracies of less structured methods may be affecting their results. This concern is compounded in studies where sites are selected purposively as exemplars of particular treatments, conditions, or local characteristics, rather than as representatives of any known universe.

INTEGRATING DIFFERENT TYPES OF DATA

A question that affects all phases of the research is the relationship between the more and less structured components of the data. Four major approaches to integration can be identified: the sequential, the parallel, the fused, and the interactive. Each of these models deals with the need to maintain some of the good characteristics of one to enhance the usefulness of the other. All involve multisite/multimethod approaches to data collection and analysis, but each represents a very distinct approach to the marriage of the two paradigms.

The Sequential Model

The most commonly used sequential model is based on the explanations typically offered by quantitative methodologists of the relationship between the two paradigms. Preliminary "knowing" is seen as crucial where the topic in question is poorly understood, where measurement techniques are not perfected, and where there is a need to identify or refine hypotheses. Within such a study qualitative data collection precedes the development of survey or testing instruments, which are perceived as yielding the "final" data for the study (see Zetterberg, 1962). A contemporary application of this approach is the "evaluability" movement, which stresses the need for a two-stage evaluation of major social programs. The first stage involves significant field-based data collection to determine whether the parameters of the treatment can be identified, and to develop a model of program operations and outcomes that will form the basis for an appropriate quantitative design in the second stage (see Rutman, 1980).

A newer sequential model, in which case material is collected after a survey or other standardized data collection, has been gaining in popularity. This approach, often drawing upon Lazarsfeld's notion of "deviant case analysis," views the purpose of the less structured data as that of illuminating the questions that may arise in preliminary analysis of the quantified data base. Or it may be used to understand the dynamics of specific types of cases or settings of particular interest to the funding agency or principal investigator. From the perspective of policy research the reversed sequential model has a great deal to offer. The research model may start with a quasi-experimental study designed to indicate whether the program is working as intended. If and when mixed results are produced in all or some of the sites, case studies can help

to illustrate some of the policy options, as well as explain the reasons for the variation (see also Lazarsfeld, 1976).

The Parallel Model

The parallel model makes no assumption about an appropriate linear relationship between qualitative and quantitative methods. Typically, such designs accept the arguments of Rist (1977) or Scriven (1972) that they represent two very different ways of knowing, and that they help the research see and illuminate different aspects of the social phenomenon under study. This new tradition for multisite/multimethod case studies has typically assumed that the most appropriate approach to maximizing the contributions of both is to allow them to develop independently but simultaneously.

Parallelism reduces problems of coordination, both of data collection and between centrally based staff and field-based staff. In addition, it retains the maximum design and analysis flexibility. Consequently, it is often preferred by paradigmatic researchers, and it is frequently employed in large studies.[3]

The Fused Model

Because of some of the limitations of the sequential and parallel approaches, an entirely different approach has been gaining emphasis in a number of recent research projects.[4] The new method fuses some of the most "valuable" features of quantitative data collection—emphasis upon standardization of data points, an emphasis upon determining causality and testing hypotheses (rather than describing), and an emphasis upon cross-case analysis—with a flexible approach to observation and an emphasis upon holistic analysis. This approach is most frequently referred to as the standardized case method (see Baltzell, 1980).

Many of the standardized case method studies have gone one step further to include coding of the completed case studies by the field staff using "survey" instruments similar to a respondent interview, and then analyzing them quantitatively.[5]

The Interactive Model

The interactive model—a new approach to integrating qualitative and quantitative data—builds upon some of the features of each of the

previous models. The major distinctive characteristics of the approach are:

(1) the merging of qualitative and quantitative data within as well as across sites;
(2) staffing patterns that involve senior researchers in both quantitative and qualitative aspects of the study;
(3) persistent attempts to triangulate data sources and interpretations;
(4) cyclical interaction between the qualitative and quantitative method during all phases of the study, including sampling, instrumentation, data collection, analysis, and report.

The interactive model is so named because it attempts to respond directly to Lazarsfeld's admonition that "the most important lesson to learn is that . . . the quantitative and the qualitative operations should be kept in continuous interchange" (1976: 57).

DESIGN OF THE ISSUE

The remainder of this issue has been designed by the editors and authors to address in greater detail some of the questions raised in the previous pages. The topics have been selected because they are among the most pressing that arise when researchers engaged in this type of project try to explain what they are doing and why.

Each of the five case articles (Herriott, Yin, Smith and Robbins, Huberman and Crandall, and Louis) focuses on three or four of the issues discussed above. The last two articles (Miles and Datta) take a somewhat different approach. Rather than describe a single case, the authors draw upon their extensive experience with policy research of different types to comment upon the collection of previous chapters as a group. Miles addresses the question of whether the approaches suggested here have a substantive payoff for other researchers in other contexts, and provides a critical review from a colleague struggling with similar problems, while Datta's article focuses on the politics of policy research and its implications for multisite/multimethod endeavors.

NOTES

1. That field-based methods should come to be thought of as efficient is an ironic turnabout from earlier periods in which public opinion surveys and other survey data collection activities were touted because of their speed and low cost.

2. This form of triangulation assumes, of course, that the survey or other standardized data are available when the field work is being completed.

3. The Rand Corporation Study of Federal Programs Supporting Educational Change and the SRI study of Teacher Corps provide some recent examples, in addition to the Abt Associates Rural Experimental Schools Study.

4. The fused model approach is perhaps best articulated by Yin (1980 and 1981a), McClintock et al. (1979), and Baltzell (1980).

5. Recent studies emphasizing this approach include Yin's studies of innovations in urban bureaucracies (1978) and of interorganizational networks among state, local, and regional educational agencies (1981); King's (1980) study of staff development in desegregation; and a study of magnet schools as mechanisms for desegregation (Royster and Baltzell, 1979).

REFERENCES

ABT, W. P. and J. MAGIDSON (1980) Reforming Schools: Problems in Program Implementation and Evaluation. Beverly Hills, CA: Sage.

ANDERSON, S. and S. BALL (1978) The Profession and Practice of Program Evaluation. San Francisco: Jossey-Bass.

BALTZELL, C. (1980) The standardized case study: A hyprid approach to the qualitative/quantitative issue. Presented at the meeting of the American Educational Research Association.

CAMPBELL, D. (1979) "Qualitative knowing in action research," in T. Cook and C. Reichardt (eds.) Qualitative and Quantitative Methods in Social Research. Beverly Hills, CA: Sage.

COOK, T. and D. COOK (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

COOK, T. and C. REICHARDT [eds.] (1979) Qualitative and Quantitative Methods in Evaluation Research. Beverly Hills, CA: Sage.

CORBETT, H. and W. FIRESTONE (1980) Constructing Case Studies on Educational Change for Policy Makers. Philadephia: Research for Better Schools.

CRAIN, R. (1977) "Racial tension in high schools: pushing the survey method closer to reality." Anthropology and Education Q. 8 (May): 142-151.

DENZIN, R. (1978) The Research Act. New York: McGraw-Hill.

FIENBERG, S. (1977) "The collection and analysis of ethnographic data in educational research." Anthropology and Education Q. 8 (May): 50-57.

GILBERT, J., R. LIGHT, and F. MOSTELLER (1975) "Assessing social innovations: an empirical base for policy," in A. Bennet and A. Lumsdaine, Evaluation and Experiment: Some Critical Issues in Assessing Social Programs. New York: Academic Press.

KING, N. (1980) Staff Development Programs in Desegregated Seetings, Santa Monica, CA: RAND.

KNAPP, M. (1979) "Ethnographic contributions to evaluation research. The Experimental Schools Program and some alternatives," pp. 118-139 in T. Cook and C. Reichardt (eds.) Qualitative and Quantitative Methods in Evaluation Research. Beverly Hills, CA: Sage.

McDONAG, E. and K. SCHWIRIAN (1981) "Changes needed in graduate sociology training for the '80s." Footnotes (October): 9.

MILES, M. (1979) "Qualitative data as an attractive nuisance: the problem of analysis." Admin. Sci. Q. 24 (December): 590-601.

PRESSMAN, J. and A. WILDAVSKY (1973) Implementation. Berkeley: Univ. of California Press.

RIST, R. (1977) "On the relations among educational research paradigms: from disdain to detente." Antropology and Education Q. 8 (May): 42-49.

ROYSTER, E. and C. BALTZELL (1979) An Evaluation of the ESAA Magnet Schools Program: Final Report. Cambridge, MA: Abt Associates.

RUTMAN, L. (1979) Designing Useful Evaluations. Beverly Hills, CA: Sage.

SCRIVEN, M. (1972) "Objectivity/subjectivity in educational research," in H. B. Dunkie et al. (eds.) Philosophical Redirection of Educational Research. Chicago: National Society for the Study of Education.

SIEBER, S. D. (n.d.) A review of analysis procedures in textbooks of qualitative methods. (unpublished)

SIEBER, S. D. (1973) "The integration of field methods and survey research." Amer. J. of Sociology 78 (Spring): 1335-1359.

Strauch, R. (1976) "A critical look at quantitative methodology." Policy Analysis 2 (Winter): 121-143.

SUNDQUIST, J. (1978) "Research brokerages: the weak link," in L. E. Lynn (ed.), Knowledge and Policy: The Uncertain Connection. Washington, DC: National Academy of Sciences.

TREND, M. (1978) "On the reconciliation of quantitative and qualitative analyses." Human Organization 37: 345-354.

WEISS, C. (1972) Evaluation Research: Methods of Assessing Program Effectiveness. New York: Prentice-Hall.

YIN, R. (1981a) "The case study as a serious research strategy." Knowledge 3 (September): 97-114.

YIN, R. (1981b) "The case study crisis: some answers." Admin. Sci. Q. 26 (March): 58-66.

YIN, R. (1978) Tinkering with the System: Technological Innovations in State and Local Services. Lexington, MA: D. C. Heath.

ZETTERBERG, H. (1965) On Theory and Vertification in Sociology. Tottowa, NJ: Bedminster.

12

Meta-Analysis

Techniques, Applications, and Functions

Michael J. Strube and Donald P. Hartmann

The literature review has always served the important function of "taking stock" of what is known so that future research can be directed more efficiently (Cooper, 1979), policy decisions can be made more effectively (Light, 1979; Pillemer & Light, 1980a, 1980b), and scientific information can be disseminated to wider audiences (Cooper & Rosenthal, 1980). Historically, the sophisticated and precise methods used in the single studies that comprise a given literature have not been duplicated when those same studies were reviewed and integrated. In fact, traditional methods of literature review have been criticized for subjectivity, imprecision, and neglect of important information contained in primary studies (cf. Glass, 1976). The impact of these weaknesses is amplified in view of the complexity resulting from the staggering numbers of studies being conducted (Glass, 1977). Schmidt (Note 1) has gone so far as to state that the production of useful cumulative knowledge is the most pressing research need of the 1980s (see also Tedeschi, Gaes, Riordan, & Quigley-Fernandez, 1981).

In response to the inadequacy of the typically qualitative and largely narrative literature review (e.g., see Jackson, 1980), psychologists have begun a rapid transition to the use and development of more quantitative procedures (e.g., Glass, 1976, 1977; Glass, McGaw, & Smith, 1981; Rosenthal, 1978, 1979a, 1980). These procedures, known collectively as meta-analysis (Glass, 1976), enjoy increasingly widespread use (e.g., Blanchard, Andrasik, Ahles, Teders, & O'Keefe, 1980; Hyde, 1981; Smith, Glass, & Miller, 1980; Strube, 1981; Strube & Garcia, 1981; Underwood & Moore, 1982; see also Glass et al., 1981). However, meta-analysis has not gone unchallenged (e.g., Cook & Leviton, 1980; Eysenck, 1978; Gallo, 1978; Leviton & Cook, 1981; Rachman & Wilson, 1980; Sohn, 1980; Strube & Hartmann, 1982). The majority of these commentaries express healthy skepticism rather than dogmatic resistance and have served the useful purpose of increasing critical self-examination among meta-analysts. Our purpose in this article is to "take stock" of what we know about meta-analysis. By doing so we hope to point out the proven strengths of the procedures, promising developments, and most important, areas in need of attention.

Meta-analysis is not simply a collection of quantitative techniques. Rather, it represents a systematic approach to the problem of integrating a common research domain (cf. Cook & Leviton, 1980; Cooper, 1979; Cooper & Arkin, 1981; Glass et al., 1981; Leviton

Thanks are extended to Joe Garcia, Bill Gardner, and Paul Vinciguerra for their comments on an earlier draft of this manuscript.

From Michael J. Strube and Donald P. Hartmann, "Meta-Analysis: Techniques, Applications, and Functions," *Journal of Consulting and Clinical Psychology*, 1983, 51(1), 14–27.

& Cook, 1981). Furthermore, as we (Strube & Hartmann, 1982) and others (e.g., Cook & Leviton, 1980; Leviton & Cook, 1981) have pointed out, the meta-analytic approach makes obvious the large number of qualitative and often arbitrary decisions that one must make during a research integration. Finally, the outcome of a meta-analysis need not be restricted to bland summary statements of research results but may be more general and creative. These three levels of examining meta-analysis—the quantitative techniques, the application of those techniques, and the more general functions of meta-analysis—are the framework for the presentation that follows.

Meta-Analytic Techniques

A wide variety of statistical techniques are available to the meta-analyst, but most fall under two basic approaches: combination of significance levels and combination of effect sizes.[1] Although significance level and effect size are typically correlated (e.g., Rosenthal & Rubin, 1979, 1980), the two provide distinct information. From a theoretical standpoint, it is important to know whether a particular result occurred due to chance. Statistical significance provides such information, and the combination of probabilities across studies allows the reviewer to determine whether a set of results could have arisen by chance. A detailed description of all available techniques for combining probabilities is beyond the scope of this article. Rosenthal (1978, 1980) provides useful summaries of the most common methods for combining probabilities, as well as guidelines for their application. Current practice (e.g., Cooper, 1979; Strube & Garcia, 1981) favors the use of the Stouffer technique (Mosteller & Bush, 1954; Rosenthal, 1978).[2] However, reviewers are well advised to use at least two methods of combining probabilities (cf. Cooper, 1979; Rosenthal, 1978), particularly with small samples. The sampling distributions of these statistics are not well understood, and the results may not always agree (see Strube, 1981).

The combination of effect sizes represents the second major approach to the summarization of results across studies (Glass, 1976, 1977; Glass et al., 1981). Because a significant result is not necessarily meaningful, it is important to examine the *magnitude* of effects across studies. This step can be particularly important in applied areas (e.g., psychotherapy outcome) in which the effectiveness of a treatment or intervention has high priority. To achieve this end, the meta-analytic reviewer has available a rather large number of methods for estimating effect size. These estimates include the proportion of variance accounted for in the dependent measure (e.g., Sohn, 1980), the percentage of overlap between treatment and control distributions (Cohen, 1977; Cooper, 1979), and the standardized difference between treatment- and control-group means (Cohen, 1977). This latter estimate, known as Cohen's *d*, is the most widely used when two groups are being compared (e.g., Cooper, 1979; Smith et al., 1980).[3]

Beyond the computation of a combined probability and of an average effect size, several additional analytic procedures are available. An important aspect of a meta-analysis concerns the stability of the combined results. Stability ultimately affects the confidence placed in the inferences and conclusions drawn from the collection of studies

---

[1] Other approaches, such as vote counting (e.g., Hedges & Olkin, 1980), will not be discussed since they have seen little current use. The brand of meta-analysis developed by Schmidt and Hunter (e.g., Hunter & Schmidt, 1978; Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman, & Shane, 1979) will also not be discussed, since it appears to require more quantitative information than is typically found in most research domains.

[2] The Stouffer technique is defined by the following formula:

$$z_s = \sum_{i=1}^{N} z_i/(N)^{1/2}$$

where z_i is the standard normal deviate corresponding to the exact one-tailed p value for a given hypothesis test (sign determined by direction of support), and N is the number of hypothesis tests combined. The resulting z_s is referred back to a table of standard normal deviates to obtain the combined probability.

[3] It should be noted that whereas researchers agree that an estimate of effect size is desirable, there has been considerable debate as to which effect-size estimate is best, and how effect size should be interpreted. The reader is referred to Cooper (1981), Gallo (1978), Rimland (1979), Rosenthal and Rubin (in press), and Sohn (1980) for various perspectives on effect-size interpretation.

reviewed. Thus, it is important to demonstrate that the addition of a few more studies would not drastically alter the combined results. Rosenthal (1979a) has presented a procedure, based on Stouffer's technique, which allows the computation of the number of unretrieved studies or hypothesis tests averaging null results that would be required to bring the overall combined probability to a chosen level (e.g., just significant, $p = .05$). This number, dubbed the "Fail-safe N" (Cooper, 1979), allows the reviewer to demonstrate whether an overall combined probability could be rendered nonsignificant by a few "file-drawer" studies averaging no effect. For effect-size estimates, the computation of confidence intervals or standard deviations allows one to estimate their stability. It is also possible to estimate the impact on an average effect size of including N file-drawer studies with a given average d (e.g., Shapiro & Shapiro, 1982). This can be a useful procedure provided there is a reasonable estimate of the file-drawer d (e.g., the average d of unpublished studies that *have* been recovered).

In forming the combined results, the meta-analyst may choose to emphasize differentially some studies or hypothesis tests over others. For example, studies differ in their quality, and poor studies can be weighted less in the analysis rather than excluded. Differential weighting of studies is easily accomplished using a modification of the Stouffer technique for combined probabilities (e.g., Cooper, 1979; Mosteller & Bush, 1954; Rosenthal, 1978). Weighting of effect sizes is also possible, although it has not been conducted routinely. Weighting of studies is simple in principle and allows the reviewer a certain degree of flexibility in the combination of results. However, it is important that the weighting strategy chosen be sensible, defensible, and preferably selected prior to data collection.

These techniques might be said to provide the basic results of a meta-analysis. However, it is possible to extract additional quantitative information from a literature review. In particular, the size of effects and significance levels obtained in the studies reviewed are likely to vary considerably. Opposite or contradictory findings may emerge, and this variability may be systematically related to differences between studies. As Light has argued, examining the variability in a set of results may lead to considerable insight ino the "why" of the underlying process (Light, 1979; Light & Smith, 1971; Pillemer & Light, 1980a, 1980b). A number of procedures and strategies exist for identifying sources of variation in either p values or effect sizes. Simple correlation of study characteristics with effect sizes or significance levels can indicate whether a given phenomenon is general or depends on specific values or levels of the variables examined. For example, Smith and Glass (1977), in their meta-analysis of psychotherapy outcome, found that effect size was significantly correlated with IQ of clients, reactivity of the outcome measure, and similarity of therapists and clients. A related approach is to disaggregate studies or hypothesis tests by levels or categories of some important variable in order to actually examine the combined results at each level or within each category. Using this approach, Shapiro and Shapiro (Note 2) found that effect sizes were greatest for phobias and lowest for anxiety and depression in their analysis of predominantly analogue studies.

A more complex analysis is called for when the joint effect of more than one mediating variable is of interest. One approach is to disaggregate studies along two or more variables. If the data in the resulting cross-classification conform to assumptions, then standard analysis of variance (ANOVA) can be used to examine main and interactive effects (e.g., Smith et al., 1980; Shapiro & Shapiro, Note 2). However, most often the mediating variables will be intercorrelated, or primary interest will lie in the independent contribution of many variables. In such cases, multiple regression analysis (e.g., Cohen & Cohen, 1975; Kerlinger & Pedhazur, 1973) can be used. Regression analyses have the advantages of providing a measure of the proportion of variance accounted for in either effect sizes or significance levels (i.e., R^2), statistically controlling for the effects of other variables, and allowing the creation of variables that represent interactions. Thus, multiple regression easily handles data traditionally analyzed via ANOVA, with some additional advantages as well.

Pillemer and Light (1980) advocate an additional approach to examining variation in study outcomes. These researchers advocate the examination of the distributions of effect sizes and significance levels in order to detect unusual variation. What is particularly novel about their approach is their emphasis on the detection and examination of "outliners"—unusually effective and ineffective treatments or programs (cf. Klitgaard, 1979). Once identified, a search for explanatory factors (key similarities or differences) may provide considerable insight into understanding why the program or treatment worked so well or so poorly. When coupled with the in-depth examination of the usual or typical findings across studies (which Pillemer and Light also advocate), the investigation of the atypical can provide a rich yield of information.

Finally, a number of recent developments in the analysis of combined results also bear mention. For example, Rosenthal and Rubin (1979, Note 3) present techniques for comparing two or more significance levels, or effect sizes, from independent studies. These techniques not only allow for the detection of significant variability within a set of studies (i.e., heterogeneity of significance levels or effect sizes) but also provide for the testing of a priori hypotheses about that variability. These techniques can be applied to small numbers of studies or hypothesis tests where more traditional analyses are not appropriate. Thus they should provide useful adjuncts to the methods described above. In addition, Glass and his colleagues (Glass et al., 1981; Smith et al., 1980) have made a number of advancements in the analysis of effect-size data. For example, Smith et al. (1980; see also Glass et al., 1981) examined the effects of drugs and psychotherapy. Using least squares procedures they were able to estimate the separate contributions of drugs and psychotherapy as well as their combined effect, even though all studies did not include both drug treatment and psychotherapy groups contrasted with a control group. Glass et al. (1981) also present procedures for fitting curves to groups of studies with quantitative independent variables (e.g., class size in the analysis of achievement or number of months to posttherapy follow-up in the analysis of psychotherapy outcome). Finally, these researchers have used multidimensional scaling techniques in both the construction of independent variables (types or classes of psychotherapy, Smith & Glass, 1977; Smith et al., 1980) and the scaling of studies of sex roles along a psychological adjustment continuum (Glass et al., 1981). These developments enhance the wide variety of techniques that the meta-analyst has at his or her disposal.

We have purposely given relatively brief attention to statistical techniques that can be used in a meta-analysis. A number of sources cited above provide detailed descriptions. Furthermore, the techniques themselves do not provide any insurmountable barriers to the conduct of a meta-analysis. They are for the most part simple in nature, or at least understandable with a little work. However, the application of the techniques to the real and often problematic data of a research domain creates a host of problems that require special attention.

Application of the Techniques: Conducting the Review

The conduct of a meta-analytic review requires a series of carefully considered decisions. No amount of precision or "quantification" in the actual techniques can spare the reviewer the thought and planning that are required of *any* review (cf. Feldman, 1971; Jackson, 1980; Taveggia, 1974). However, the systematic nature of a meta-analysis does serve to make a number of those decisions and their implications more explicit. Following is a discussion of some of the more bothersome problems that challenge the reviewer (see also Cooper, in press). Central to this discussion is the assumption that the reviewer has clear, precise, and explicit hypotheses or questions that he or she wishes to address.

Sample Biases

As we and others (e.g., Shapiro & Shapiro, 1982; Smith et al., 1980) have noted, the results from individual studies that are aggregated in meta-analysis are analogous to the responses from individual subjects in traditional studies. However, like most analogies, the resemblance is imperfect. Unlike the tra-

ditional research design, the meta-analytic sample is likely to be nonrandom and biased, and the individual datum may lack independence. Many search procedures draw their samples from published sources, often on the tenuous assumption that published research is of higher quality (cf. Cooper, 1979; Glass et al., 1981) and because retrieving published sources is simply easier. The most serious effect of this selective sampling is to bias the estimates of both significance levels and effect sizes. Publication policy is clearly biased toward the reporting of significant findings (Bakan, 1966; Chase & Chase, 1976; Greenwald, 1975; Smart, 1964; Sterling, 1959). Thus, the magnitude of both significance levels and effect sizes may be overestimated. Indeed, data from simulation studies and past meta-analyses validate the existence of this bias. In two studies using Monte Carlo techniques Lane and Dunlap (1978) and Underwood and Dickson (Note 4) both found evidence for a bias in effect sizes as a function of selection criteria based on statistical significance. Glass et al. (1981) review 11 meta-analyses that included data from journals and other less traditional sources such as books, theses, and unpublished studies. Across all the analyses, the published studies yielded the greatest average effect size ($d = .64$), followed by unpublished studies ($d = .58$), theses and dissertations ($d = .48$), and studies reported in books ($d = .30$). The magnitude of this bias is not typical and reviewers should recognize that their selection strategy can have a considerable impact on their combined results.

An equally important implication of biased selection is that the published literature contains most, if not all, of the Type I errors of inference (cf. Bakan, 1966; Cohen, 1962; Greenwald, 1975; McNemar, 1960). Estimates like Rosenthal's fail-safe number are helpful, but they are not complete solutions. The fact remains that such estimates are based on sample values that may be unrepresentative of the population, a population with essentially unknown parameters.

Finally, biased selection of studies may preclude sampling of important methodological and theoretical variables. Published sources tend to contain studies that "worked" and that "fit" into current scientific thought (Zeitgeist). Studies that use novel methods or that debunk a popular theory are given closer scrutiny, which will more likely reveal the fatal flaw that keeps them out of print. Despite our claims to objective science, the publication process is not entirely apolitical (Glass et al., 1981), and studies that contain important information may be excluded for irrelevant reasons. For example, Smith (1980) examined the presence of sex bias in psychotherapy and found a substantial tendency for counselors and therapists to stereotype women and view them more negatively than men, but *only* in published studies. Unpublished studies showed a substantial difference in the *opposite* direction. As this study makes clear, biased selection can affect the direction as well as the magnitude of the results that one obtains in a meta-analysis.

There is no simple solution to the biased-selection problem. In some research areas it is possible to collect and use as large and complete a sample as possible. In other areas it may be necessary to either narrow one's theoretical focus or to restrict the sampling to keep the analysis manageable. For example, Miller (reported in Glass et al., 1981) examined the literature on the psychological effects of drug therapy. Even though he restricted his sample to published sources and made further restrictions based on research design and patient characteristics, 2,963 studies were located. A representative subsample were read and coded for the meta-analysis. Where large portions of a literature are omitted, the possibility of bias certainly exists. The effects of such bias can be offset somewhat if reviewers take pains to describe their sampling procedures in detail (Glass et al., 1981). This description allows the reader to assess the conclusions made in a review in light of the quality of the sample.

Data Retrieval

Problems in the reporting of results in individual studies can also produce biases and distortions in a review's data base (cf. Feldman, 1971). Selective reporting of results is particularly troublesome. Researchers are likely to report and emphasize findings that are significant and in the expected direction. Null or unexplainable results may be regarded as unimportant and unworthy of mention since they detract from the "clean"

presentation of the results. Unfortunately such suppression systematically biases a meta-analysis, and the detection of the bias may be impossible, particularly when crucial dependent measures are not even mentioned in method sections (cf. Shapiro & Shapiro, 1982).

Another problem in reporting that can distort the data retrieved for a meta-analysis concerns the completeness and accuracy of results presented in written reports. The meta-analytic techniques use quantitative information and can provide precise, accurate results only to the extent that precise, accurate information can be retrieved. Often the necessary statistical information can be reconstructed. For example, Glass and his colleagues (Glass, 1980; McGaw & Glass, 1980; Glass et al., 1981) present a variety of procedures for estimating effect sizes when limited statistical information only is available. In respect to significance levels, the conservative approach taken is usually to assume $p = .50$ when "no difference" is reported, and $p = .05$ when all that is reported is "significant differences favoring Group X were obtained." However, when one must resort to reconstruction or estimation, the data become less precise because of the unverifiable, and sometimes tenuous, assumptions that must be made (e.g., equality of treatment- and control-group variances).

Reporting accuracy applies not only to effect sizes and significance levels that serve as the "dependent variables" in a meta-analysis but also to the host of methodological variables and study characteristics that can serve as "independent variables." For example, adequate descriptions of sample characteristics (age, sex, ethnicity, IQ, etc.), treatments (type, how implemented), settings (home, office, university), therapists (experience, training), and procedures (controls, type of outcome measures), to name but a few, are *necessary* if the mediating impact of these variables on outcome is to be assessed adequately. Imprecise reporting of these variables is especially serious since they may be impossible to estimate.

Data Quality

A vital issue in the conduct of a meta-analysis is the quality of the data base. At some point, the reviewer must decide which studies from the sample will be included in the analysis. The problem of inclusion criteria has caused considerable debate (e.g., Eysenck, 1978; Rachman & Wilson, 1980). At one extreme is the argument to include all studies in a meta-analysis, under the assumption that if the "flaws" do not even out (cf. Cook & Leviton, 1980), then the analysis will indicate where any biases exist. Coding study quality (e.g., Smith et al., 1980; Shapiro & Shapiro, Note 2) and assessing the relationship of quality to outcome represents an empirical solution to the problem that has been used with some success. Glass et al. (1981) review 12 studies that examined the relationship of internal validity to study outcome. The results were mixed and depended on the content area being examined. In some cases low-quality studies yielded higher effect sizes than high-quality studies, sometimes the opposite was found, and sometimes no relationship between study quality and outcome was obtained. The advantage to coding the quality of study and examining the relationship of quality to study outcome is that it allows an examination of important methodological variables. This examination, in turn, can aid future research by pointing out potential procedural problems.

At the other extreme are research domains in which an adequate theoretical test must satisfy very restrictive conceptual as well as methodological requirements (e.g., sleeper effect, see Cook & Leviton, 1980). Inclusion of inappropriate (not simply weak) tests could produce misleading conclusions, particularly where only a small sample of studies satisfy the requirements of the theory (cf. Jackson, 1980). In these cases, clear decision rules are required a priori to insure that the combined analysis adequately addresses the substantive issues. This requirement does not necessarily mean that inappropriate tests are not analyzed. A useful middle ground might be to contrast the analyses of appropriate and inappropriate tests in order to explicate the key methodological and theoretical differences (and perhaps discover a few new ones).

Regardless of one's preference for level of inclusion, the strategy used has another implication. In a strict sense, the conclusions drawn from a meta-analysis apply only to the sample of studies examined; the results can

be generalized to studies differing in various ways only with some (perhaps extreme) caution. If the samples are based on inappropriate or poor quality data, the conclusions become suspect. Whether one opts for restrictive inclusiveness, or coding of data quality (or some middle position on this inclusiveness continuum), one should have clear guidelines by which to make decisions. We have previously outlined three steps in examining data quality (Strube & Hartmann, 1982) that bear repeating here. These steps require considering three types of validity: conceptual, methodological, and statistical.[4]

Conceptual validity. Conceptual validity refers to the necessity that putting a theoretical construct into operation constitutes a valid test of the phenomenon under study (see the discussion of construct validity by Cook & Campbell, 1979). If a study does not test what it purports to test, then it should not be included in the analysis. Related to this point is the fact that similarly labelled treatments or programs may not test the same underlying process (cf. Pillemer & Light, 1980a). The reviewer must consider carefully whether a particular treatment was implemented appropriately and must not take the investigator's interpretation as the sole basis for a decision. The requirement of conceptual validity can work another way. It is possible to identify studies that were designed for other purposes but that implicitly test the phenomenon of interest. These studies should be included. The consideration of conceptual validity again points up the role of human thought in the review process. A clear strategy (whether for inclusion, exclusion, or coding) should be available so that reviewer decisions are reliable and defensible. If clear decision rules cannot be derived, then a research area is not "ready" for meta-analysis, since it lacks the development to define its domain (cf. Cooper, 1979).

Methodological validity. Provided that a study satisfies at least minimal requirements of conceptual validity, the reviewer must next examine its methodological validity. This includes examination of internal validity (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979) as well as the many methodological and procedural variables that are endemic to a particular research area (e.g., see the special issue on methodology in the *Journal of Consulting and Clinical Psychology*, 1978; Hartmann, Roper, & Gelfand, 1977; Strube & Hartmann, 1982). Essentially, examination of methodological validity entails determining the quality of the hypothesis test. Critical examination at this stage identifies potentially important mediating variables that should be examined in subsequent analyses, particularly if inclusion criteria are less restrictive. For example, Smith et al. (1980) found that the reactivity of the outcome measures, and the number of months posttherapy for the follow-up assessments, were important methodological mediating variables in the analysis of psychotherapy outcome.

Statistical validity. A final source of validity that must be examined is statistical validity. Although a study may be conceptually and methodologically sound, it is still possible to apply improper statistical techniques in testing the hypothesis (see Cook & Campbell, 1979, for a discussion of statistical conclusion validity). Not only must the important assumptions associated with a given statistic be satisfied (e.g., uncorrelated errors) but the particular comparison or contrast reported must parallel the question that is being addressed. An additional level of complexity arises when the primary study does not provide the exact comparison or contrast that is of interest to the meta-analyst (even though it does address the primary investigator's question). In this case, the meta-analyst may be able to reconstruct the appropriate comparison *if* full reporting of results is available. The point is that the meta-analyst must be certain that the appropriate statistical information is retrieved from individual studies.

Examination of data quality, deciding what studies should be included, and how those studies should be coded, represent several of the more subjective decision points in a meta-analysis. The development of clear decision

[4] Cook and Campbell (1979) provide a thorough presentation of the various types of validity that should be examined at the single-study level. Most, but not all, of their discussion is also applicable to meta-analysis. However, a meta-analysis requires several additional considerations and thus a somewhat different nomenclature is used.

rules and categorization schemes (for study characteristics, methodological and theoretical variables) serves to systematize the process and at least makes the reviewer's selection and classification procedures available for scrutiny by others (e.g., Smith et al., 1980). Clear rules, when coupled with further analyses through disaggregation or multiple regression, should enhance the detection of both important biases and mediating variables (see Strube & Hartmann, 1982 for a further discussion of these issues in respect to the psychotherapy literature).

Independence

The meta-analytic techniques described earlier require independence of the individual hypothesis tests combined (e.g., Jones & Fiske, 1953; Rosenthal, 1978). Unfortunately, this requirement often may go unmet because several types of nonindependence can occur in a meta-analysis. Quite commonly, subjects provide responses on several dependent measures, each of which is analyzed statistically and used in a subsequent meta-analysis (e.g., Smith et al., 1980). Solutions to the problem of multiple-correlated statistical tests from a single study include differentially weighting the various measures or selecting the one most appropriate in relation to the hypothesis (e.g., Strube, 1981; Strube & Garcia, 1981). However, when the meta-analysis focuses on the relationship of different types of measures to study outcome, neither of these solutions is appropriate. The approach then taken is to analyze the results under the assumption that the data are independent. As Glass et al. (1981) point out, this solution is practical but risky (and the assumption is patently false). The result is to ignore the effect of complex interdependencies in the data that can drastically affect the standard errors of parameter estimates. Glass et al. illustrate the problem using data from 14 studies on class size and achievement, which gave rise to 108 different comparisons. Using Tukey's jackknife technique (Mosteller & Tukey, 1968), Glass et al. found that the confidence interval for the regression coefficient was over 350% wider when interdependencies were taken into account. An even more accurate procedure appears to be generalized least squares (Glass et al., 1981), which allows correlated errors (dependencies) and is better known than the jackknife procedure. In either case, meta-analysts are well advised to consider the complex interdependencies that may exist in their data lest they place too much confidence in their point estimates.

A second source of "nonindependence" arises when a given researcher or research team conducts multiple-hypothesis tests (cf. Rosenthal, 1976). In this case, results may be "correlated" due to common methodology, similar subject samples, or any other of a host of variables that are idiosyncratic to a given research lab. This kind of bias can be detected. For example, Strube and Garcia (1981) in their review of Fiedler's contingency model of leadership effectiveness disaggregated their results according to whether or not the researchers came from Fiedler's "camp." The results from both groups supported Fiedler's model but the results from Fiedler and his associates were somewhat more supportive. This type of nonindependence affects both our confidence in the results of the meta-analysis and the external validity of the analysis (cf. Rosenthal, 1976). Results that are replicated by a number of independent researchers using different procedures and samples bolster our confidence in the actual existence of an effect. Likewise, replication across methods and samples increases the external validity of the results.

Technique-Produced Variability and Statistical Considerations

As outlined previously, meta-analytic investigators have a veritable arsenal of statistical techniques at their disposal. A wide variety of techniques exist for calculating and interpreting effect sizes (e.g., Glass, 1980; Glass & Hackstian, 1969; Mitchell & Hartmann, 1981; see also Cooper, 1981) and combining probabilities (e.g., Rosenthal, 1978, 1980). Further analysis of retrieved values can proceed along a number of lines. Such choice leaves open the possibility that the results of a meta-analysis may vary depending on the specific techniques used (see Strube, 1981). This variability can be particularly problematic when a meta-analysis is con-

ducted on a small data set in which the results may be quite unstable (see also Cook & Leviton, 1980). Although recommended "packages" of techniques may be useful (e.g., Cooper, 1979; Strube & Hartmann, 1982), the meta-analyst should not use these suggestions as a crutch. For every meta-analysis, the reviewer must decide what set of techniques will most adequately address the questions of importance. Within the general recommendation that a review concern itself with the overall combined probability, average effect sizes, estimates of stability, and analysis of mediating variables, there is considerable latitude for decisions by the reviewer.

A second area of caution concerns the application of inferential techniques (e.g., regression analyses, ANOVA) to data from a meta-analysis. Inferential statistics require the random and independent sampling of units from a known population or, at least, the random assignment of units to conditions of an experiment (Glass et al., 1981). As pointed out earlier, meta-analytic samples will rarely be random samples from a known population. The individual studies and hypothesis tests will also not be randomly assigned to their "levels" of study characteristics or of methodological and theoretical variables. A meta-analysis is at best a quasi-experimental design and is most commonly a nonexperimental design. Whereas the use of inferential statistics aids the meta-analyst in his or her quest for precise delineation of effects, the nature of the "design" should temper his or her zeal for making unwarranted causal conclusions. In addition, the meta-analyst should be wary of the number of hypothesis tests conducted and the dangers of inflating the review-level Type I error-rate.

The problems that arise from the application of meta-analytic techniques are, for the most part, not limited to meta-analytic studies (cf. Cook & Leviton, 1980; Leviton & Cook, 1981). Any study must be concerned with sample selection, validity of measures and manipulations, and sources of bias (e.g., see Feldman, 1971; Taveggia, 1974). However, the desire of the meta-analytic reviewer to make more precise, probabilistic statements based largely on statistical information forces a closer look at the implications of such application. This closer look, while organizing the review process, also makes explicit the large numbers of important problems that can arise.

The Functions of Meta-Analysis

We now turn to a consideration of the many uses or functions that a meta-analysis may serve. Again we will find that many useful applications of meta-analysis are not inherent in the techniques. However, by explicating these uses, future meta-analyses may capitalize on the greater power and precision of the approach.

One major purpose that a research review should serve is what can be called a *descriptive function*. By description we mean not only the general summarizing of results, but also the categorization and summarization of study characteristics, including methodological, procedural, and theoretical variables. A research review should not only tell us what we know but also how we obtained that knowledge. An often neglected aspect of a research review is the external validity of the data on which inferences are based. The claim that "research overwhelmingly supports Theory A" should always be qualified by consideration of the methods, subjects, settings, and measures used in the sample of investigations (cf. Cooper, 1979; Leviton & Cook, 1981). These ends are met by a thorough description of the sample, which is facilitated by the systematic procedures of meta-analysis.

A detailed description can facilitate two additional purposes of a review. First, the accumulation of research evidence in a given area carries with it an accumulation of research methods and procedures. A researcher embarking on a new investigation should be able to consult a review and find what methods have worked best in the past, or ideally, what methods best match the question under study. The evolution of theory proceeds most efficiently when the accompanying methodology is continuously improved (Cooper, 1979). Inappropriate or inadequate methods can be discarded most efficiently when precise appraisal of their past effectiveness is

compiled. A rather important corollary to this use of meta-analysis must be mentioned. The fact that a meta-analysis indicates that methodological variables do not mediate study outcome *must not* be interpreted as meaning that any methodology can be used successfully in future research. The failure to find a correlation between a coded methodological variable and study outcome may be due to restriction in the range of levels of that methodology. Alternately, it may be that a wide array of methods has been used and that they all were particularly well matched to the research question. The result would be that all methods appear equally effective—but only because considerable thought went into their selection by individual investigators. No less thought should be given to their examination at the review level.

In addition, a thorough description aids the application of theory to real-world problems. A reviewer serves the very important role of gatekeeper in the transmission of information to practitioners. Whether a theory or treatment is ready for application, and its proper domain of application, can best be determined when the information in a review is presented in detail. This detail allows practitioners to determine whether broad generalizations are warranted or whether more detailed, specific applications are called for. Thorough description allows a further advantage. A substantial amount of empirical research is laboratory-based. The increased control and precision of the laboratory is bought with increased artificiality and presumably lower external validity. However, as Henshel (1980) convincingly points out, beneficial effects produced in the laboratory *may* be reproducible in the real world, even if the appropriate conditions do not presently exist. The precise accumulation of laboratory results can help specify the exact conditions that must be created to produce real-world benefits.

Consistent with the goal of "taking stock" of what we know in a given area is the parallel goal of determining what we do not know. Any review (but a meta-analysis in particular) can serve a diagnostic function by identifying gaps in our knowledge. A theory typically rests on several key postulates that lead to testable hypotheses. A meta-analysis can provide a precise appraisal of the studies that have tested each hypothesis. Such analyses usually uncover glaring weaknesses in the empirical assessment of a given theory. For example, Strube and Garcia (1981) reviewed evidence testing Fiedler's contingency model of leadership effectiveness. The model provides testable hypotheses for eight different situational contexts. Overall support for the model was strong, but the review also identified several of the eight contexts in which additional study was needed. A meta-analysis can thus help to identify "holes" in the "nomological net" of multiple empirical relationships that constitute the tests of a theory (cf. Cook & Leviton, 1981; Cronbach & Meehl, 1955). An additional advantage of identifying areas in need of research is that fewer research efforts will be directed toward areas that have been heavily investigated. The result will be more efficient use of research resources and less accumulation of redundant information (Rosenthal, 1979b; Smith & Glass, 1977).

A review can also serve a predictive function. This is perhaps the most neglected of a review's potential uses but is one that is particularly well suited to the more statistical orientation of meta-analysis. By predictive we mean the ability to examine the plausibility of hypotheses that have not been tested in single studies (cf. Feldman, 1971; Pillemer & Light, 1980a, 1980b; and the exploratory mode of reviewing outlined by Cook & Leviton, 1980). Because each data point in a meta-analysis is a study with its own methodological and theoretical characteristics, it is possible to "construct" variables and test their relationship to study outcome. As an example, one might determine why or under what conditions subjects sought psychotherapy and thus classify studies along a "commitment to change" dimension. Relating this new variable to study outcome might reveal interesting results to be followed up with subsequent research. Furthermore, the use of regression analysis allows one to predict or estimate study outcome given specific values of independent variables (e.g., Smith & Glass, 1977). The values used in a regression analysis need not have existed in any one study.

Given a sufficient number of actual values, one can estimate the outcome for hypothetical values. The vital corollary to this use of accumulated data is that one *must* follow up such data snooping with an empirical test in which the values of the independent variables exist in the *same* study. The predictive use of meta-analysis establishes plausible hypotheses, it does not actually test them. However, if used cautiously, an accumulated data base can provide the foundation for considerable exploration of a theory's uncharted domain.

Obviously, the above functions are not independent of one another. Thorough description is required in order to identify gaps in empirical research and test new hypotheses. The predictive function of a meta-analysis might estimate the likely results in these uncharted areas. The important point is that each function provides an important emphasis that can be brought to bear on the data. Theory development and application will proceed most effectively by using all the information contained in accumulated research.

Unresolved Issues and Future Directions

The use of meta-analysis will likely become standard practice in the conduct of future literature reviews. This should not lead potential reviewers to believe that the approach is trouble free. As outlined in this article, the approach contains all the problems of the more traditional narrative review, plus its own unique set of statistical-based problems. Meta-analysis offers greater precision in the summarization of research evidence, but this precision will be illusory if the statistical techniques are misapplied.

The issue of sampling is one of the more vexing problems in the conduct of a meta-analysis, since biased sampling limits the generalizability of results. An ambitious solution to this problem would be the establishment of central repositories for studies in well-defined content areas (cf. Rosenthal, 1976; Smart, 1964). To simplify the process, such studies might be summarized in brief or abstract form so that reviewers could judge the relevance of the study and then obtain more detailed information from the primary investigators.

The solution to inaccurate reporting in published research rests with the primary investigators and journal editors. As more reviews use the meta-analytic approach, sensitivity to reporting accuracy should increase. In the meantime, examination of the validity of methods for estimating missing or partial data should be examined, given the impact of these procedures on the precision of results.

Finally, we would like to advocate the expanded use of meta-analysis to data bases unrestricted by traditional theoretical boundaries. We thus propose a *generative* function for meta-analysis that is an extension of the predictive function described earlier. Several important theoretical frameworks have been generated from data bases collected for other purposes. Examples include Fiedler's (1967) development of the contingency model of leadership effectiveness and Zajonc's (1965) drive-theory explanation for social facilitation. There is no logical reason why existing data cannot be reinterpreted. What we propose is analogous to the use of secondary analyses (Cook, 1974) of primary data (cf. Feldman, 1971). The creative investigator can explore hypotheses through clear conceptualization of a phenomena and judicious choice of studies (see McGuire, 1973, for an insightful discussion of how such hypotheses might be generated). An advantage to using a meta-analytic approach is that the greater precision could identify the more promising areas on which to focus initial empirical investigation. As a result, the identification and development of new theoretical frameworks could proceed more efficiently. Obviously, several of the problems inherent in a traditional meta-analysis are exacerbated in the generative use of the approach. Selection of the sample becomes more difficult because studies will not be classified according to the proposed reinterpretation. The subjective decisions of the traditional meta-analysis are compounded by the reinterpretation that the investigator wishes to impose on the data. These and other problems will require careful attention.

Conclusions

We have outlined the major techniques and uses of meta-analysis in the hope of clarifying both the advantages and limitations that currently exist. As has been noted throughout this presentation, meta-analysis consists of a series of complex, subjective, and sometimes arbitrary-seeming decisions. The potential meta-analyst should not be seduced by the quantitative nature of the approach. Rather, the meta-analytic review should be approached with the same care and thoughtfulness of any scholarly endeavor. The effective solution of the problems made obvious by the systematic approach of meta-analysis will ultimately enhance the utility of our ever-growing research literatures.

Reference Notes

1. Schmidt, F. L. *Moderator research and the law of small numbers.* Paper presented at the Conference of Moderator Research, University of Maryland, March, 1977.
2. Shapiro, D. A., & Shapiro, D. *Meta-analysis of comparative therapy outcome studies: A replication and refinement.* Paper presented at the annual meeting of the Society for Psychotherapy Research, Aspen, Colorado, June, 1981.
3. Rosenthal, R., & Rubin, D. B. *Comparing effect sizes of independent studies.* Unpublished manuscript, Harvard University, August, 1981.
4. Underwood, B., & Dickson, P. *The significance bias and estimation of effect size.* Unpublished manuscript, Southwestern Data Consultants, Austin, Texas, 1981.

References

Bakan, D. The test of significance in psychological research. *Psychological Bulletin*, 1966, *66*, 432–437.

Blanchard, E. B., Andrasik, F., Ahles, T. A., Teders, S. J., & O'Keefe, D. Migraine and tension headache: A meta-analytic review. *Behavior Therapy*, 1980, *11*, 613–631.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1963.

Chase, L. J., & Chase, R. B. A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, 1976, *61*, 234–237.

Cohen, J. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 1962, *65*, 145–153.

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press, 1977.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Wiley, 1975.

Cook, T. D. The potential and limitations of secondary evaluations. In M. W. Apple et al. (Eds.), *Educational evaluation: Analysis and responsibility.* Berkeley, Calif.: McCutchan, 1974.

Cook, T. D., & Campbell, D. T. (Eds.), *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally, 1979.

Cook, T., & Leviton, L. Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, 1980, *48*, 449–471.

Cooper, H. M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 1979, *37*, 131–145.

Cooper, H. M. On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology*, 1981, *41*, 1013–1018.

Cooper, H. M. Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, in press.

Cooper, H. M., & Arkin, R. M. On quantitative reviewing. *Journal of Personality*, 1981, *49*, 225–230.

Cooper, H. M., & Rosenthal, R. Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 1980, *87*, 442–449.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281–302.

Eysenck, H. J. An exercise in mega-silliness. *American Psychologist*, 1978, *33*, 517.

Feldman, K. A. Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, 1971, *44*, 86–102.

Fiedler, F. *A theory of leadership effectiveness.* New York: McGraw-Hill, 1967.

Gallo, P. S., Jr. Meta-analysis—A mixed metaphor? *American Psychologist*, 1978, *33*, 515–517.

Glass, G. V. Primary, secondary, and meta-analysis research. *Educational Researcher*, 1976, *5*, 3–8.

Glass, G. V. Integrating findings: The meta-analysis of research. *Review of Research in Education*, 1977, *5*, 351–379.

Glass, G. V. Summarizing effect size. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science: Quantitative assessment of research domains.* San Francisco: Jossey-Bass, 1980.

Glass, G. V., & Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, 1969, *6*, 403–414.

Glass, G. V., McGaw, B., & Smith, M. L. *Meta-analysis in social research.* Beverly Hills, Calif: Sage Publications, 1981.

Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 1975, *82*, 1–20.

Hartmann, D. P., Roper, B. L., & Gelfand, D. M. Evaluation of alternative modes of child psychotherapy. In B. Lahey & A. Kazdin (Eds.), *Advances in child clinical psychology* (Vol. 1). New York: Plenum Press, 1977.

Hedges, L. V., & Olkin, I. Vote-counting methods in research synthesis. *Psychological Bulletin*, 1980, *88*, 359–369.

Henshel, R. L. The purpose of laboratory experimentation and the virtues of deliberate artificiality. *Journal of Experimental Social Psychology*, 1980, *16*, 466–478.

Hunter, J. E., & Schmidt, F. L. Differential and single group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology*, 1978, *63*, 1–11.

Hyde, J. S. How large are cognitive gender differences? A meta-analysis using ω^2 and *d*. *American Psychologist*, 1981, *36*, 892–901.

Jackson, G. B. Methods for integrative reviews. *Review of Educational Research*, 1980, *50*, 438–460.

Jones, L. V., & Fiske, D. W. Models for testing the significance of combined results. *Psychological Bulletin*, 1953, *50*, 375–382.

Journal of Consulting and Clinical Psychology, 1978, *46*, 595–838.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research*. New York: Holt, Rinehart, & Winston, 1973.

Klitgaard, R. Identifying exceptional performers. *Policy Analysis*, 1979, *3*, 529–547.

Lane, D. M., & Dunlap, W. B. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 107–112.

Leviton, L. C., & Cook, T. D. What differentiates meta-analysis from other forms of review. *Journal of Personality*, 1981, *49*, 231–236.

Light, R. J. Capitalizing on variation: How conflicting research findings can be helpful for policy. *Educational Researcher*, 1979, *8*, 8–11.

Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 1971, *41*, 429–471.

McGaw, B., & Glass, G. V. Choice of the metric for effect size in meta-analysis. *American Educational Research Journal*, 1980, *7*, 325–337.

McGuire, W. J. The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, 1973, *26*, 446–456.

McNemar, Q. At random: Sense and nonsense. *American Psychologist*, 1960, *15*, 295–300.

Mitchell, C., & Hartmann, D. P. A cautionary note on the use of omega squared to evaluate the effectiveness of behavioral treatments. *Behavioral Assessment*, 1981, *3*, 93–100.

Mosteller, F. M., & Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 1). Cambridge, Mass.: Addison-Wesley, 1954.

Mosteller, F. M., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (2nd ed., Vol. 2). Reading, Mass.: Addison-Wesley, 1968.

Pillemer, D. B., & Light, R. J. Benefits from variation in study outcomes. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science: Quantitative assessment of research domains*. San Francisco: Jossey-Bass, 1980. (a)

Pillemer, D. B., & Light, R. J. Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review*, 1980, *50*, 176–195. (b)

Rachman, S. J., & Wilson, G. T. *The effects of psychological therapy: Second enlarged edition*. New York: Pergamon Press, 1980.

Rimland, B. Death knell for psychotherapy? *American Psychologist*, 1979, *34*, 192.

Rosenthal, R. *Experimenter effects in behavioral research* (Enlarged ed.). New York: Irvington, 1976.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 1978, *85*, 185–193.

Rosenthal, R. The "file-drawer problem" and tolerance for null results. *Psychological Bulletin*, 1979, *86*, 638–641. (a)

Rosenthal, R. Replications and their relative utilities. *Replications in Social Psychology*, 1979, *1*, 15–23. (b)

Rosenthal, R. Summarizing significance levels. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science: Quantitative assessment of research domains*. San Francisco: Jossey-Bass, 1980.

Rosenthal, R., & Rubin, D. B. Comparing significance levels of independent studies. *Psychological Bulletin*, 1979, *86*, 1165–1168.

Rosenthal, R., & Rubin, D. B. Summarizing 345 studies of interpersonal expectancy effects. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science: Quantitative assessment of research domains*. San Francisco: Jossey-Bass, 1980.

Rosenthal, R., & Rubin, D. B. A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, in press.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 1977, *62*, 529–540.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, 1979, *32*, 257–291.

Shapiro, D. A., & Shapiro, D. Meta-analysis of comparative therapy outcome research: A critical appraisal. *Behavioral Psychotherapy*, 1982, *10*, 4–25.

Smart, R. G. The importance of negative results in psychological research. *The Canadian Psychologist*, 1964, *5*, 225–232.

Smith, M. L. Integrating studies of psychotherapy outcomes. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science: Quantitative assessment of research domains*. San Francisco: Jossey-Bass, 1980.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, *32*, 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. *The benefits of psychotherapy*. Baltimore, Md.: Johns Hopkins University Press, 1980.

Sohn, D. Critique of Cooper's meta-analytic assessment of the findings on sex differences in conformity behavior. *Journal of Personality and Social Psychology*, 1980, *39*, 1215–1221.

Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of American Statistical Association*, 1959, *54*, 30–34.

Strube, M. J. Meta-analysis and cross-cultural comparison: Sex differences in child competitiveness. *Journal of Cross-Cultural Psychology*, 1981, *12*, 3–20.

Strube, M. J., & Garcia, J. E. A meta-analytic investigation of Fiedler's contingency model of leadership effectiveness. *Psychological Bulletin*, 1981, *90*, 307–321.

Strube, M. J., & Hartmann, D. P. A critical appraisal of meta-analysis. *British Journal of Clinical Psychology*, 1982, *21*, 129–139.

Taveggia, T. C. Resolving research controversy through empirical cumulation: Toward reliable sociological knowledge. *Sociological Methods & Research*, 1974, *2*, 395–407.

Tedeschi, J. T., Gaes, G. G., Riordan, C., & Quigley-Fernandez, B. Social psychology and cumulative knowledge. *Personality and Social Psychology Bulletin*, 1981, *7*, 161–172.

Underwood, B., & Moore, B. Perspective-taking and altruism. *Psychological Bulletin*, 1982, *91*, 143–173.

Zajonc, R. B. Social facilitation. *Science*, 1965, *149*, 269–274.

13

Archival Data in Program Evaluation and Policy Analysis

James W. Luckey, Andy Broughton, and James E. Sorensen

Accountability through the evaluation of human service programs has received increased emphasis in recent years. The social security, health, mental health and rehabilitation acts, among others, have either mandated or strongly urged built-in evaluation systems. Concurrent with these requirements has come increasing use of data-based support by policy analysts.

Despite the growing overlap in function between ongoing data systems, evaluation research and policy analysis, gaps still remain among the three. One method to reduce these gaps has been the design of management information systems with the goal of evaluation specifically in mind (e.g., Chapman, 1976; Sorensen & Elpers, 1978). Another strategy has been retrospective analysis of existing data bases. This article focuses on the potential problems of the latter approach.

Evaluators historically have avoided using archival data, preferring instead to collect their own data to insure control over the content and process of data collected. However, there are some real advantages to using existing data sources. The most obvious of course, are the potential savings in time, money and effort achieved by sidestepping the original data collection process. These considerations are likely to make archival data more attractive to evaluators as demands for analyses increase and resources decrease. Archival data also has some methodological advantages, one being the information is generally non-reactive to the specific purpose of the present evaluation (Webb, Campbell, Schwartz, & Sechrest, 1966). Another is

From James W. Luckey, Andy Broughton, and James E. Sorensen, "Archival Data in Program Evaluation and Policy Analysis," *Evaluation and Program Planning*, 1982, 5(3), 247–253.

that archival data are often available for extended periods of time and for a variety of populations. Depending on the nature of the data and the ingenuity of the analyst, either of these may allow for a variety of quasi-experimental designs in answering evaluation and policy-related questions (Campbell & Stanley, 1963; Cook & Campbell, 1979). Although strong arguments have been marshalled for the use of randomized experiments in evaluation (e.g., Boruch, 1976; Apsler, 1977), prospective, randomized experiments may not always be possible because of ethical considerations or feasible because of resource or time constraints. Often the evaluator or policy analyst does not enjoy the luxury of enough time to produce research results to influence the decision-making process. The only alternative is to utilize data that are already available.

This article first identifies types of problems encountered in using archival data for policy analysis and evaluation. Following this general discussion, an example will be presented to illustrate specific difficulties and potential pitfalls of using existing data sets. While other authors such as Weiss (1974) have discussed the inadequacies of existing data sets, this article goes beyond a description of these potential problems and will present a series of strategies for anticipating, assessing and overcoming problems with archival data.

GENERAL PROBLEMS WITH ARCHIVAL DATA

Evaluations using archival data risk severe limitations. The general considerations in using an existing data collection system not specifically designed with evaluation in mind include the appropriateness, accuracy and accessibility of the information contained in the system.

Appropriateness

The primary concern about archival data is appropriateness for the particular evaluation effort. Appropriateness embodies both the type and form of the information available. When evaluators have the luxury of collecting their own data, the type and format can be tailored to the purpose of the evaluation. When existing data are used, however, there is the temptation to tailor the evaluation to the data.

Purpose of Data Collection. The appropriateness of the existing data is frequently problematic because the original rationale for collecting the data was different from the reason for the current evaluation. Data collected by fiscal intermediaries to document reimbursement for accounting reports are shaded by the reimbursement purpose (e.g., some services qualify for reimbursement while others, although equally acceptable in some fields of practice, do not). Data collected for research purposes are usually free of this possible contamination. In general, the less congruence between the two purposes, the less likely the data will be of real use in an evaluation. The evaluator is then faced with a decision of altering the evaluation, generating new data or scrapping the evaluation effort. Given the external pressure for accountability the first option frequently is tempting.

Quantity Oriented Data Bases. Appropriateness of the information is a common problem where data collected by governmental agencies emphasizes documenting the quantity of effort (e.g., number and kinds of persons served). Such information may have some use in process evaluations but is usually of dubious value in outcome evaluations without the use of questionable assumptions. For example, following the introduction of a new law intended to protect the rights of involuntarily admitted patients, psychiatric hospital utilization patterns were examined (Luckey & Berman, 1979). The average length of stay decreased following this intervention and could be interpreted as an indication of successfully decreasing infringement on the rights of patients. However, an alternative conclusion could be a "revolving door" phenomenon exchanging a few longer hospitalizations for many short ones with more frequent disruptions in the patient's life. This latter conclusion was supported by a significant increase in the number of readmissions. Utilization patterns support both interpretations; an assessment of the desirability of these changing patterns of care could not be determined by this information alone.

Data Format. A related but less obvious difficulty has to do with the form of data collection or storage. Information may be aggregated, for example, across time, programs, geographic units, or in other ways different from those required for evaluation. For example, in the CHAMPUS study cited later, the data provides an example where the unit of interest for the evaluation project was the inpatient psychiatric admission while the data had been collected and stored by provider reimbursement claims.

Accuracy

Accuracy of the archival data includes the quality of the data (reliability and validity) and the significance of the information.

Reliability. Quality of the data is a major concern in a large data collection system. One aspect is mechanical. The greater the number of discrete steps involved in going from original source to data analysis (i.e., collection, coding, keying, etc.), the greater the potential

for the introduction of error into the system. When using archival data, the user has no control over the reliability of the information in the system. The only alternative is to check on the system to arrive at some estimate of the accuracy of the various steps.

Validity. A second aspect of quality is the validity of the data. A problem arises from both the number of levels involved in the data acquisition process and the lack of control the evaluator has over the system. Generally, the protocol for the data is provided by the administrative branch of the organization (i.e., data management personnel) while the information in the system is generated at the program level. Congruence between these levels in the perception of the meaning of a particular piece of information is important, especially when the evaluator has to rely on the description provided by the data management personnel. Factors other than differing perception of the meaning of the data impinge on the validity of the data such as response bias, perception of how the information is to be used, stability of the characteristics included, the number of possible alternatives for a given item and the level of abstraction of the information. The evaluator may be easily seduced into relying on a view of their immediate source of the information, the data system personnel, which may or may not correspond with the view of the people at the program level.

Time Span. The longer the time period of the evaluation, the greater the concerns with the reliability and validity of archival data. Because of growth of the data system and turnover in personnel, both common occurrences, the passage of time can increase the number of people involved both at the program and data management levels. Staff turnover has the tendency to increase the possibility of problems with both the reliability and validity of the data. Long time spans also increase the possibility of system-wide changes, either through refinements of the data management system, abrupt changes in the system reporting requirements or alternation of the meaning of an item through changes in factors external to the system (e.g., changes in the diagnostic system). While current personnel should be aware of the present status of the system, historic changes in instrumentation pose potential problems.

Experience with state hospital admission data provides an example of the problems with changes in instrumentation over time. One key variable was the type of commitment used to admit a patient. Since commitment laws had been changed three times over the ten year data period, coding schemes were suspected to have also changed. This suspicion was bolstered by a visual examination of the data revealing discontinuities in the commitment codes coincident with the legal changes. It was only when these observations were made that the data management personnel could be asked to search their files and eventually were able to locate documentation on prior codes. Only with this additional information could valid conclusions be drawn.

Significance. Concerns about the quality of the data relate to the data acquisition and storage process and are internal to the system. Concerns about the significance of the data on the other hand are evaluation based. One frequent limitation of archival data is the availability of a small amount of information for a large population of cases and sometimes leads the evaluator to impute more significance to the data than is justified—a problem similar to operational definitions in traditional experimental designs. In mental health data systems, for instance, one almost universally available item is diagnosis. In the absence of additional information, diagnosis may be equated with either severity of illness or level of functioning. Diagnosis as a surrogate measure for severity or functioning would be attaching unrealistic significance to the available data.

Understanding the Program. A major difficulty with evaluators relying solely on archival data is the potential for an evaluation to become an exercise in the manipulation of large numbers of numbers without a real understanding of the program being evaluated. Basing perceptions on official program descriptions or administrative viewpoints only can result in a limited and biased view of the actual program objectives and operations.

Accessibility

While archival data has already been collected, existence does not assure accessibility. One issue is the confidentiality or the right to privacy of the individual's information in a system. In general, access to most existing data systems requires removal of all identifying information for individuals. Tracking of individuals through the system may not be possible (e.g., matching repeat episodes for the same person) and matching the data at the individual level with other data sources may be nearly impossible. A frequent method used to assure confidentiality is for the system to provide only aggregate data to the evaluator. Checking the reliability and validity of the information in aggregate form is more difficult, however. In addition, aggregated data can mask important information and prevent an examination for any sub-aggregate trends.

Political Barriers. While not unique to archival data, negative exposure from an evaluation is a further barrier in gaining access to archival data. Political aspects of evaluation have been widely discussed in the

literature (e.g., Downs, 1971; Rossi, 1972). Refusal of access to data often (under the guise of confidentiality) can be a hidden reaction to this threat.

Hardware and Software Barriers. The final problem involved in accessibility is the electromechanical and software aspects of the system involved. Compatibility and capability of the varying systems used to collect and analyze the data are potential problems since many existing data sets have massive dimensions by social science standards.

CHAMPUS—AN EXAMPLE

The Civilian Health and Medical Program of the Uniformed Services (CHAMPUS) is a reimbursement system for health care services for both dependents of military personnel and those retirees who do not yet qualify for Medicare. Our evaluation experience with the CHAMPUS program focused on efforts to contain inappropriate utilization of mental health services through concurrent peer review. CHAMPUS has extensive mental health coverage and provides for both inpatient services with minimal copayment and almost unlimited outpatient services. Because of this extensive coverage, CHAMPUS has been offered by some as a model or prototype for the inclusion of mental health services for all carriers including any proposed national health insurance. However, increasing costs and reports of abuse raised concerns about such extensive coverage.

The CHAMPUS data system is enlightening for two reasons. First, it provides specific examples of the types of problems encountered with utilization of existing data systems for research. Second, these problems raise concerns about policy recommendations based on descriptive information from this system which has appeared in recent literature, particularly those about mental health benefits (e.g., Dorken, 1976; 1977; 1980).

Concurrent Peer Review Project

The CHAMPUS example is based on an evaluation of two concurrent peer review demonstration projects. Though the focus was on beneficiaries diagnosed as schizophrenic, the comparison group covered all mental health diagnoses (Note 1).

The experience with the CHAMPUS data base was broad-based since it cut across all mental health diagnoses, inpatient and outpatient services and four different locations in three states over a five-year period (FY 1974–1978).

The initial data request made to CHAMPUS was for all claims, physical and mental health, for all those who received psychiatric care during the period of study. The result was 11 computer tapes with some 1.8 million claims. Clearly, an initial accessibility problem was the size of the data set and the possibility of consuming large amounts of resources just to manipulate it.

The problems encountered were not a result of poor cooperation by CHAMPUS staff. Both data processing and managerial personnel were extremely helpful by providing information about the data system and CHAMPUS procedures. They facilitated access to other sources of information, offered useful suggestions and provided validation for many of our observations.

Data Set Problems

Several technical difficulties were experienced with the data set. Foremost, the system was designed as an accounting system for reimbursement of insurance claims. This raised concerns about the appropriateness of using this system for an evaluation because of both data format problems and also using a system for purposes other than those for which it was designed. Second, the CHAMPUS data set was a compilation of many sets of similar information originating from several sources often using varying coding schemes. Variations arose because CHAMPUS used a system of insurance companies as fiscal intermediaries (FI) and did not reimburse claims directly. Rather, claims were forwarded to the FI by the provider; the FI codes, keys, processes, and forwards the data to CHAMPUS in periodic batches. Because of the number of subsystems involved in the data collection process over the period of study, the accuracy of the data was a major concern. Third, the system of making adjustment entries was in transition during the period of study which also raised concerns about the accuracy of the data over time. Finally, the size of the data set presented some accessibility problems. Resource constraints required careful planning to avoid depleting the entire computer budget by just extracting the necessary data.

Faculty Claims Data. Evaluation of the effects of peer review focused on utilization and reimbursement data. Even such straightforward variables presented difficulties. The first problem was the determination of what claims to include in the analysis (i.e., distinguishing original claims from adjusting entries for those claims). Because the system was designed for accounting purposes, an auditing trail was required. Initially, any adjustments for a claim were entered without deleting the original claim, but CHAMPUS changed the method of entering adjustments midway through the study period.

Visual inspection of the data suggested adjustments were not always clearly identified. This lack of iden-

tification required a set of decision rules to exclude those claims which appeared to be adjustments only or to include only the relevant portion of the adjustment. Failing to examine the raw information visually or only having aggregated data would have resulted in overlooking this problem and would have resulted in double or possibly triple counting of values for an episode thereby inflating both reimbursement and utilization values.

Unit of Analysis. Another major acceptability problem was an incompatibility between the data collected by CHAMPUS and the unit of analysis desired for evaluation. Inpatient admissions was the desired unit for evaluation, but as a reimbursement system, CHAMPUS works with individual claims. The number of claims involved for a given admission varied with the length of time the patient was in the hospital and the billing procedures used by the facility. Again, it was visual inspection of the raw claim data which revealed the difficulties entailed in creating an admissions file. For most claims, the day portion of the date was missing, thereby precluding an exact determination if any two claims were contiguous (i.e., for the same admission). Two other variables in the data set could be used to define an admission but in many cases the two contradicted each other.

Once identified, the resolution of the admission file dilemma came from information external to the data set. This information was obtained during week-long visits to the demonstration project sites included as part of the evaluation procedure. The visit to one location uncovered utilization data collected independent of the CHAMPUS system. Though this manual system was insufficient for the evaluation, it did serve as a criterion to assess the relative accuracy of the two possible methods of generating admissions data. Use of one variable clearly minimized discrepancies between the local information and the CHAMPUS data set, though differences remained. Without careful scrutiny of the raw data, the discrepancies in the data would not have been discovered. An arbitrary choice between the two variables to create an admissions file had a 50% chance of generating erroneous information.

Other Problems. Another difficulty resulted from the use of five FIs in the CHAMPUS data subset. There were different FIs across the provider locations and also changes in FIs over time. These variations created difficulties in identifying patients and facilities involved in the review since each FI used its own coding scheme for certain items. For example, inspection of the raw data revealed one FI used a non-numeric coding scheme for age. Failure to detect this coding would have biased the sample of patients included. Also, knowledge of the number of eligible beneficiaries in each location would have been useful, but CHAMPUS did not have this information. The number of military personnel at each site had to be used as a surrogate measure.

Both scrutiny of the raw data and site visits were crucial in uncovering data set problems. In addition to the discovery of a validation data source, the site visits yielded invaluable information about the various procedures used by the hospitals for filing claims, the methods used by the FIs for processing claims and most important, a detailed view of the scope, purpose and functioning of the demonstration projects.

CHAMPUS serves as a useful example of archival data including problems of size, control, compatibility, purpose of the data system and changes over time. Further, visual inspection of the data and site visits lead to discovery and methods of addressing problems.

STRATEGIES FOR DEALING WITH ARCHIVAL DATA

Despite limitations archival data can be used for evaluation, managerial or policy decisions. Valuable information does exist and has been either underutilized or ignored. Caution and common sense are required and several strategies are useful in coping with the limitations of archival data. Experiential hindsight can be an asset to future forays into archival data.

Assessment of the Data Source

The first commandment in working with archival data is "know thy data." A continual learning process begins at the start of an evaluation and continues until the final report. One striking feature of working with data collected by others is the occurrence of "insights" about the information. To keep last minute surprises and traps to a minimum, a reasonable understanding of the data is required before becoming committed to the project.

View of Data System. A cross-sectional view of the data system is needed early to effectively assess the appropriateness and accuracy of the data. Objective measures may not be possible at this time, but subjective impressions from a variety of sources can provide insights into the operation and acceptability of the data system. A useful first step is to obtain copies of official documents on the data system, including all forms on which the data are collected, key codes used for punching the data, definitions of variables, training manuals and other information related to the system.

Besides the official documentation describing the

system, early exposure to the data itself is highly desirable. Old printouts of aggregated data, monthly reports, for example, identify information included and may yield some insight into the utility and/or accuracy of the data. These reports can also provide an indication of the time lag involved in data processing by comparing the time of the event reported with the date of the report. Such reports may also provide the evaluator with an intuitive check on the data (i.e., do the figures make sense?); several monthly reports can form an initial cross-sectional check on validity and reliability. Similar checks can be done if other documents containing the same information collected independently are available.

Data Manager Perceptions. Besides documentation and sample data, another important variable is the perception of the personnel involved. Higher level management can provide a view of how the system is intended to work; the data processing personnel are more likely to have detailed knowledge of the actual functioning of the system and be sources of information on weaknesses with the system. Their knowledge of the foibles of the system will provide a view of both the reliability of the overall system and the trustworthiness of individual items.

Key Decision-Maker Perceptions. The initial assessment of the system should also include the perceptions of key decision-makers in the organization. If the evaluation is to result in the implementation of changes, the credibility of the data base should be tested early. Frequently, unpopular evaluation results are attacked on methodological grounds (Rossi, 1972), but a similar strategy can be criticisms of the data source. The involvement of the decision-makers at this stage reduces the likelihood of the latter kind of criticism.

Feasibility. The appropriateness, accuracy, as well as accessibility of the data are to be considered in assessing a potential data source. The simplest part of the accessibility is purely mechanical: compatibility of machines (e.g., tape density), size of the data set, the form in which the data can be released and the type of information to be included. The more difficult aspect of assessing the accessibility question is political in nature because of the potential threat of negative exposure resulting from evaluation efforts. For an outside evaluator, a long term negotiation process is required to address issues such as the purpose of the evaluation, who can release the results, and who is to pay for the data extraction process. The unique aspect in archival data is the introduction of an additional agent in the data collection process. Collecting one's own data generally implies control over the timetable. Requesting data from others places the evaluator in the provider's timetable with the data extraction often being done as an addition to the provider's regular workload. Depending on the size of the request and the workload of the facility, considerable time delays may ensue.

Working Relationship. The preliminary assessment of the data source is a first step; if the data appears acceptable and accessible, a detailed assessment of the system is advisable. Since the cooperation of a variety of people will be required, building relationships with data processing personnel becomes critical. Often these personnel may feel their efforts are not fully appreciated because of under-utilization of the data in the past and a request for data creates an additional workload for them. A good working relationship is invaluable, particularly when a major stumbling block with the data is encountered after it has been acquired. Because of their day-to-day knowledge of the system, they are the ones most likely to have the solution.

Data Acquisition

A thorough preliminary assessment improves the acquisition process. If one has to make early decisions about data and format (especially with a large data set requiring substantial data processing and long lead times), obtaining the *raw data* or major subsets is highly desirable. Advantages include continued assessment of the data and flexibility in the design and form of the evaluation. Independent sources of the same information (e.g., manual records) may be discovered after the data is in the hands of the evaluator but only cover a subset of the evaluation data source. If the information from the original data system is aggregated so a subset is not extractable, using other discovered information as a check on the data is not possible. In addition to external checks, a visual review of the raw data can be enlightening. Missing data or unexplained discontinuities over time or shifts will often raise important questions about data processing, recording methods or programmatic changes.

Data Type and Format. Size of the data set, confidentiality or other practical considerations, may require decisions about the type and form of the information extracted. The ideal situation is to obtain all information in raw form. If a decision is necessary to limit acquisition to some subset of the available variables, the two key criteria are usefulness to the evaluation effort and the accuracy of the information, both reliability and validity. A preliminary assessment of the data system may provide a tentative estimate of the accuracy of individual items. The level of abstraction involved in a particular item also may be a determinant. For instance, basic demographic information tends to be more accurate than some global measure (e.g., sex *vs.* level of functioning) but there is a trade-off because, generally the more subjective information tends to be more useful.

Timing. A final consideration in the acquisition process is timing. In addition to the lead time required for data processing personnel to honor the evaluator's request, time is required for data to be processed through the information system. Depending on the size of the system and the number of steps involved, the time lag between the event and final processing of information about the event may be a few days or several months. A data set must be complete for the time frame desired at the time of the data acquisition by the evaluator to avoid bias because of a selection artifact. Suppose, for example, length of psychiatric hospitalization was being assessed for all patients admitted to a facility during a given calendar year. Acquiring the data the following February will skew the distribution. For those admitted early in the year, discharge information will be available for almost a full year. But for those admitted in December, discharge data will only be available for those who left the hospital within 60 days of their admission; the longer length of stay for those admitted late in the year will not be included because the patients still remained in the hospital.

Site Visit/Case Studies

Up to this point, the discussion has focused on the end point of the system, the data processing division. To perform a thorough assessment of the data and to understand what significance should be placed on any results of the evaluation, an evaluator should consider the organizational level of the individual program plus all administrative and data processing levels between the individual program and the focus of the evaluation. While evaluation literature urges evaluators to understand the workings of a program being evaluated, this requirement is often overlooked, especially in large scale programs with a variety of levels involved. Site visits to all or a selected sample of programs can lead to an understanding of the perceptions of purpose, operation, scope, origins and outcome of the program at each level. Official documentation only provides a view of the head of the elephant; one has to consider the legs also (i.e., that which makes it move).

Besides a process evaluation, site visits provide an additional opportunity to assess the quality of the data. One method is to physically follow the information through the system. Insights emerge by talking to the people who filled out the forms and assessing what the information means to them. Questions may focus on unavailable data, timing problems, and importance or meaninglessness of data. This simple-minded approach of following the form through each step of the system will often provide more insight into the accuracy and meaning of the data than sophisticated and expensive reliability and validity studies. Independent auditors (such as Certified Public Accountants, CPA's) often use the foregoing approach in evaluating the internal controls operating to insure the accuracy and completeness of information produced by an information system. An inexpensive and quick check on the reliability of the system and its time lag is to feed several test cases into the system and then monitor the speed and accuracy of the output of those cases. The test-case approach and tracing single transactions throughout the entire system are popular techniques with independent auditors as well.

CONCLUSIONS

The use of archival data can be important to program evaluators and policy analysts. Through an empirical example potential problems are outlined, including safeguarding the confidentiality of the data and the appropriateness, accuracy, and accessibility of the data. Strategies based on a full knowledge of the program and its data system are presented as possible ways of addressing these problems.

Final evaluation reports should document the problems encountered and strategies used to cope with them. Explicit statement should be made if, for example, specific variables were *not* used in the evaluation because of concerns about their accuracy. For the variables used, similar statements should be offered if there is either objective information or subjective impressions about their relative accuracy. These statements will allow the report reader to weight various results and conclusions appropriately. A detailed assessment of the acceptability and accuracy of various pieces of information may also serve as an impetus for the improvement of the data system facilitating future evaluation efforts.

REFERENCE NOTE

1. SORENSEN, J. E., ZELMAN, W. N., BROUGHTON, A., CLOW, H. K., LUCKEY, J. W., MEILE, R. L., & YOUNG, E. H. *CHAMPUS experience with concurrent peer review: Case studies, utilization and cost-effectiveness analysis.* Department of Defense Contract #MDA906-80-C-0003. National Institute of Mental Health Contract #278-78-0078 (OD). March, 1980.

REFERENCES

APSLER, R. In defense of the experimental paradigm as a tool for evaluation research. *Evaluation,* 1977, *4,* 1-18.

BORUCH, R. On common contentions about randomized field experiments. In G. Glass, (Ed.). *Evaluation studies review annual: Volume 1.* Beverly Hills: Sage Publications, 1976.

CAMPBELL, D. T., & STANLEY, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally, 1963.

CHAPMAN, R. L. *The design of management information systems for mental health organizations: A primer.* (DHEW Publication No. ADM 76-333) Washington, D.C.: U.S. Government Printing Office, 1976.

COOK, T. D., & CAMPBELL, D. T. *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally, 1979.

DORKEN, H. CHAMPUS ten-state claim experience for mental disorder. In H. Dorken and Associates, *The professional psychologist today: New developments in law, health insurance and health practice.* San Francisco: Jossey-Bass, 1976.

DORKEN, H. CHAMPUS ten-state claim experience for mental disorder: Fiscal year 1975. *American Psychologist,* 1977, *32,* 697-710.

DORKEN, H. Mental health services to children and adolescents under CHAMPUS: Fiscal year 1975. *Professional Psychology,* 1980, *11,* 12-14.

DOWNS, A. Some thoughts on giving people economic advice. In F. Caro, (Ed.). *Readings in evaluation research.* New York: Russell Sage Foundation, 1971.

LUCKEY, J. W., & BERMAN, J. J. Effects of a new commitment law on involuntary admissions and service utilization patterns. *Law and Human Behavior,* 1979, *3,* 149-161.

ROSSI, P. H. Booby traps and pitfalls in the evaluation of social action programs. In C. Weiss, (Ed.). *Evaluating action programs: Readings in social action and education.* Boston: Allyn and Bacon, Inc., 1972.

SORENSEN, J. E., & ELPERS, J. R. Developing information systems for human service organizations. In C. C. Attkisson, W. A. Hargreaves, M. J. Horowitz, and J. E. Sorensen (Eds.), *Evaluation of Human Service Programs.* New York: Academic Press, 1978.

WEBB, E. J., CAMPBELL, D. T., SCHWARTZ, R. D. & SECHREST, L. *Unobtrusive measures.* Skokie, Ill.: Rand McNally, 1966.

WEISS, C. *Evaluation research.* New York: Prentice Hall, 1974.

14

Cost-Effectiveness
A Review

Paul M. Wortman

Evaluation researchers, especially psychologists, have been largely concerned with determining the effectiveness of innovative social programs. The last decade has witnessed tremendous progress in developing research methods for accomplishing this goal, as the preceding discussion has indicated. However, as the recent debate over federal policy for mental health services has indicated (Kiesler 1980, 1982, Saxe 1982), the demonstration of program effectiveness is not sufficient to affect decision and other policymakers. The next step after the evaluation of program effectiveness is the determination of whether the programs are efficient in their use of scarce economic resources (McKinlay 1981). The methods involved in making this determination are called cost-benefit and cost-effectiveness analysis. Although an earlier *Annual Review of Psychology* chapter (Perloff et al 1976) contained a brief section on "benefit-cost analysis," the methods have undergone such rapid development in both conception and importance as to warrant a second, more detailed, discussion.

In a seminal paper appearing in the first major treatise on evaluation research, Levin (1975) described the rationale and general approach in conducting a cost analysis. Levin noted that the evaluator's excitement in discovering significant effects for a program may be "misleading" since it ignores cost considerations. Levin demonstrated the cogency of this comment by an actual example drawn from the evaluation of an innovative educational program using computer-aided instruction (CAI). While the CAI program was able to produce mathematics gains for elementary school children, it did so at a cost of about $150 per year. Using data from the control group, Levin found that the same gains could be achieved by adding approximately a half hour of additional instruction at a cost of only $35 per year. Thus, the statistically significant findings demonstrating the effectiveness of the CAI program cost four times as much as a simple alternative involving only a small increase in teaching.

Editors' Note: The following article is one section from a longer paper that also has major reviews of the topics of social experimentation and meta-analysis.

From Paul M. Wortman, "Evaluation Research: A Methodological Perspective," pp. 246–260 in *Annual Review of Psychology*, Vol. 34, edited by Mark R. Rosenzweig and Lyman W. Porter.

Definitions

Current practitioners view the methods of cost-effectiveness and cost-benefit analysis as being similar and complementary processes (Warner & Luce 1982). For this reason, recent reports (Office of Technology Assessment 1980) have referred to them by their initials as CEA/CBA, respectively. CEA/CBA is viewed as a dicision-making tool for allocating public funds to programs that are more efficient.

The terms CBA and CEA are defined as: "formal analytical techniques for comparing the positive and negative consequences of alternative ways to allocate resources" (OTA 1980). The results of research studies and other applied findings are often used in conjunction with elaborate mathematical and other models to determine or compare the costs and benefits of the program under consideration. A wide variety of techniques can be used in conducting a CBA and CEA.

The principal difference between these two approaches lies in the "valuation" of benefits. In CBA all benefits as well as all costs are valued in monetary terms. This permits the decision maker to compare projects or programs of different kinds (such as mental hospitals with drug addiction treatment programs). On the other hand, CEA does not value benefits in terms of monetary value but measures them in some other unit (such as quality of life-years or years of life gained). As a consequence, CEA does not result in a net monetary value (that is, of benefits minus costs), but instead the amount of dollars or costs involved in achieving some desired effect. Therefore, CEA only allows a decision maker to compare programs that have similar objectives.

Principles

Despite this difference, the methods are considered as generally similar in both "concept and purpose." CEA is now viewed as an important adjunct to CBA. Moreover, analysts have become more sophisticated and this has allowed them to combine the two. For example, monetary benefits are also included in a CEA by transforming them into negative costs. This allows one to compute "the *net* cost per unit of effectiveness." While there is no standard method, OTA (1980) did find "general agreement on a set of 10 principles of analysis that could be used to guide the conduct, evaluation, or use of CEA/CBA studies." These principles, listed in Table 2, are consistent with the description of other authors (Thompson 1980, Warner & Luce 1982) and are discussed briefly below.

Table 2 Ten general principles of analysis (for CEA/CBA methodology)[a]

1. Define problem	6. Differentiate perspective of analysis
2. State objectives	7. Perform discounting
3. Identify alternatives	8. Analyze uncertainties
4. Analyze benefits/effects	9. Address ethical issues
5. Analyze costs	10. Interpret results

[a] Reprinted with permission from the Office of Technology Assessment (1980).

PROBLEM DEFINITION The first step in the conduct of a cost analysis (#1) involves the specification of the problem including its scope. For example, if one is concerned with the anguish and suffering of those who are mentally ill, then it is appropriate to consider alternative forms of treatment such as the most effective psychotherapeutic procedures. On the other hand, if one takes a broader perspective that includes future mental illness, then prevention becomes a relevant concern and a larger set of alternatives should be considered.

Evaluation research, however, has been program, rather than problem, focused. The perspective is admittedly a narrow and limited one. In these situations, evaluators have been primarily concerned with either program operation (or process) or program outcome (or impact). A CEA/CBA would focus on whether these outcomes are achieved "efficiently." While most CEA/CBAs have been, in fact, program or technology focused in the past few years, they run the risk, according to Warner & Luce (1982), of "missing the forest for the trees." That is, analytical rigor may be achieved at the price of ignoring the broader perspective involving a comparison with alternative programs. However, in evaluation research this is often not possible for there are seldom alternative programs available for comparison except in a prospective way. While CEA/CBA is appropriate for analyzing prospective programs, it places a greater burden on the analyst in estimating future costs and outcomes.

OBJECTIVES Once the problem has been specified, the next task (#2) is to define concrete objectives. This also poses many problems. The major one concerns measurement. It may simply be impossible to measure or even specify the objectives that are of major interest. For example, psychotherapy may have as its major objective diminished mental anguish or, alternatively, improved mental health. Thus, the alleviation of pain, suffering, and incapacity may have many associated variables that are difficult, if not impossible, to quantify. A useful outcome measure might be an increase in the number of additional high quality years of life, but as yet there are few

measures available to provide quality of life indices for mental illness prevention and treatment.

More commonly there will be multiple objectives and these should be noted. It is important that they represent the most important dimensions of the problem and capture the impacts of the program. This should include "non-measurable goals and outcomes." Currently, CEA/CBA analysts tend to ignore such difficult problems and focus instead on the quantifiable objectives. This means that the difficult problems are not getting appropriate consideration. On the other hand, it may not be worthwhile to invest considerable resources in quantifying such objectives. In some cases it may be possible to achieve a meaningful analysis using secondary measures. For example, the cost of maintaining mental patients in noninstitutionalized settings may be cheaper than the old mental hospital, and it may be agreed that the quality of patients' lives is definitely as great, if not greater, than when they were institutionalized. In this case the deinstitutionalization program would be seen as cost-effective.

IDENTIFICATION OF ALTERNATIVES When considering this issue (#3), analysts typically are thinking of "explicit programs with budgets, organization, inputs, and outputs." However, as noted above, in evaluation there are often no alternative programs to be considered. One also should keep in mind "nonprogrammatic" alternatives that may not be amenable to CEA/CBA since their costs and benefits may be difficult to assess. For broadly defined problems, the analyst may be forced to reduce the number of alternatives considered. In this case, he or she should select a variety of programs that are considered to be "potentially cost-effective." If possible, these should be representative of the various possible alternatives. When choosing among similar alternatives, one should select the program that is most effective.

MEASUREMENT AND ANALYSIS OF BENEFITS AND COSTS The next steps (#4 and 5) involve the description of what is called a "production function." The production function is used to relate economic resources or "inputs" to benefits or outcomes (sometimes called "outputs"). This is often a mathematical procedure involving techniques such as linear programing, Markov processes, and computer simulations. For example, in perhaps the first reported cost-benefit analysis (Guess 1981), Bernoulli in 1760 employed a differential equations model to determine the effectiveness of smallpox vaccine on mortality. More recently, Albritton (1978) performed an interrupted time-series analysis to estimate the impact of a federally sponsored measles vaccination program (i.e. the "number of cases averted").

Schoenbaum et al (1976) used the Delphi technique to obtain several estimates needed to determine the likely effectiveness of the swine flu vaccine program on various subgroups of the population. According to Warner & Luce (1982), this is "one of the most technically challenging aspects of analysis." In many cases an analyst can borrow from comparable programs in other locations to determine the relationships. A variety of external validity factors should be considered, including the scale of the program, whether the technology or program has changed, whether the program being used as a model is really efficient, and whether it has some unique resources that cannot be duplicated in the program under consideration.

Once this is accomplished the next step is the "identification, measurement, and evaluation of the costs and benefits associated with the production process" (Warner & Luce 1982). As noted above, in CEA cost savings are included not as a benefit but as a negative cost. This allows one to examine net cost changes compared to "all net health benefit changes measured in some non-monetary units." There are both direct and "indirect" costs that must be taken into account in calculating the CBA/CEA. The direct costs refer to actual resources used in producing outcomes while indirect costs involve resources not directly affecting the outcomes; for example, the value of a patient's time that may be used in other productive activities rather than waiting for services to be delivered. These so-called "opportunity costs" should follow directly from the specification of the production function.

Once the resources involved have been identified, their cost or "valuation" must be determined. This, too, is generally a straightforward procedure when one uses current wage rates and other monetary expenditures. However, caution is recommended in using the actual, rather than the billed, charges for costs, for in many cases these charges are adjusted to subsidize other unprofitable activities in the organization.

A similar process occurs on the benefits or effectiveness side involving the outputs of the process. First, one identifies the potential objectives or the benefits that one wants to attain and then assigns a dollar value to them if one is dealing in a cost-benefit approach. Otherwise, for CEA the process ends at this identification step. The benefits involve personal, institutional, and societal outcomes. For example, this could include reduced days of illness, decreased days in institutions, and increased work as a result of better health. Also one has to take into account the "intangible," "nonquantitative" benefits such as reduced pain and suffering.

According to Warner & Luce (1982), one of the major problems for CBA in the areas of health and mental illness concerns the assessment of benefits involving "the value of human life"—specifically such benefits as "the avoidance of premature mortality or unnecessary morbidity." The tradi-

tional way of handling this in CBA has been to use a procedure called the "human capital approach." Briefly, this involves using labor market values such as work productivity or earnings as the measure for such benefits. This approach has been criticized for being too restrictive in omitting values of personal enjoyment and quality of life above and beyond the economic rewards of work. Moreover, wage measures can be biased in that white males are likely to be given a higher value than females or minorities. Despite these criticisms and recent developments using other procedures, the human capital approach still is the most widely used method for CBA.

CEA has provided a useful alternative procedure to dealing with this problem. It can include the "nonmarket value of life," using such quantitative measures as "deaths averted" or "life-years saved." However, CEA does not entirely avoid the issue of the value of human life; for once money is allocated to a program, it is possible to derive a minimum or maximum value of life. Thus Warner & Luce (1982) observe that there is not as large a "conceptual difference" between CBA and CEA as may appear. The former explicitly places values on life while the latter allows decision makers to accomplish this implicitly and perhaps avoids the ethical problems associated with valuing life (Kelman 1981).

Weinstein & Stason (1977) have recommended the use of "quality-adjusted life years" as a more sensitive measure of effectiveness than the more traditional improvements in mortality and morbidity. Mosteller (1981) concurs that "we need to assess the quality of life" since many treatment innovations have this as their major goal. One approach to deriving such measures involves the use of health-status indexes (Stewart et al 1981). The index acts as a "weighting scheme" to adjust changes in mortality (i.e. increased life expectancy) for the quality of life during that time.

PERSPECTIVE Most CEA/CBAs in program evaluation would be conducted from a governmental perspective. It is important to note this (#6) because perspective effects the resources considered in the analysis. If the program were run by a charity, certain costs, reimbursed or exempted by state and federal governments, may not have to be considered, for example. Fischhoff (1977) also warns that the perspective of the analyst could affect the fairness of the results. In particular, care must be taken to ensure that a societal perspective is not subverted by a special interest.

DISCOUNTING Since all costs and benefits do not occur at the same time, the analyst must devise a method of converting them into a common metric (#7). This procedure is called "discounting" and is based on the assumption that "current and future dollars are not the same." In fact, a preference for present dollars over future dollars is presumed, since they can be in-

vested to yield a profit. The procedures for discounting are straightforward. However, the results of a CEA/CBA may be sensitive to the actual discount rate used.

Warner & Luce (1982) note that while there is agreement that discounting should be employed, there is "less consensus" on the exact rate. The importance of small changes in the rate should not be ignored or underestimated. The authors provide an example where one program that is superior (in net benefits) to another at 0% and 2% discount rates, becomes inferior at a 4% rate. Such reversals are likely when large benefits are delayed many years, as they often are in screening and other preventive programs (e.g. Head Start).

SENSITIVITY ANALYSIS Since the choice of a discount rate and other assumptions made in CEA/CBA make the results somewhat uncertain and open to criticism, analysts have developed a technique known as "sensitivity analysis" to test the robustness of their findings (#8). The procedure, in brief, allows the analyst to determine whether different assumptions yield different results. Thus, for example, do changes in the value of human life, different personnel configurations, costs of tests, or the discount rate used to assess the value of future benefits and costs affect the result? One approach is to perform a so-called "worst case" analysis, employing values for those variables that will most bias the results against the findings obtained.

Another approach is known as "break even analysis." In this procedure the analyst compares alternatives involving nonmonetary measures of effect. For example, a program that costs a million dollars could be compared to another program that costs only $900,000 but saves 10 lives. In this case a life would be have to be valued at $10,000 to make the two programs equal in cost. Since most would agree that that is a relatively low value for human life, the second program would be viewed as preferable. Thus, sensitivity analyses allow the investigator to determine whether the results are dependent upon a particular assumption, for what range of values for a particular variable the program is cost effective, and also to identify issues needing further research.

ETHICAL ISSUES It is important that ethical issues be included, or at least noted, in the analysis (#9). These include not only problems in valuing human life, but issues concerning the potential harmful side effects of a program, the deprivation in withholding services, and "equity" issues in providing services. The OTA report (1980) notes that CEA is efficiency oriented and thus may not take such ethical issues as equity into account. It is possible, for example, that cost-effective decisions could systematically favor the well-to-do. The OTA report candidly acknowledges that

"CEA/CBA is weak in the areas of equity and other ethical considerations."

This is essentially an external validity issue, and Levin (1975) has made a cogent recommendation for handling it. He suggests that the analysis be broken down by subpopulation. If equity is a relevant concern, then different weights can be used so that, for example, the effects for low-income people are counted more heavily than those for the more affluent. This would yield an "effectiveness index."

RESULTS The final step in the process (#10) is the presentation and interpretation of findings. The analyst should indicate the key variables and assumptions, including the limitations of the analysis and sensitivity analyses if they have been performed. In particular, Warner & Luce (1982) note that a "very popular misconception" is that a cost-benefit ratio is an adequate measure or index of a program's performance. In fact, they recommend that a CBA should report net benefits, not the ratio, since the latter can sometimes be misleading. They cite as an example a program that produces $4000 worth of benefits from $2000 in costs, as opposed to a program that produces $3 million in benefits for $2 million dollars in cost. In the former case the ratio is 2 and in the latter 1.5. However, the former program only produces $2000 in net benefits as opposed to $1 million in the latter. Assuming that resources were available and that programs cannot be expanded or contracted in the same cycle, then the latter program is clearly preferable to the former even though it has a smaller benefit-cost ratio. Moreover, they note that these ratios are also sensitive to whether economic benefits are treated as negative costs. On the other hand, CEA, lacking an economic measure of benefits, does report a cost-effectiveness ratio as an index of program performance. However, the authors warn that this measure should not be accepted uncritically for often one can save more lives with an increase in cost that is not considered inappropriate.

In some cases it may not be possible to report the analysis in terms of a single measure of effectiveness or dollars for either CEA or CBA. In such cases analysts recommend that multiple outcomes be presented. For example, such measures of effectiveness as "quality-adjusted life years," morbidity days saved, and disability days may represent a set of measures that cannot be combined. When alternative programs differ in both cost and benefits, it is impossible to arrive at an objective ranking. However, when the costs are the same, some comparisons are possible, particularly in those cases where one alternative clearly dominates another; that is, it has larger or smaller benefits than another alternative. In some cases, the analyst will be required to use subjective judgments to determine whether slight differ-

ences in benefits are superior and thus allow one program to be ranked higher than another.

Levin (1975) notes that there are at least three measures of CEA: "total cost for achieving a certain level of effectiveness; average cost per unit of effectiveness; and marginal cost for additional units of effectiveness." Total costs are appropriate when equal levels of effectiveness are achieved. When this is not the case, average cost per unit effectiveness, as noted earlier, is appropriate. There is, however, one serious problem with this measure. It does not take into account the scale of the program and thus assumes that these results can be extrapolated, an assumption that may or may not be true. For example, if a program has a large amount of costs that are fixed, it will show a high average cost when, in fact, one could enlarge it at small costs and get more benefits. In these cases, the marginal costs are the appropriate measures to be used. For program evaluators, this is calculated by subtracting the effects for the control group from the results of the experimental programs and then dividing that number (e.g. number of people returning to work, number of people cured by therapy, etc) into the additional costs in mounting these alternative programs (obtained by subtracting the cost of the control program from these alternative programs).

The Potential Application of CEA/CBA to Psychotherapy

Saxe (1980) maintains that CEA/CBA can be "potentially" useful to decision makers in allocating scarce fiscal resources for mental health services. He illustrates how this approach could be applied to assessing psychotherapy. The discussion reveals the difficulty of using these methods. Major obstacles for CEA/CBA occur in identifying and measuring benefits and finding useful empirical estimates. As Boruch (1982) has recently noted: "The absence of formal cost-benefit analyses of evaluations, including experiments, is remarkable. Part of the problem lies in defining benefits." These problems are briefly reviewed in the remainder of this section.

BENEFITS According to Saxe, "The unique problems of CEA/CBA of psychotherapy have to do with the difficulty of comprehensively assessing and valuing the effects of psychotherapy" (e.g. "reduction of pain and suffering," "improved well-being"). He notes that the identification process, itself, is problematic. It is often difficult to determine which effects or benefits are attributable to psychotherapy. This requires having good research data, but such information is often "inadequate." This is due to either poor research designs or the absence of appropriate measures (see below). Similar problems beset the assignment of benefits to those associated with patients, such as families, friends, employers, and even society. Generally either "willingness to pay" or earnings are used as proxies for such

benefits as reduced unemployment payments and increased productivity. However, the former is somewhat controversial since sick patients may not be able to make such cost decisions (Kelman 1981). Benefits, thus, are the most problematic component in the analysis. One may either not fully identify all benefits that should be considered or underestimate their value when they are properly identified.

EMPIRICAL EVIDENCE Saxe briefly reviews the few empirical studies dealing with the impact of psychotherapy on the utilization of medical services. Nine studies (including one review of 25 studies) are referenced and briefly discussed; all show that psychotherapy lowers the utilization of medical services. This has led Kiesler (1980) to claim that, "We know that adding the option of psychological services to an existing physical health care system is cost-effective and reduces the use of the physical health system." While this evidence is provocative, it is apparently not as conclusive as Kiesler's remark indicates.

There are a number of serious methodological problems that characterize these studies that necessarily add caution to their interpretation. The most critical are those affecting internal validity. As Saxe notes, many of the studies did not include "appropriate control group conditions." In others, patients who were high utilizers of services were studied, introducing the possibility of statistical regression artifacts. Another major problem stems from construct validity concerns. The timing of the psychotherapeutic intervention usually occurred after an initial period of high utilization and thus is potentially confounded with the normal temporal pattern of service utilization. As is customary with such reviews, it concludes by calling for "additional research."

Utility of CEA/CBA

Despite the problems noted above, Saxe (1980) concludes that psychotherapy is "scientifically assessable" using CEA/CBA in combination with other rigorous research methods. In this he is in agreement with Kiesler's (1980) analysis of the research needed to provide a strong mental health policy analysis. However, this does not mean that definitive CEA/CBA studies can be conducted that will be the sole basis for the policy. At present, the OTA report (1980) notes that most CEA/CBA studies are "academic exercises" of little policy relevance.

The OTA report concluded that CEA/CBA has "too many methodological and other limitations" to be used as the only evidence for decision-making. Mosteller (1981) has also observed that health-care goals often conflict. For example, in minimizing lives lost and days hospitalized for an appendectomy, many unnecessary surgeries are required for the former and

some deaths in the latter. Mosteller notes that this is an instance of "the classical mathematical dilemma that we cannot expect to maximize two functions at the same time." For such reasons OTA recommended that the CEA/CBA techniques be used to "structure" the problem, to obtain the information, and then to present or "array" the data elements that would be included in making a decision. Such caution has also been urged by others (cf Roid 1982) including the American Public Health Association (1981).

In conclusion, knowledge and use of CEA/CBA may be essential for the very survival of evaluation research. The decade of the 1970s witnessed the rapid development of methods for assessing the effectiveness of innovative social programs. This was an area that psychologists felt most comfortable with since it required only moderate extensions and refinements of skills basic to the profession. Issues concerning costs and policy analysis have been foreign to the training and experience of psychologists/evaluators. Given the recent fiscal and political climate, cost considerations may be essential in the current decade for appropriate and useful evaluation. Such analyses represent a logical step in the development of methods for evaluation research. In some quarters the debate on establishing formal organizations to gather cost-effectiveness data has already begun (Relman 1982, Bunker et al 1982).

CONCLUSION

Evaluation research is a multidisciplinary activity that is united by its concern for sound methods that can be used to obtain valid information. It is appropriate to consider the role of psychology in this young field. Psychologists have played a central role in the development and creative adaptation of rigorous experimental and quasi-experimental designs for evaluating innovative social programs. They have made their evaluation colleagues aware of the problems posed by the improper conduct and analysis of evaluative studies. In doing so, they have provided a critical logic along with the statistical skills needed to refine other methods of potential use in evaluating programs. Despite these accomplishments, there remains important work to be done. The repertoire of methods should be expanded to include more qualitative methods that can illuminate and corroborate the statistical findings. The search for better outcome measures of program benefits focusing on quality-of-life and other, primarily psychological, outcomes must continue. The initial development of methods for synthesizing the expanding varieties of information into a coherent assessment must be completed. Finally, psychologists must learn that they cannot do it all

themselves. There must be a greater spirit of social science ecumenism in their evaluation approaches and staffs.

ACKNOWLEDGMENTS

The author thanks Fred Bryant, Judy Garrard, Leonard Saxe, William Shadish, and Bill Yeaton for their careful reading and thoughtful comments on earlier drafts of this chapter. Thanks are also in order for Rita Page for her conscientious and patient typing of the many drafts and revisions. The work on this chapter was supported by grants from the National Institute of Education (NIE-G-79-0128) and the National Center for Health Services Research (HS 04849-01).

Literature Cited

Albritton, R. B. 1978. Cost-benefit of measles eradication: Effects of a federal intervention. *Policy Anal.* 4:1–22

Alkin, M. C., Daillak, R., White, P. 1979. *Using Evaluations: Does Evaluation Make a Difference?* Beverly Hills/London: Sage. 269 pp.

American Public Health Association. 1981. Use of cost-benefit analysis in public health regulation. *Nation's Health* 11(9):3

Baum, M. L., Anish, D. S., Chalmers, T. C., Sacks, H. S., Smith, H., Fagerstrom, R. M. 1981. A survey of clinical trials of antibiotic prophylaxis in colon surgery: Evidence against further use of no-treatment controls. *N. Engl. J. Med* 305:795–99

Berk, A. A., Chalmers, T. C. 1981. Cost and efficacy of the substitution of ambulatory for inpatient care. *N. Engl. J. Med.* 304:393–97

Boruch, R. F. 1982. Experimental tests in education: Recommendations from the Holtzman report. *Am. Stat.* 36:1–8

Boruch, R. F., Cordray, D. S., Pion, G., Leviton, L. 1981a. A mandated appraisal of evaluation practices: Digest of recommendations to the Congress and the Department of Education. *Educ. Res.* 10:10–13, 31

Boruch, R. F., Gomez, H. 1977. Sensitivity, bias, and theory in impact evaluations. *Prof. Psychol.* 8:411–34

Boruch, R. F., McSweeny, A. J., Soderstrom, E. J. 1978. Randomized field experiments for program planning, development, and evaluation: An illustrative bibliography. *Eval. Q.* 2:655–95

Boruch, R. F., Wortman, P. M. 1979. Implications of educational evaluation for evaluation policy. In *Review of Research in Education,* ed. D. C. Berliner, 7:309–61. Washington DC: Am. Educ. Res. Assoc.

Boruch, R. F., Wortman, P. M., Cordray, D. S., et al. 1981b. *Reanalyzing Program Evaluations.* San Francisco: Jossey-Bass. 403 pp.

Bunker, J. P., Fowles, J., Schaffarzick, R. 1982. Evaluation of medical-technology strategies: Part II. Proposal for an Institute for Health-Care Evaluation. *N. Engl. J. Med.* 306:687–92

Burger, J. M. 1981. Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. *Psychol. Bull.* 90:496–512

Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L. et al. 1976. Randomized clinical trials: Perspective on some recent ideas. *N. Engl. J. Med.* 295:74–80

Campbell, D. T. 1979a. Assessing the impact of planned social change. *Eval. Program Plann.* 2:67–90

Campbell, D. T. 1979b. "Degrees of freedom" and the case study. In *Qualitative and Quantitative Methods in Evaluation Research,* ed. T. D. Cook, C. S. Reichardt, 1:49–67. Beverly Hills/London: Sage (Res. Prog. Ser. Eval.). 160 pp.

Campbell, D. T. 1976. Focal local indicators for social program evaluation. *Soc. Ind. Res.* 3:237–56

Campbell, D. T., Erlebacher, A. 1970. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In *Compensatory Education: A National Debate,* ed. J. Hellmuth, 3:185–210. New York: Bruner/Mazel. 225 pp.

Campbell, D. T., Stanley, J. C. 1966. *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally. 84 pp.

Chalmers, T. C. 1981. The clinical trial. *Milbank Mem. Fund Q.* 59:324–39

Chalmers, T. C., Block, J. B., Lee, S. 1972. Controlled studies in clinical cancer research. *N. Engl. J. Med.* 287:75–78

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., et al. 1981. A method for assessing the quality of a randomized control trial. *Controlled Clin. Trials* 2:31–49

Cohen, J. 1962. The statistical power of abnormal-social psychological research: A review. *J. Abnorm. Soc. Psychol.* 65:145–153

Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic. 474 pp. Rev. ed.

Comroe, J. H. Jr. 1978. The road from research to new diagnosis and therapy. *Science* 200:931–37

Cook, T. D., Campbell, D. T. 1979. *Quasi-experimentation: Design & Analysis Issues for Field Settings.* Chicago: Rand McNally. 405 pp.

Cooper, H. M. 1979. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *J. Pers. Soc. Psychol.* 37:131–46

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., et al. 1980. *Toward Reform of Program Evaluation: Aims, Methods, and Institutional Arrangements.* San Francisco: Jossey-Bass. 438 pp.

Davis, B. G., ed. 1982. Boruch and McLaughlin debate. *Eval. News* 3:11–20

DeSilva, R. A., Hennekens, C. H., Lown, B., Casscells, W. 1981. Lignocaine prophylaxis in acute myocardial infarction: An evaluation of randomized trials. *Lancet* ii:855–58

Elashoff, J. D. 1978. Combining results of clinical trials. *Gastroenterology* 75:1170–72

Eysenck, H. J. 1978. An exercise in megasilliness. *Am. Psychol.* 33:517

Fischhoff, B. 1977. Cost benefit analysis and the art of motorcycle maintenance. *Policy Sci.* 8:177–202

Fletcher, R. H., Fletcher, S. W. 1979. Clinical research in general medical journals: A 30-year perspective. *N. Engl. J. Med.* 301:180–83

Gallo, P. S. Jr. 1978. Meta-analysis—A mixed meta-phor? *Am. Psychol.* 33: 515–17

Gehan, E. A., Freireich, E. J. 1974. Non-randomized controls in cancer clinical trials. *N. Engl. J. Med.* 290:198–206

Gilbert, J. P., Light, R. J., Mosteller, F. 1975. Assessing social innovations: An empirical base for policy. In *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* ed. C. A. Bennett, A. A. Lumsdaine, pp. 39–193. New York: Academic. 553 pp.

Gilbert, J. P., McPeek, B., Mosteller, F. 1977. Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In *Costs, Risks, and Benefits of Surgery,* ed. J. P. Bunker, B. A. Barnes, F. Mosteller, pp. 124–69. New York: Oxford Univ. Press. 401 pp.

Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. *Educ. Res.* 5:3–8

Glass, G. V. 1977. Integrating findings: The meta-analysis of research. In *Review of Research in Education,* ed. L. S. Shulman, 5:351–79. Itasca, Ill: Peacock. 399 pp.

Glass, G. V. 1978. Reply to Mansfield and Busse. *Educ. Res.* 7:3

Glass, G. V., Ellett, F. S. Jr. 1980. Evaluation research. *Ann. Rev. Psychol.* 31:211–28

Glass, G. V., McGaw, B., Smith, M. L. 1981. *Meta-analysis in Social Research.* Beverly Hills/London: Sage. 279 pp.

Glass, G. V., Smith, M. L. 1978. Reply to Eysenck, *Am. Psychol.* 33:517

Glass, G. V., Smith, M. L. 1979. Meta-analysis of research on class size and achievement. *Educ. Eval. Policy Anal.* 1:2–16

Guess, H. A. 1981. Bernoulli's cost-benefit analysis of smallpox immunization. *N. Engl. J. Med.* 305:347

Horwitz, R. I., Feinstein, A. R. 1981. Improved observational method for studying therapeutic efficacy. *J. Am. Med. Assoc.* 246:2455–59

House, E. R. 1980. *Evaluating with Validity.* Beverly Hills/London: Sage. 295 pp.

House, E. R., Glass, G. V., McLean, L. D., Walker, D. F. 1978. No simple answer: Critique of the Follow Through evaluation. *Harvard Educ. Rev.* 48:128–60

House, E. R., Mathison, S. 1981. Book review of *Toward Reform of Program Evaluation. Eval. News.* 2:314–20

Hunter, J. E., Schmidt, F. L., Jackson, G. 1982. *Meta-analysis: Cumulating Research Findings Across Studies.* Beverly Hills/London: Sage. In press

Jackson, G. B. 1980. Methods for integrative reviews. *Rev. Educ. Res.* 50:438–60

Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., Skon, L. 1981. Effects of

cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychol. Bull.* 89: 47–62

Judd, C. M., Kenny, D. A. 1981. *Estimating the Effects of Social Interventions.* Cambridge/London: Cambridge Univ. Press. 243 pp.

Kelman, S. 1981. Cost-benefit analysis: An ethical critique. *Regulation* Jan/Feb: 33–40

Kenny, D. A. 1979. *Correlation and Causality.* New York: Wiley. 277 pp.

Kiesler, C. A. 1980. Mental health policy as a field of inquiry for psychology. *Am. Psychol.* 35:1066–80

Kiesler, C. A. 1982. On cost effectiveness. *Am. Psychol.* 37:95–96

Knox, R. A. 1980. Heart transplants: To pay or not to pay. *Science* 209:570–75

Landman, J. T., Dawes, R. M. 1982. Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *Am. Psychol.* 37:504–16

Levin, H. M. 1975. Cost-effectiveness analysis in evaluation research. In *Handbook of Evaluation Research,* ed. M. Guttentag, E. L. Struening, 2:89–122. Beverly Hills/London: Sage. 736 pp.

Mansfield, R. S., Busse, T. V. 1977. Meta-analysis of research: A rejoinder to Glass. *Educ. Res.* 6:3

Marshall, E. 1980. Psychotherapy works, but for whom? *Science* 207:506–8

McKinlay, J. B. 1981. From "promising report" to "standard procedure": Seven stages in the career of a medical innovation. *Milbank Mem. Fund Q.* 59:374–411

McSweeny, A. J., Wortman, P. M. 1979. Two regional mental health treatment facilities: A reanalysis of evaluation of services. *Eval. Q.* 3:537–56

Mosteller, F. 1981. Innovation and evaluation. *Science* 211:881–86

Mosteller, F., Gilbert, J. P., McPeek, B. 1980. Reporting standards and research strategies for controlled trials. *Controlled Clin. Trials* 1:37–58

Nisbett, R., Ross, L. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment.* Englewood Cliffs, NJ: Prentice-Hall. 334 pp.

Office of Technology Assessment. 1978. *Assessing the Efficacy and Safety of Medical Technologies.* Washington DC: GPO (No. 052–003–00593–0). 133 pp.

Office of Technology Assessment. 1980. *The Implications of Coss-Effectiveness Analysis of Medical Technology.* Washington DC: GPO (Stock No. 052–003–00765–7). 219 pp.

Office of Technology Assessment. 1982. *Technology Transfer at the National Institutes of Health.* Washington DC: GPO. 188 pp.

Partlett, M., Hamilton, D. 1976. Evaluation as illumination: A new approach to the study of innovatory programs. In *Evaluation Studies Review Annual,* ed. G. V. Glass, 1:140–57. Beverly Hills/London: Sage. 672 pp.

Patton, M. Q. 1978. *Utilization-focused Evaluation.* Beverly Hills/London: Sage. 303 pp.

Patton, M. Q. 1980. Making methods choices. *Eval. Program Plann.* 3:219–28

Perloff, R., Perloff, E., Sussna, E. 1976. Program evaluation. *Ann. Rev. Psychol.* 27:569–94

Porter, A. C., Schmidt, W. H., Floden, R. E., Freeman, D. J. 1978. Practical significance in program evaluation. *Am. Educ. Res. J.* 15:529–39

Posavac, E. J. 1980. Evaluations of patient education programs: a meta-analysis. *Eval. Health Profs.* 3:47–62.

Posavac, E. J., Carey, R. G. 1980. *Program Evaluation: Methods and Case Studies.* Englewood Cliffs, NJ: Prentice-Hall. 350 pp.

Presby, S. 1978. Overly broad categories obscure important differences between therapies. *Am. Psychol.* 33:514–15

Reichardt, C. S., Cook, T. D. 1979. Beyond qualitative *versus* quantitative methods. See Campbell 1979b, 1:7–32

Relman, A. S. 1982. An institute for healthcare evaluation. *N. Engl. J. Med.* 306:669–70

Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K. Jr., et al, eds: 1974. *Social Experimentation: A Method for Planning and Evaluating Social Intervention.* New York: Academic. 339 pp.

Roid, G. H. 1982. Cost-effectiveness analysis in mental health policy. *Am. Psychol.* 37:94–95

Rosenthal, R. 1978. Combining results of independent studies. *Psychol. Bull.* 85: 185–93

Rossi, P. H., Freeman, H. E. 1982. *Evaluation: A Systematic Approach.* Beverly Hills/London: Sage. 352 pp. 2nd ed.

Saxe, L. 1980. *Background Paper #3: The Efficacy and Cost Effectiveness of Psychotherapy.* Washington DC: Off. Technol. Assess. (GPO Stock No. 052–003–007853–5). 93 pp.

Saxe, L. 1982. Public policy and psychotherapy: Can evaluative research play a role? *New Dir. Program Eval.* 14:73–86

Saxe, L., Fine, M. 1981. *Social Experiments.* Beverly Hills/London: Sage. 267 pp.

Schoenbaum, S. C., McNeil, B. J., Kavet, J. 1976. The swine-influenza decision. *N. Engl. J. Med.* 295:759–65

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., Yeaton, W. 1979. Some neglected problems in evaluation research: Strength and integrity of treatments. In *Evaluation Studies Review Annual,* ed. L. Sechrest, S. G. West, M. A. Phillips, R. Redner, W. Yeaton, 4:15–35. Beverly Hills/London: Sage. 766 pp.

Sechrest, L., Yeaton, W. 1981. Assessing the effectiveness of social programs: Methodological and conceptual issues. *New Dir. Program Eval.* 9:41–56

Smith, M. L., Glass, G. V. 1977. Meta-analysis of psychotherapy outcome studies. *Am. Psychol.* 32:752–60

Stewart, A. L., Ware, J. E. Jr., Brook, R. H. 1981. Advances in the measurement of functional status: Construction of aggregate indexes. *Med. Care* 19:473–88

Strube, M. J., Garcia, J. E. 1981. A meta-analytic investigation of Fiedler's model of leadership effectiveness. *Psychol. Bull.* 90:307–21

Thompson, M. S. 1980. *Benefit-Cost Analysis for Program Evaluation.* Beverly Hills/London: Sage. 310 pp.

Tukey, J. W. 1977. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 198:679–84

Tversky, A., Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–31

Wagner, J. L. 1979. Toward a research agenda on medical technology. In *Medical Technology,* ed. J. L. Wagner, pp. 1–12. Washington DC: GPO (PHS-79-3254, NCHSR Res. Proc. Ser.) 120 pp.

Warner, K. E. 1975. A "desperation-reaction" model of medical diffusion. *Health Serv. Res.* 10:369–83

Warner, K. E. 1975. Luce, B. R. 1982. *Cost-Benefit and Cost-Effectiveness Analysis in Health Care: Principles, Practice, and Potential.* Ann Arbor: Health Admin. Press. 384 pp.

Weinstein, M. C., Stason, W. B. 1977. Foundations of cost-effectiveness analysis for health and medical practices. *N. Engl. J. Med.* 296:716–21

Weiss, C. H., ed. 1977. *Using Social Research in Public Policy Making.* Lexington, Mass: Heath. 256 pp.

Williams, W. 1976. Implementation problems in federally funded programs. In *Social Program Implementation,* ed. W. Williams, R. F. Elmore, pp. 15–40. New York: Academic. 299 pp.

Wortman, P. M. 1981a. Consensus development. In *Methods for Evaluating Health Services,* ed. P. M. Wortman, 8:9–22. Beverly Hills/London: Sage (Res. Prog. Ser. Eval.) 143 pp.

Wortman, P. M. 1981b. Randomized clinical trials. See 1981a, 8:41–60

Wortman, P. M., King, C. S., Bryant, F. B. 1982. Meta-analysis of quasi-experiments: School desegregation and black achievement, Part I—Retrieval and coding. Ann Arbor: Inst. Soc. Res. (tech. rep.) 22 pp.

Wortman, P. M., Reichardt, C. S., St. Pierre, R. G. The first year of the Education Voucher Demonstration: A secondary analysis of student achievement test scores. *Eval. Q.* 2:193–214

Wortman, P. M., St. Pierre, R. G. 1977. The educational voucher demonstration. *Educ. Urban Soc.* 9:471–92

15

The Application of Social Impact Assessment to the Study of Criminal and Juvenile Justice Programs

A Case Study

Merry Morash

Because the United States criminal and uvenile justice systems are highly decen-ralized, with many functions delegated to ity and county agencies, local programs re frequently developed as innovations nd used as models for replication in ther jurisdictions. This process of local rogram development and replication is ormalized and encouraged by the Na-onal Institute of Law Enforcement, which designates selected programs as Exemplary Projects. Exemplary Projects are considered to be noteworthy innovations, and in order to encourage their widespread adoption, they are highly publicized by the National Institute.

For these Exemplary Projects, and other local programs which are to serve as models for replication, there are some special informational needs to be met through research.

rom Merry Morash, "The Application of Social Impact Assessment to the Study of Criminal and uvenile Justice Programs: A Case Study," *Journal of Criminal Justice*, 1983, 11(3), 229-240.

Those who develop the program need feedback in the early stages of operation in order to justify continued support for the project and to identify the need for modifications. Policy makers, program administrators, and concerned citizens who contemplate replicating a program need an organized body of information about one or more similar programs which are already implemented. For purposes of support, modification, and replication, this information should reveal a wide spectrum of the program's probable impacts on individuals, organizations, and the community. Furthermore, there would be comparable information on the impacts of important alternative programs, or the option of no program at all.

Unfortunately, the research strategies commonly used to study newly developing programs and programs which are to be replicated, including those designated as Exemplary, do *not* produce information on a full range of impacts in any systematic way. Instead, the focus is typically quite narrow. For example, studies of most correctional programs are usually limited to the intended effects of the target group of offenders (e.g., Sechrest, White, and Brown, 1979). Although primarily theoretical work has brought some unintended impacts of correctional programs to our attention, program evaluation results are rarely in a form that lends itself to simultaneous consideration of both desired and undesired program results. Again using correctional programs as an example, an ideal comparison of the effects of institutional programs would include information not only on the reduction of recidivism, but also on common undesired results, such as "sensory deprivation, labeling, the fostering of dependency, and the destruction of family ties" (Barton and Sarri, 1979:164). Compared with existing evaluation results, the availability of a varied and systematically generated set of information about the impacts of alternative programs could lead to quite different conclusions about the desirability of a program. For instance, a program that seems worthy of replication based on its desired impact on the offender may be rejected on the basis of accompanying unanticipated negative results which do not occur with alternative programs.

The probable impacts of criminal and juvenile justice programs on individuals other than those in the target group, on organizations, and on the community are even less well documented than the range of impacts on target group members. Writing about social action programs in general, Cain and Hollister (1972:128, 129) described effects which go unstudied as "external" or "third-party effects of the program." It is of considerable importance to have information about third-party effects of criminal and juvenile justice programs. Third parties who are positively affected can be cultivated to support the program and thereby promote continued implementation and successful replication. Third parties who think that they will be negatively affected, regardless of whether or not their expectations are well grounded, are likely to block the implementation of a program (e.g., Bardach, 1977; Lermack, 1977; Wycoff and Kelling, 1978). Advance information on program impacts can be used to assure some third parties that they will not be adversely affected by the program. Moreover, such advance information can demonstrate the need for program modifications, or for strategies such as cooptation to neutralize those who are likely to be adversely affected.

The undesirable narrow focus on a limited number of positive impacts is found in studies not only of correctional programs, but also of law enforcement and court programs. The narrow focus is inconsistent with a generally recognized conceptualization of law enforcement, courts, and corrections as an interdependent system located within the larger social system (e.g., Duffee, 1980). Munroe (1971:621) voiced a similar conclusion, writing that "the research effort is primarily oriented, in spite of the many brave words about a criminal justice system, towards specific agency and/or topical concerns." The focus of the achievement of a select few, intended program objectives precludes examination of unintended and/or

highly negative effects on people outside of the target population.

The purpose of this article is to show how the research approach, social impact assessment, can be used to correct the narrow focus of much research and to generate information on a wide variety of criminal or juvenile justice program impacts. The procedure for carrying out a social impact assessment is illustrated with information about a traditional screening program and the Community Arbitration Project (CAP), an Exemplary Project which is unique because it was studied with the social impact assessment method. Although a social impact assessment was used to compare traditional court screening with CAP, this was not reported in the official literature on CAP, most notably the Exemplary Project announcement entitled *Community Arbitration Project, Anne Arundel County, Maryland* (Blew and Rosenblum, 1979). The discussion which follows explains the social impact assessment method. It also demonstrates the process used to select the CAP and traditional screening impacts to be compared, the type of information resulting from social impact assessment, and the utility of such information in making informed decisions about program support, modification, and replication.

THE METHOD OF SOCIAL IMPACT ASSESSMENT

Social impact assessment involves the prediction and comparison of a wide range of impacts of two or more program options. Predictions are based on data drawn from "the literature, experts, project data, and direct experience" (Finsterbusch, 1980:22). Social impact assessment has been used primarily to study the impacts of large, publicly financed construction projects (e.g., Finsterbusch, 1980; Finsterbusch and Wolf, 1977). This concentration of applications has occurred because social impact assessments are legally required before government approval is given for construction of dams, power plants, airports, and tunnels, and for urbal renewal and transportation related construction. Although there is no legal basis for extending social impact assessment research to criminal or juvenile justice settings, social systems theory provides a theoretical basis.

Social systems theory (Forrester, 1969) has provided a theoretical rationale for the development of the social impact assessment research methodology. It also serves as the theoretical underpinning for contemporary conceptualizations of the criminal and juvenile justice system and for explanations of behavior in the system (Duffee, 1980). Both social impact assessment, and social systems theory as applied to the criminal and juvenile justice system, borrow from general social systems theory the concept of unexpected effects of changes in one part of the system on other parts. In this vein, Forrester (1969) has noted that social systems in general are characteristically counterintuitive; that is, a change in one part of the system is often followed by numerous causal reactions that include unplanned results.

The original purpose in legally requiring social impact assessments of public construction projects was to anticipate and in some cases avoid undesirable, counterintuitive effects of the projects. There is a clear utility in similarly anticipating undesirable effects of criminal and juvenile justice programs, either in the early stages of program operations or before a program is even chosen for replication. Two experts (Nagel and Neef, 1976) have noted so many unintended effects of criminal and juvenile justice programs that they entitled an article "Department of Unintended Consequences" (also see Vinter, 1979:158). The criminal justice literature is replete with examples of unintended program effects felt throughout the criminal and juvenile justice systems. For example, Intensive Special Probation projects, which were intended to reduce recidivism, unintentionally increased recidivism by increasing surveillance on probationers (Banks, 1977). In another example, incarceration of parents led to increased anti-social behavior of their children

(Sack, Seidler, and Thomas, 1976). As a final example, improved physical conditions in a prison had the unintended effect of disrupting relations among staff and inmates (Smith and Swanson, 1979). Such diverse research findings as those in the above examples exposed certain unintended effects of criminal and juvenile justice programs *after* they occurred. Social impact assessment is a technique for predicting many such effects before they occur, or at least in the early stages of a program's operation. The results of social impact assessment can thus be used by criminal justice decision makers to weigh undesired but probable effects against those which are intended. Moreover, the results can be used as a basis for modifying the program design to eliminate some undesired effects, to cultivate the support of groups which are positively affected, and to develop strategies for dealing with negatively affected groups.

AN APPLICATION OF SOCIAL IMPACT ASSESSMENT

The social impact assessment comparing CAP with a more traditional program was chosen to illustrate the approach to studying the effects of alternative programs. There are no other published examples of the formal application of social impact assessment in a criminal or juvenile justice setting known to the author. In fact, perhaps because the social impact assessment approach is so atypical in criminal justice research, the National Institute for Law Enforcement's publication on CAP (Blew and Rosenblum, 1979) omitted any description of the rationale for the choice of impacts to be studied, and the resulting gestalt of desired and undesired impacts of CAP and its alternative. Instead, the publication reorganized the research results to stress the positive effect of CAP on recidivism and on police and court operations. Only brief mention was made of some of the other impact areas, including negative impacts. Whereas the present description of the social impact assessment application to CAP allows the decision maker to weigh desired impacts against undesired impacts, the National Institute report attempted to persuade us that CAP is superior to a more traditional program, primarily because of its desired effects on recidivism and efficiency in case processing.

CAP is an Anne Arundel County, Maryland program designed to improve the procedure for selecting and informally treating juveniles whose offenses are not serious enough to warrant formal court attention. Blew and Rosenblum (1979) have compared the design of CAP, which was the first nationally recognized program of its type, with that of a more traditional screening hearing:

1. Police arrest youths to be sent to CAP by giving them a citation, similar to a traffic ticket. For the traditional screening hearing, police arrest youths by taking them to a police station.
2. The CAP hearing is conducted in a courtroom setting rather than in an office.
3. Victims are more often encouraged to voluntarily take part in CAP.
4. The arbitrators presiding at CAP hearings are attorneys who focus their attention on standards of legal sufficiency, whereas the personnel for the traditional screening hearing have social science backgrounds.
5. Volunteer community resources provide work experiences for offenders involved with CAP, and this gives them a way to redress an act against the community.
6. In CAP, work experiences, restitution, counseling and educational programs are used singly or in some combination in most of the cases for which there is sufficient evidence. In the traditional screening hearing, most youths are warned and no other activity is required.

These six innovations, described in more detail elsewhere (Blew and Rosenblum, 1979; Larom, 1976; Morash, 1978a), were intended to reduce the time between arrest and the hearing, rehabilitate offenders, and

divert a larger proportion of minor offenders from formal court.

Step 1: Identification of Possible Impacts

The first and key phase in carrying out a social impact assessment is the identification of a wide range of possible program impacts throughout the social system. Guidelines which have been developed to study the impacts of construction projects are useful in accomplishing this first phase. As one guide, Finsterbusch and Motz (1980:92–93) have published an "impact category tree" to suggest the possible location of impacts throughout the entire social system. To illustrate, the first major "branch" of the tree includes households. Within this branch, impacts on households can include, among other things, effects on consumption, work, well-being, education and socialization, leisure, social relations, and cultural and moral systems. With this impact category tree as a guide, the researcher uses interviews with local and other experts and the literature to generate a comprehensive list of the possible effects of each program option on each subcategory of each branch of the social system. The variety of effects is increased by considering the other major branches of the impact category tree in addition to households: communities, organizations and groups, and social institutions and systems.

To conduct the social impact assessment of CAP and the traditional screening program, this author's first step was to conduct a series of interviews with program staff and experts in the delinquency program area. The staff included the administrators responsible for designing both programs, supervisors of the programs, and administrators at a higher level who were familiar with juvenile justice system operations. These higher level administrators were from police and prosecutor's offices, as well as the juvenile services department. The experts were three university professors with knowledge of juvenile justice system operations and research.

Each person was shown a diagram of the impact category tree and asked for predictions of the impacts of CAP and traditional screening on each element. From this procedure and a review of the literature on juvenile justice programs it became clear that programs of the CAP and traditional screening type may differentially affect offenders, offenders' families, victims, victims' families, law enforcement personnel, court personnel, and correctional personnel. The programs also might differentially affect offenders' peer groups, other social service agencies that provide services to offenders, schools attended by offenders, community perceptions of crime, and citizens proximate to the program target area or to the offenders. The use of the impact category tree alerted the research staff and people interested in assessing the results of CAP to a range of program effects much broader than achievement of the explicit program objectives.

This researcher's role in the interviews was to inquire systematically about impacts in each branch of the impact category tree, and to ask follow-up questions regarding effects on biological, psychological, social, consumer, transportation, and other needs that might be either positively or adversely affected (Finsterbusch, 1980:23). These interviews not only set the stage for the choice of impacts to be studied, but they also served to educate everyone involved about the tremendous range of effects that could occur. To illustrate, the possible impacts of CAP and traditional screening on just victims included: physical safety from future harm by the offender, feelings of powerlessness as a victim, loss of time from work, fears of leaving the house to shop, potential damage by the offender to the victim's residence, needs for trasportation to the hearing, feelings that "justice was done," and fear of attending cultural or leisure time activities in the neighborhood (Morash, 1978b). This type of list is considerably more comprehensive than the usual program effects considered in most juvenile delinquency program evaluations.

Step 2: Identification of Probable Impacts for Study

Constraints on time, resources and information necessitate the selection of a limited number of the many possible impacts for study. In the case of the CAP-traditional screening impact assessment, the research results were intended for use by the administrators of CAP and by local juvenile justice system administrators throughout the nation who were considering shifting from traditional screening programs to CAP. Therefore, local administrators in Anne Arundel County and experts familiar with the operations of juvenile justice programs nationally judged the importance of including each impact in the study. The impacts which were chosen for study were considered to be (1) important reflections of achievement of explicit juvenile justice objectives (e.g., reduced recidivism), (2) potential evidence that third parties who could influence program implementation would benefit from one of the program options, and/or (3) potential indicators that third parties who could block program implementation would be adversely affected by one of the program options. The three local administrators and three experts who were surveyed shared considerable consensus on the impacts which met these criteria. Three impacts were chosen for study because they were related to highly valued, explicit objectives of the juvenile justice system: reduction of recidivism, exposure of offenders to the value that the law is legitimate, and punishment for illegal behavior. Because increasing youth involvement with the justice system was sharply at odds with the objective of a screening program and has occurred in numerous screening programs (reviewed by Allison, 1979), the proportion of juveniles arrested and the proportion diverted from formal court were two additional impacts chosen for study. In the larger social system, police, victims, and community groups, as well as other elements in the juvenile court system, could potentially interrupt the implementation of CAP or traditional screening. Alternatively, they could provide valuable support for the programs. Thus, effects on police, victims, the community and other elements of the juvenile court system were chosen for study.

Step 3: Collection of Data to Reflect Impacts

Impacts were assessed either by generalizing the results of prior research on programs similar to CAP and traditional screening or, if prior research was inadequate, by collecting new data. Prior research (Horney, 1980) showed that citations similar to those used to make arrests for CAP had influenced police to arrest higher proportions of suspects. Similarly, several studies (Allison, 1979) showed that screening (i.e., diversion) programs have tended to recruit more rather than fewer clients for the juvenile justice system. Based on this prior research, CAP was likely to result in increased arrests.

Because there was no existing literature on many of the impacts of interest, most of the information on CAP and traditional screening was generated from an experiment to compare the two programs as they operated in Anne Arundel County. The experiment was managed by instituting the CAP program for only half of the eligible youths during the first year of operation. For the year, a randomly assigned group of offenders eligible for CAP took part in traditional screening instead.

For each of the probable impacts that had been chosen for study, documentation was assembled in summary form. Table 1 is a prototype of a resulting social impact statement summary, in which the relative effects of CAP and traditional screening are assessed on the basis of prior research and direct study of the two programs. The table indicates that, in comparison to traditional screening, CAP programs are likely to result in less recidivism, a reduced allocation of police time to juvenile work, an increase in arrests, a decrease in youths referred to court, a reduced juvenile court staff work load, and increased citizen-offender interactions after the hearing. For the other impacts studied, there were no differences.

UTILIZATION OF IMPACT ASSESSMENT RESULTS

In the case of the CAP-traditional screening impact assessment, results could be used by two different audiences for two distinct purposes. First, administrators of CAP could use the results to develop support for the program and to correct negative impacts by making program modifications. Second, persons who were in the process of deciding to replicate CAP could use the results of the impact assessment as the basis for their decision.

The most complete information about the utilization of the results of the CAP-traditional screening assessment pertains to use by CAP administrators in the formative stages of the program. The finding that the citation system resulted in increased arrests was judged to be a serious threat to the success of the program; although it was somewhat offset by the subsequent high rate of diverting the arrested youths from court, CAP was intended to produce less, not more involvement of offenders with the juvenile justice system. The negative finding was made known to supervisors in the city, county, and state police departments, and a police liaison officer was hired to work for the juvenile court department as a link with the police departments. By alerting police to the findings, and communicating with them on a regular basis about the appropriateness of their arrests, an effort was made to offset the increase in referrals of youths from the police department to the juvenile services department. A preliminary study (Morash and Anderson, 1977) showed that the initial increase was offset. Although we cannot conclude with certainty that the efforts to slow the arrest of juveniles were the cause of the decrease, the attempted application of impact assessment results is a good illustration of their potential utility.

Another way in which CAP administrators used social impact assessment results was to prevent the negative reaction of parties who thought they were adversely affected by the program. The juvenile court department in Anne Arundel County includes the intake workers, who are the staff for traditional screening. There was a potential threat to the intake workers when their caseload was divided with CAP, since one could argue that fewer intake workers were needed. The threat did not materialize in Anne Arundel County, primarily because the juvenile court department had a strong orientation toward delivering treatment services. Therefore it was consistent with department philosophy and staff training to shift traditional screening staff's resources away from the large group of minor offenders served by CAP and towards intensive treatment for status and serious offenders remaining in their caseloads. The shift was documented in the social impact assessment as a positive result, and this information was used to encourage support for CAP.

The desired and positive impacts of CAP on youths, police and the community (Table 1) were publicized to develop ongoing support for the program. This information was conveyed to the general public through the media, and to the police through both formal and informal channels. The police in particular were positive about the CAP impacts which coincided with their goal of doing the "real police work" of "crook catching" (Manning and Van Maanen, 1978:5), made possible by the reduced allocation of time to juvenile work. Also, positive information about CAP was used to promote the program among state officials with jurisdiction over juvenile justice operations.

There are no systematic data on the use of results of the CAP-traditional screening assessment by persons in other jurisdictions who were making decisions about replicating CAP. As was mentioned above, the results of the full impact assessment were not widely distributed, but instead were reorganized in the nationally distributed Exemplary Project report (Blew and Rosenblum, 1979) to emphasize the greater benefit of CAP as opposed to traditional screening. It is possible, however, to speculate on how the results of a social impact assessment could be used as input into the decision to replicate a program. This speculation is

TABLE 1

PROTOTYPE OF A SUMMARY OF SOCIAL IMPACT ASSESSMENT RESULTS: COMPARISON OF CAP AND TRADITIONAL SCREENING

Location of Impact and Impact Statement	*Basis for Making the Impact Statement*
I. Offenders	
1. *Recidivism* Although CAP does not differ from traditional screening in effect on youths' attitudes, it does result in less recidivism, especially for property offenders.	*1a. Youths randomly assigned to CAP or traditional screening did not differ in their view of the law, the legal system, themselves, or the victim (Morash, 1978a, 1978b). †1b. In another comparison of randomly assigned youths, 482 CAP youths had a statistically lower recharge rate per youth than 342 youths in traditional screening, particularly if they were property offenders (Blew and Rosenblum, 1979:57).
2. *Explanation of Law's Legitimacy* The CAP hearing exposes youths to reasoning about the law more than does traditional screening.	*2a. CAP youths were: 20% more likely to report that the law had been explained to them; 24% more likely to report they were told they hurt society; and, 40% less likely to report that hearing time was spent to discover family problems.
3. *Severity of Punishment* CAP youths perceived that they are treated more harshly than do traditional screening youths.	*3a. The CAP courtroom-like setting produced fears of "being sent away," with 16% of CAP youths and 3% of traditional screening youths reporting this fear. *3b. 19% of CAP youths as opposed to 10% of those in traditional screening said they were scared or nervous during their hearing. *3c. 57% of CAP youths compared with 74% in traditional screening said they were pleased with the hearing outcome. In CAP the major reason for displeasure was the work or counseling assignment; 23% of youths gave this reason.
4. *Proportion of Youths Arrested* In comparison to traditional screening hearings, CAP is likely to promote increased involvement of youths with the justice system.	4a. Citations like those used to make an arrest for CAP have influenced police to arrest a greater proportion of offenders (Horney, 1980). 4b. Several precourt screening programs similar to CAP have unintentionally recruited clients who previously would have just been warned by police (Allison, 1979). 4c. In the year after CAP began, police arrested 19% more offenders than they had when the traditional screening was used (Morash and Anderson, 1977).
II. Victims	
1. *Participation* Victims participate more frequently in CAP than in traditional screening.	*1a. 50% of CAP victims and 15% of traditional screening victims participated in the hearing.
2. *Satisfaction* Victims in CAP and traditional intake do not see either procedure as better meeting their needs than the other.	2a. Randomly assigned groups of 45 CAP and 46 traditional screening victims were not different in anticipation of further difficulties with the offender; anticipation of future physical assault by the offenders or their friends; evaluation of whether the offender was treated justly; satisfaction that they could "speak their piece" at the hearing; agreement with the hearing outcome; attitude toward offenders or their families; and view of juvenile justice personnel as cooperative (Morash, 1978b).

Location of Impact and Impact Statement	*Basis for Making the Impact Statement*
3. *Future Interactions with Youths* Victim correctly perceived that CAP does not differ from traditional screening in its effect on future victim-of-fender interactions.	3a. A minority of youths in both CAP (18.4%) and traditional intake (12.2%) felt they had hurt the victim or society. Nearly half (47%) of CAP youths and 40% of traditional screening youths stated they would not repair damages to the victim. 10% of CAP offenders and 5% of traditional hearing offenders felt the victim deserved to be physically hurt after the hearing.
III. Police Department	
1. *Proportion of Time on Juvenile Work* Police spend less time on juvenile work with the CAP procedures than traditional hearing procedures.	1a. Interviews with 87 police officers showed that when citations and the CAP hearing procedure replaced traditional screening, police needed less time to complete paperwork, attend juvenile court, transport juveniles, supervise arrested juveniles, and placate angry victims. A similar before and after comparison in neighboring counties showed that this increase was not due to more general changes in law enforcement practices (Morash and Anderson, 1977).
IV. Juvenile Court Department	
1. *Diversion from Formal Court* CAP diverts 10% more youths from formal court than does traditional screening.	*1a. CAP diverted 99% of the youths, and traditional screening diverted 89% (Morash, 1978a). This offset increased arrests by police.
2. *Caseload* The caseload for traditional screening staff was substantially reduced, allowing them to work intensively with status and serious offenders.	2a. Before CAP began, the traditional screening staff handled all minor offenders, status offenders, and serious offenders. Their caseload was reduced by half when responsibility for all minor offenders was shifted to CAP. 2b. The traditional screening staff in the county where CAP was located shifted their resources to an intensive counseling program for status offenders and an innovative detention program in which youths were allowed to remain in their own homes but were given intensive supervision.
V. Community	
1. *Community Involvement with Juvenile Offenders* CAP stimulates more community involvement with offenders than does traditional screening.	1a. Over 100 community groups and agencies regularly provide counseling or work settings for juveniles referred by CAP, whereas traditional screening did not result in any community-offender involvement.
2. *Offenders' Contribution to Community* CAP results in juveniles' volunteering their efforts in community betterment projects, whereas traditional screening does not.	2a. CAP administrative records show that during 4½ years, CAP offenders completed 15,000 hours of work for community groups and agencies, whereas no work was required of traditional screening offenders.

*The comparison is of 201 cases randomly assigned to CAP or traditional screening.
†The comparison is of 824 cases randomly assigned to CAP or traditional screening.

based on general knowledge of the criminal and juvenile justice system and on studies of the social impact assessment approach applied in other policy areas.

When social impact assessment results are to be used as input into choosing between program options (in our case, deciding to replicate CAP), the information is circulated to interest groups and individuals who have some interest in the decision outcome. Their reaction is included in a final summary report, and becomes a part of the input into the decision making. The decision outcome depends on a weighing of the negative and positive expected impact.

Weighing the probable impacts as identified by social impact assessment differs somewhat from the classical problem of the relative importance of multiple program objectives (Cain and Hollister, 1972:112). If the impact assessment has been thorough, and is not biased towards one or another group, results will reflect different valuations of impacts as seen by the different people affected. In the CAP-traditional screening case, for example, juvenile offenders would negatively value the increased punishment by CAP, and juvenile officials and elements of the public and police would positively value it. The results of a social impact assessment should clarify such interest group differences. Given the relativity of weights attached to probable impacts, it is not accurate to view a social impact statement as definitive support for the choice of a particular program. This is a major difference from the Exemplary Project publication (Blew and Rosenblum, 1979), which presented evidence in favor of one option over the other. Rather, impact assessment results are a valuable input into debates between individuals over the appropriate program choice.

In evaluating the feasibility of using social impact assessment results as input into decision making about program choices, it is essential to consider whether the model is compatible with the way program choices currently are made. The dynamics of and influences on the choice of one criminal or juvenile justice program over another are points of contention within the criminological community. One major framework, the pluralistic view, is however, highly consistent with the above description of the use of social impact assessment results. According to the pluralistic view, the choice of programs depends on interactions between special interest groups, citizens, officials, and organizations. Research on many types of public policy (Clark, 1973; Lindbloom, 1968; Lowi, 1961, 1971) and on criminal and juvenile justice policy in particular (Allison, 1979; Lemert, 1970; Lemert and Dill, 1978) have provided support for the pluralistic framework.

With this pluralistic explanation of program choice in mind, social impact statements are potentially useful in making people aware of their "stake" in program choices. This opening up of the decision-making process takes place when the impact assessment results are used for public construction projects, as documented by Finsterbusch (1980:29–30). He wrote that the social impact assessment "opens the decision making process to environmentalists and negatively affected parties. At the same time, the [social impact assessment] provides opportunities for active publics to input into the decision making process." Based on a review of several hundred public construction program assessments, and involvement in one hundred, Friesema and Culhane (1976:340) similarly concluded that requirements to carry out a social impact assessment "seems to have created a new, complex political process which can be and has been used very effectively to improve the social and environmental sensitivity of government decision makers." Furthermore, they (1976:349) wrote that the subjection of each social impact statement to review by interest groups "gives increased access to environmental, *ad hoc* community and public interest groups, particularly those groups which might not otherwise have close, informal access to decision makers." It should be noted that Finsterbusch and Friesema and Culhane are not saying that the quality of research is improved by social impact assessment, but

instead that the quality of decision making is improved. Additionally, this improvement occurs not only because the social impact assessment approach is used, but also because funding and approving agencies require a satisfactory resolution of interest group conflict before allowing for program implementation.

Social impact assessment would be useful in the juvenile justice system, where the need to open up decision making is particularly acute. Sarri and Vinter (1979:169) concluded from their national survey that "juvenile justice appears to be a relatively marginal area of government activity everywhere; juvenile justice has no general constituency among the states; and, few interest groups regularly attend to it." In another article, Vinter (1979:158) describes the need for the type of information produced by social impact assessment: "Those who participate in the formulation and implementation of justice policy must develop much greater factual knowledge of the actual patterns and of the systemic relations that characterize this area." It is interesting that Vinter concluded that, regardless of whether public interest groups' attention was positive or negative, it "often provides the vital margin for change." It is possible that social impact assessment information could stimulate public involvement in juvenile justice system decision making.

Although adult focused law enforcement, court, and correctional programs are more often in the public limelight than juvenile programs, it would be inaccurate to describe decision making about the adult programs as "open." Musheno, Palumbo, and Levine (1976) contended that staff often affect program choice by blocking implementation. This has been shown to be true of large-scale programs, in which agency administrators played an inordinately large role in shaping program choice and other interest groups have not been able to exert influence (e.g., Wycoff and Kelling, 1978). There appears to be a need to use information, such as that provided by social impact assessment, to bring interest groups into the criminal justice policy making process.

It would be naive to equate the use of social impact statements with the quality of decision making about criminal and juvenile justice programs. On the other hand, the typical practice of closed decision making without benefit of information on a full range of likely impacts is even more difficult to defend.

CONCLUSION

This article has demonstrated that the method of social impact assessment, which was developed to study probable effects of public construction projects before a program is implemented, could usefully be applied to ensuring successful program development and to decision making about choices between alternative criminal or juvenile justice programs. The wide scope of information that can be produced with social impact assessment was illustrated by the prototype social impact statement for CAP and traditional screening, two juvenile justice program options. Applied to law enforcement, court, and correctional programs for both adults and juveniles, the social impact assessment could similarly be used to predict desired and undesired and planned and unplanned effects throughout the social system. Such information also could alert those who seek to implement a program to the need for program modifications to prevent undesired effects, to potential allies of the program, and to groups who will try to block the program due to adverse effects on them. Additionally, the information could highlight the concerns of various affected interest groups, thereby stimulating their input into decisions about the choice of programs for local replication.

ACKNOWLEDGMENTS

The author would like to thank the following staff at the Anne Arundel County, Maryland Department of Juvenile Services for their extensive help in carrying out the social impact assessment of CAP and traditional screening: David Larom, Regional Director; Kay Peacock, Director of the Community

Arbitration Project; and, Debra Choulis, Research Assistant. Additionally, Dr. Jack Greene, of the Michigan State University School of Criminal Justice, provided extremely helpful comments on an earlier draft of this article.

REFERENCES

Allison, R. S. (1979). LEAA's impact on criminal justice: A review of the literature. *Crim. Just. Abstracts* 11:608–648.

Banks, J. (1977). *Evaluation of intensive special probation projects*. Washington, DC: U.S. Government Printing Office.

Bardach, E. (1977). *The implementation game*. Cambridge, MA: MIT Press.

Barton, W. H., and Sarri, R. (1979). Where are they now? A follow-up study of youths in juvenile correctional programs. *Crime and Delinq.* 25:162–176.

Blew, C. H., and Rosenblum, R. (1979). *An Exemplary Project: The Community Arbitration Project*. Washington, DC: U.S. Government Printing Office.

Cain, G. G., and Hollister, R. G. (1972). The methodology of evaluating social action programs. In *Evaluating social programs*, ed. P. H. Rossi and W. Williams. New York: Seminar Press.

Clark, T. N. (1973). *Community power and policy outputs: A review of urban research*. Beverly Hills: Sage.

Duffee, D. E. (1980). *Explaining criminal justice*. Cambridge, MA: Oelgeschlager, Gunn and Hain.

Finsterbusch, K. (1980). *Understanding social impacts*. Beverly Hills: Sage.

Finsterbusch, K., and Motz, A. D. (1980). *Social research for policy decisions*. Belmont, CA: Wadsworth Publishing Company.

Finsterbusch, K., and Wolf, C. P., ed. (1977). *The methodology of social impact assessment*. Stroudsberg, PA: Dowden, Hutchinson, and Rose.

Forrester, J. W. (1969). *Urban dynamics*. Cambridge, MA: MIT Press.

Friesema, H. P., and Culhane, P. J. (1976). Social impacts, politics, and the environmental impact statement process. *Natural Resources J.* 16:339–356.

Horney, J. (1980). Citation arrest. *Criminol.* 17:419–434.

Larom, D. (1976). The arbitration experience. *Juv. Just.* 27:13–17.

Lemert, E. M. (1970). *Social action and legal change: Revolution within the juvenile court*. Chicago, IL: Aldine.

Lemert, E. M., and Dill, F. (1978). *Offenders in the community*. Lexington, MA: Lexington Books.

Lermack, P. (1977). Hookers, judges, and bail forfeiters: The importance of internally generated demands on policy-implementing institutions. *Admin. and Soc.* 8:459–468.

Lindbloom, C. E. (1968). *The policy making process*. Englewood Cliffs, NJ: Prentice-Hall.

Lowi, T. J. (1961). *The end of liberalism*. New York: W. W. Norton.

——— (1971). *The politics of disorder*. New York: Basic Books.

Manning, P. K., and Van Maanen, J., ed. (1978). *Policing: A view from the street*. Santa Monica, CA: Goodyear.

Morash, M. (1978a). *Factors related to the development of legal reasoning: Implications for juvenile justice policy*. Ph.D. dissertation, University of Maryland.

——— (1978b). Delivering justice to victims of juveniles' misdemeanors: A comparison of intake procedures and a description of victims. In *The evolution of criminal justice*, ed. J. Conrad. Beverly Hills: Sage.

Morash, M., and Anderson, E. (1977). Impact assessment: A technique for evaluating criminal justice programs. *Crim. Just. Rev.* 2:23–34.

Morrison, P. A. (1978). Overview of demographic trends shaping the nation's future. Testimony before the Joint Economic Committee, U.S. Congress, May 31.

Munroe, J. L. (1971). Towards a theory of criminal justice administration: A general systems perspective. *Pub. Admin. Rev.* 31:621–631.

Musheno, M.; Palumbo, D.; and Levine, J. (1976). Evaluating alternatives in criminal justice: A policy impact model. *Crime and Delinq.* 22:265–282.

Nagel, S. S., and Neef, M. G. (1976). Department of unintended consequences. *Policy Analysis* 2:356–359.

Sack, W. H.; Seidler, J.; and Thomas, S. (1976). The children of imprisoned parents: A psychosocial exploration. *American J. of Orthopsychiatry* 46:618–628.

Sarri, R. C., and Vinter, R. D. (1976). Justice for whom? Varieties of juvenile correctional approaches. In *The juvenile justice system*, ed. M. W. Klein. Beverly Hills: Sage.

Sechrest, L.; White, S. O.; and Brown, E. D. (1979). *The rehabilitation of criminal offenders: Problems and prospects*. Washington, DC: National Academy of Sciences.

Smith, D. E., and Swanson, R. M. (1979). Architectural reform and corrections: An attributional analysis. *Crim. Just. and Behav.* 6:275–293.

Vinter, R. D. (1979). Trends in state correction: Juveniles and the violent young offender. *Crime and Delinq.* 25:145–161.

Wycoff, M. A., and Kelling, G. L. (1978). *The Dallas experience: Organizational reform*. Washington, DC: Police Foundation.

IV

PROGRAM MODELING

One standard goal of evaluation research is assessing whether a program makes a difference or not. Those working in the evaluation field, however, are increasingly recognizing that it is as important (and sometimes more so) to determine why a difference does or does not occur. Program sponsors, once they understand the process by which their program components resulted in outcomes, are able to give separate consideration to the strengths and weaknesses of the program, to develop problem-focused improvement schemes, and to make more informed decisions about the program's future. Whereas any planned treatment program has an underlying model of change, the content and structure of the model are often not well defined by program personnel, and in any case tend to become obscured by the day to day realities of program operation. Thus, a major task for the evaluator is to explicate the process by which a program affects outcomes. This is a particularly challenging task, because even programs that have identical goals for similar target populations are more likely than not to vary in treatment content, program format, service delivery, and, especially, the context in which they operate.

One way evaluators may meet this challenge is by conducting what Chen and Rossi call theory-driven evaluations. Chen and Rossi believe that evaluators must be cognizant of how human organizations work and how social problems are generated. They strongly advocate constructing plausible and defensible models of program processes before evaluation begins. A general process model would include characteristics of the delivered treatment variables, related exogenous factors, intervening processes and outcome variables, as well as hypotheses regarding relationships among these components. In brief, modeling requires the evaluator to explicate as much as possible the causal environment in which the program operates. If the components of the model are considered throughout the design, measurement, and analysis phases of the evaluation, then the evaluation report may go beyond giving a blanket conclusion about the efficacy of the program, to characterize how, why, and for whom the intervention is or is not effective. This increased specificity helps evaluators avoid the narrow and sometimes distorted understanding of programs with which Chen and Rossi are concerned. Moreover, the increased specificity attained by modeling provides more points with which to assess the comparability of apparently similar programs, thereby providing more rigorous criteria for the integration of evaluation data.

Moos and Finney provide an excellent review of how a conceptually based, process-oriented framework can be fruitfully applied to the evaluation of

alcoholism treatment programs. Like Chen and Rossi, Moos and Finney believe that traditional summative evaluations do not adequately capture the complexity of either treatment programs or the extra-treatment and contextual factors that can mediate program effects. Some of the complexities uncovered by process-oriented evaluations include the finding that the perceived quality of the treatment program (e.g., cohesiveness, organization) and characteristics of the patient's work environment (e.g., job satisfaction) can affect the outcome of an alcoholism intervention. Moos and Finney believe that the process-oriented model they illustrate can help generate new and potentially more effective intervention strategies and, by accounting for within-group differences in treatment response, help match individuals to optimal treatments.

A key component of both the Chen and Rossi and the Moos and Finney models is conceptualizing a program's implementation system as an integral part of the treatment process. As Moos and Finney convincingly argue, few programs are so simple or direct that their implementation can be assumed to occur exactly as planned. Indeed, characteristics of the sponsoring agency, personnel, treatment facilities, and clients may all affect how much and what specific forms of a given treatment are delivered. These are elements beyond the evaluator's control that will nevertheless affect the validity and interpretability of the evaluation data. Thus it is important that evaluators gather empirical evidence of the extent to which program components are actually implemented. Conrad and Eash's longitudinal study of compensatory education programs provides a recent example of how implementation measurement has been included as part of an evaluation project. In this study, trained observers used the Classroom Observation Rating Scale to record program implementation.

According to Scheirer and Rezmovic, standard methodological paradigms for constructing implementation measures have not yet been developed, in spite of the critical importance of assessing program implementation for the evaluation of program outcomes and examination of innovation processes. Based upon their review of 74 previous studies that used implementation measures, Scheirer and Rezmovic discuss the conceptual and methodological issues surrounding the measurement of degree of implementation. A particularly important issue they discuss is whether the different techniques for measuring degree of implementation yield comparable findings. They also question the adequacy of measuring tools now being used to assess the degree of implementation. The basic conclusion reached by Scheirer and Rezmovic is that much work remains to be done in the young field of implementation research; they offer several suggestions for the development of more adequate measures.

Program modeling requires extra time and effort from the evaluator to conceptualize the elements of the intervention, its implementation system, and the extraneous and mediational factors that may affect these elements. In the long run, however, this process should improve the power of evaluation designs and enhance the validity of evaluation findings.

16

Evaluating with Sense

The Theory-Driven Approach

Huey-tsyh Chen and Peter H. Rossi

For more than two decades discussions about the appropriate methodology for estimating the net effects of social programs have been dominated by the paradigm of the randomized controlled experiment. For some evaluation commentators (e.g., Suchman, 1969; Campbell and Stanley, 1966; Cook and Campbell, 1979) alternative designs for impact assessment are valued to the extent that such designs mimic the validity advantages of randomized experiments. For others (e.g., Scriven, 1972; Guba and Lincoln, 1981; Deutscher, 1977) the paradigm is used as an example of what not to do in assessing the effects

---

AUTHORS' NOTE: *This article is a revised version of a paper presented at the 1982 meeting of the Evaluation Research Society. Preparation of this article was aided by a grant from the National Science Foundation (Grant SES-8121745), of which P. H. Rossi is the Principal Investigator.*

From Huey-tsyh Chen and Peter H. Rossi, "Evaluating with Sense: The Theory-Driven Approach," *Evaluation Review*, 1983, 7(3), 283-302.

of programs—arguments that often stress the artificiality of standardized treatments and accompanying data collection strategies, especially for labor-intensive human services programs.

The domination of the experimental paradigm in the program evaluation literature has unfortunately drawn attention away from a more important task in gaining understanding of social programs, namely, developing theoretical models of social interventions. A very seductive and attractive feature of controlled experiments is that it is not necessary to understand how a social program works in order to estimate its net effects through randomized experiments, provided that the goals and objectives of a program can be specified in reasonably measurable terms. Thus cookbook evaluation manuals (e.g., Morris et al., 1978) can outline how to proceed in an evaluation with scarcely a mention of any theory underlying the programs in question. Or evaluability assessments (Wholey et al., 1975) can concentrate on whether or not goals are sufficiently defined to permit the application of experimental or quasi-experimental evaluation methods. Even the critics of the experimental paradigm have had their attention distracted from seriously considering the theoretical issues in social programs by concentrating on the false issue of artificial data collection methods and the old social science finding that the actual goals of human organizations are often not their professed goals.

An unfortunate consequence of this lack of attention to theory is that the outcomes of evaluation research often provide narrow and sometimes distorted understandings of programs. It is not usually clear whether the recorded failures of programs are due to the fact that the programs were built on poor conceptual foundations, usually preposterous sets of "causal" mechanisms (e.g., the Impact Cities program); or because treatments were set at such low dosage levels that they could not conceivably affect any outcomes (e.g., Title I); or because programs were poorly implemented. Note that the emphasis in the above statements is on deficiencies in the theoretical underpinnings of the treatment or of the treatment delivery systems.

The purpose of this article is to bring theory back into program evaluation. Our aim is not to make a case for basic research—there is enough justification for that goal—but to make a case that neglect of existing theoretical knowledge and of thinking theoretically has retarded both our understanding of social programs and the efficient employment of evaluation designs in impact assessment. This perspective on

evaluation research we have called elsewhere (Chen and Rossi, 1980) the "theory-driven" approach to evaluation—a perspective, we believe, that has promise to yield better information on social programs, as well as rich yields to the basic social science disciplines.

Of course the kind of theory we have in mind is not the global conceptual schemes of the grand theorists, but much more prosaic theories that are concerned with how human organizations work and how social problems are generated. It advances evaluation practice very little to adopt one or another of current global theories in attacking, say, the problem of juvenile delinquency, but it does help a great deal to understand the authority structure in schools and the mechanisms of peer group influence and parental discipline in designing and evaluating a program that is supposed to reduce disciplinary problems in schools. Nor are we advocating an approach that rests exclusively on proven theoretical schema that have received wide acclaim in published social science literatures. What we are strongly advocating is the necessity for theorizing, for constructing plausible and defensible models of how programs can be expected to work before evaluating them. Indeed the theory-driven perspective is closer to what econometricians call "model specification" than are more complicated and more abstract and general theories.

Nor do we argue for uncritically using the theories that may underlie policymakers' and program designers' views of how programs should work. Often enough policymakers and program designers are not social scientists and their theories (if any) are likely to be simply the current folklore of the upper-middle-brow media. The primary criterion for identifying theory in the sense used in this article is consistency with social science knowledge and theory. Indeed theoretical structures constructed out of social science concerns may directly contradict what may be the working assumptions of policymakers and program designers.

It is an acknowledged embarrassment to our viewpoint that social science theory is not well enough developed that appropriate theoretical frameworks and schema are ordinarily easily available "off the shelf." But the absence of fully developed theory should not prevent one from using the best of what is already at hand. Most important of all, it is necessary to think theoretically, that is, to rise above the specific and the particular to develop general understandings of social phenomena.

A GENERALIZED MODEL FOR PROGRAM EVALUATION

A useful general scheme for identifying the main components of any social program is shown in Figure 1. The main causal relationships that would have to be worked out in any useful model of a program are shown in that diagram as arrows connecting the main components.

Central to the diagram are "delivered treatment variables," which constitute the program to be evaluated insofar and inasmuch as has been delivered. Note that treatment is conceived not as designed but as actually delivered, the delivery being affected by an implementation system that includes organizations, personnel, facilities, clients, and regulations concerning eligibility.

At the far right of the diagram are "outcome variables," those aspects—intended and unintended—that are affected either directly by treatment variables and/or mediated through a set of intervening variables, represented by the box so labeled. Whether or not intervening processes are present is a matter of conceptualization. Thus the provision of transfer payments as a treatment directly affects the incomes of clients, but the introduction of a new payment incentive system into a work organization will only affect incomes if the appropriate intervening processes postulated come into play.

The classic statement of the problem of inferring treatment effects centers around the contents of the box labeled "exogenous variables." The outcomes of social programs are rarely, if ever, attributable solely to treatments and intervening processes but are also determined by other sources: those exogenous variables correlated with the treatment variables and stochastic disturbances that are independent of the treatment exogenous variables. The exogenous processes may or may not be correlated with treatment, but they often are. Confounded estimates of the treatment effects usually result from the failure to control for correlated exogenous variables, which leads to a correlation between the disturbance and the treatment. Indeed, one of the great virtues of the classic Campbell and Stanley (1966) statement is the identification of some very general ways in which exogenous processes may be correlated with treatment variables, thereby confounding estimates of effects except under special circumstances. Indeed it is the special (but not unique) quality of the classical randomized experiment that its use can make exogenous variables uncorrelated with treatment variables.[1]

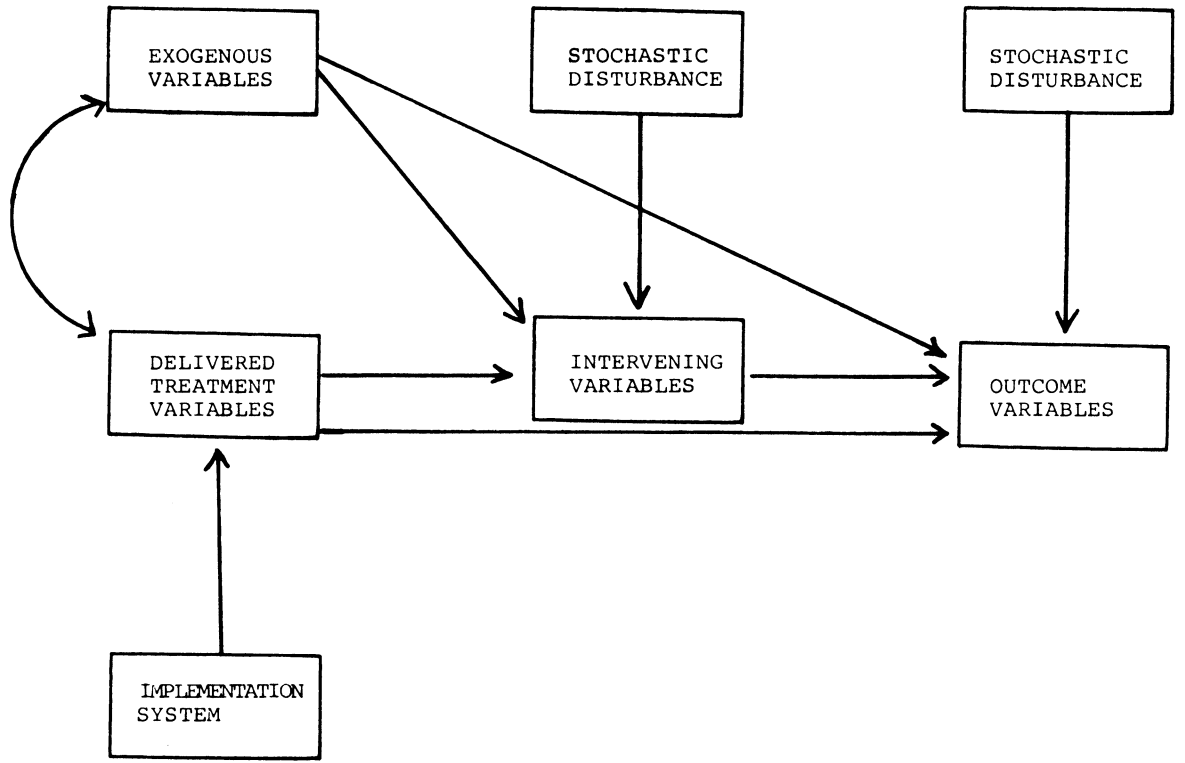


Figure 1: Schematic Representation of Program Models

OUTCOME VARIABLES: SPECIFYING THE GOALS OF PROGRAMS

The problem of specifying outcome variables is usually phrased as goal specification—the determination of those outcomes that policymakers and/or program designers envisaged as the intended outcomes of the program. This issue constitutes one of the important distinctions between basic and applied social research. In basic research, outcome variables express the disciplinary interests of the researcher; in applied social research, outcome variables are those of interest to policymakers or other sponsors of applied work.

As traditionally viewed, goal specification in evaluation research tends to be a search for appropriate operational definitions of the intended effects of programs. Since such definitions are sometimes cloaked in obscure and ambiguous statements, goal specification can be a separate empirical research enterprise of its own, as Wholey's (1975) evaluability assessments exemplify. Some of the consequences of searching for those intentions that are measurable have been noted by many commentators (e.g., Deutscher, 1977; Scriven, 1972; Chen and Rossi, 1980). First of all, outcome variables tend to be narrower than the connotative intentions of program designers and/or policymakers. Thus the goals of Head Start were defined by the evaluators (Cicirelli et al., 1969) in cognitive terms mainly because such goals could be more easily operationalized than others that were more vaguely formulated. Second, there is a large gap between enabling legislation and the actual design of programs. Designers may narrow the goals of legislators, elaborate upon them, or substitute entirely different goals. Sometimes the enabling legislation deliberately fosters diversity in the processes of implementation, as in Head Start. In other cases local conditions may appear to require extensive adaptation, as, for example, in the Planned Variation Education Program. Some of the critics of the experimental paradigm were among the first to note these program changes in execution (e.g., Scriven, 1972; Deutscher, 1977; Chen and Rossi, 1980) recommending that the goals also be inferred from actual program operations, rather than solely from policymaker statements or legislative intents. Indeed in the case of some programs, program operators are encouraged to develop specific goals as the program is designed (e.g., High Impact and Head Start); hence the goals of programs cannot be described in any specific sense through a consideration of legislative intent or policymaker statements alone.

Most of the commentators who have advocated specifying goals through empirical observation of programs in operation have been

strangely silent on how goals should be inferred through observation. Some come close to advocating connoisseurial approaches, i.e., "Anyone with any experience and smarts will obviously see that . . . "

We believe that there is some wisdom to the admonition that goal specification be empirically based. However, the process whereby it is possible to go from observations to goal specification is through social science theory and knowledge, not by the craftlore of experts and consultants—a point to which we will return at a later point in this article. Nor are we advocating ignoring the goals of policymakers in filling out the content of the goals specified in an evaluation. Indeed it is useful to conceptualize goals as falling into one or the other of the following three classes.

Policy-Directed, Plausible Goals

These are goals explicitly formulated by those who designed and/or authorized the program and that are plausible in the following senses: (1) The goals are consonant with current knowledge of the problem to which the program is directed; (2) the program is designed so that it can be implemented without heroic efforts; (3) the resources allocated to the program are sufficient to deliver the treatment at reasonable dosage levels.

For example, the policy goal of the current speed limit 55 miles per hour was to lower the consumption of gasoline by motor vehicles. The goal was plausible in the sense that prior knowledge of the gas consumption characteristics of typical gas engines indicated that consumption would be so lowered. The mechanism of implementation involved tying the passage of appropriate state laws to continuing federal highway fund allocations—a reasonable strategy.

Policy-Directed, Implausible Goals

These are goals specified explicitly by those who designed/authorized a program and which are *not* plausible in the following senses: (1) The goal is so vague that a relatively large number of specific operationalizations, some of which are mutually contradictory, are possible; (2) the goals are not consonant with current knowledge about the problem in question; (3) the program is not designed so that it can be implemented by the agency given responsibility; (4) the resources allotted to the program are not sufficient to deliver the treatment at reasonable dosage levels.

For example, although the major goals of the 1968 Federal Firearms Regulation Act were specific enough (i.e., to restrict access to firearms on the part of felons, the insane, and certain other categories of persons), the mode of implementation, namely, requiring the registration of gun dealers and forbidding them to sell to the proscribed social categories, was bound to fail since it was based on the assumptions that gun-using criminals obtained their weapons through gun dealers, that gun dealers could discern which of their potential customers fit into the proscribed categories, and that gun registration records could be easily accessed to trace gun ownership. None of these assumptions was tenable and some went quite contrary to existing established knowledge (Wright et al., 1983).

Theory-Derived Goals, Not Specified by Policy

These are goals that are plausible but not specified in policy directives, and which can be discerned either through the a priori examination of policy and program or through the empirical study of the program (Chen and Rossi, 1980).

For example, in the TARP experiments (Rossi et al., 1980), official goals included the reduction of recidivism among released felons through the extending to this group eligibility for unemployment benefits payments. Consideration of the potential work disincentive effects of such payments, as strongly suggested in the writing of labor economists, led the investigators to postulate work disincentive effects.

Except for political reasons, it obviously makes little sense to evaluate a program whose goals were only policy directed and implausible; but there is no way to decide whether a program's goals fall into such a category without careful consideration of whether or not existing social science theory and knowledge would support such a judgment of plausibility. The main point of making such distinctions among goals, however, is to highlight the fact that programs may be accomplishing some things that were not intended by their designers and that such effects may be either desirable or undesirable, may sometimes (as in the case of the TARP experiments) produce effects that offset those intended, and that a good evaluation should take into account inferred effects as well as those directly intended.

As indicated above, the judgment whether or not a set of policy-directed goals is plausible depends on examining the total program in

light of existing social theory and knowledge. We turn to that task in the next two sections of this article.

SPECIFYING THE TREATMENT MODEL

The treatment model consists of the treatment-as-delivered characteristics, other related exogenous factors, intervening processes, and outcome variables along with the postulated relationships among these component parts. An appropriate treatment model lays out in detail how delivered treatments work (see Figure 1).

As shown in Figure 1 the treatment variables[2] are likely correlated with exogenous variables, which may also independently affect intervening processes and/or outcome variables. This is simply a formal statement of the truism that most social phenomena are the outcomes of many interrelated processes. For example, whether or not a person quits smoking may depend not only on a particular antismoking program to which he or she may be exposed, but also on factors such as the participant's personal health problems believed to be caused by smoking, whether family members and friends are smokers, and past smoking history, any or all of which may be correlated with participation in the program. Hence if we are modeling the outcome of an antismoking program, we have to take such exogenous factors "into account" in estimating the effects of that antismoking program upon participants. From a technical viewpoint, unbiased estimates of a program's effects can only be obtained when treatment variables are adequately purged of correlated exogenous variables.

It is the real and present danger that treatment variables are correlated with exogenous variables that affect intervening processes and outcome variables which make the randomized controlled experiment so attractive. By randomly assigning target units of a program (usually persons) to experimental and control groups, the naturally existing correlations between treatment variables and exogenous variables are forced to be essentially zero. If the only concern in a program evaluation is to obtain unbiased treatment effects, then a randomized controlled experiment that maintains its integrity need not be designed with any knowledge of the relationships among exogenous, treatment, intervening, and outcome variables. Hence randomized experiments can be designed as "black box" researches in which *how* treatments affect outcomes is unknown.

But black box randomized experiments are not the only realization of the experimental paradigm and, indeed, may often be an inefficient form of that paradigm. This arises because advocates of the black box experimental paradigm often neglect the fact that after randomization exogenous variables are still correlated with outcome variables. Knowing how such exogenous factors affect outcomes makes it possible to construct more precise estimates of experimental effects by controlling for such exogenous variables. For example, an experiment on the recidivism of released prisoners can estimate treatment effects with smaller standard errors by taking into account the fact that age, education, and previous work experiences of the released prisoners ordinarily affect tendencies to recidivate. For a given N, a randomized experiment that takes into account existing theory and knowledge can have considerably more power than a black box randomized experiment.[3]

The black box paradigm also dominates classical discussion of nonexperimental approaches (Campbell and Stanley, 1966). Such discussions center around what are the inherent dangers of black box quasi-experimental approaches. This may be appropriate if one is estimating the effectiveness of a program for which there is no underlying sensible rationale, but it is not sensible to ignore existing knowledge when its use can increase the power of the research design.

Indeed, at best, it may be possible to obtain unbiased estimates of effects from quasi-experimental approaches if one can model with some degree of accuracy the relationships among all the elements of the treatment model. For example, an evaluation of an unemployment insurance program in California (Rauma and Berk, 1982) was able to control for the exogenous factors that determined the size of a released prisoner's benefit eligibility, because such benefits were completely determined by the number of days worked while in prison.[4] By holding constant the number of days worked while in prison, it was possible to hold constant the exogenous factors that determined receiving the treatment and hence to construct unbiased estimates of the effects of the treatment on subsequent recidivism.

The general issue of controlling for self-selection bias has been discussed thoroughly in recent literature (Barnow et al., 1980) and more recently by Berk and Ray (1982). How successfully such approaches can be applied in particular cases is determined by how well known are the exogenous processes and how well they can be measured. Furthermore, it is somewhat obvious, but bears emphasis, that knowledge and theory

concerning the effects of exogenous processes need to be built into evaluations *ab initio* and not constructed *ad hoc* from the selections available in a given data set.

The dangers of black box quasi-experiments are real, but they flow from the fact that they are black box efforts and not from their quasi-experimental character. Theory-driven randomized and quasi-experiments both are superior to their black box counterparts in power and efficiency. At best the distinction between randomized experiments and quasi-experiments becomes blurred to the extent that correctly specified theory-driven treatment models are employed. This last statement has a number of important implications. Randomized experimental designs applied to field situations have a distressing tendency to deteriorate rapidly into quasi-experiments. For example, one may randomly assign persons to treatment groups, but if treatment acceptance depends even partially on target population cooperation, differential cooperation can easily change the research design into a quasi-experiment. Witness the effects that attrition rates have had on the income maintenance and housing allowance experiments (Watts and Rees, 1976). Randomized experiments are difficult to install and carry out except on proposed but not yet enacted programs. Existing, full-coverage programs can usually only be evaluated for impact assessment by quasi-experimental designs.

Finally, theory-driven treatment modeling can meet the objections of many evaluators who are concerned that programs once in place develop goals that replace those officially proclaimed by policymakers and program designers. The truism that every program has some effects can be given some substance if treatment modeling can be used to uncover them.

THE PROBLEM OF GENERALIZATION

A given social program ordinarily is a complex bundle of specific items lumped together as a treatment. Even very simple-appearing treatments can become quite complex in implementation. For example, although the transfer payments were conceived as the treatments in income maintenance or housing allowance experiments, the treatments as delivered consisted of the payments of varying amounts, methods of establishing and validating eligibility, housing inspections, and so on through the entire apparatus of the experiments that dealt directly with the families in the experiment. Without careful specification of the

treatment as delivered, interpretation of treatment effects may become very muddy indeed. More important, an experiment may be fatally flawed by confounding the intended treatment with administrative trappings that might nullify intended treatment effects. A priori analysis of the treatment as delivered should lead to an experimental design that can separate out the effects of various components of the treatment as delivered. A very good example exists again in the TARP experiment in which the administrative regulations of the unemployment benefit systems of the states of Georgia and Texas negated the beneficial effects of the payments (Rossi et al., 1980).

Of course any treatment as delivered can be broken down analytically into a very large number of identifiable components, the vast majority of which may have trivial impacts upon outcomes. Identifying the important compenents is again the task of applying a priori knowledge and theory. Thus in the income maintenance experiments, the guarantee level and the implicit tax rates were identified on the basis of microeconomic theories concerning labor force participation as critically important components and hence were systematically varied within the experimental design. Similarly in the housing allowance experiments, the use of housing standards as a criterion for eligibility was conceived to be an important device and hence built into the experimental design.

These considerations, it should be noted, apply with equal force to quasi-experiments, especially those in which the design of treatments can be influenced by the evaluation researcher.

One of the main benefits of departing from the black box treatment-as-unit approach to evaluation is an enhanced ability to generalize from the researches in question to other circumstances. The end result of a black box evaluation is to know whether or not a given treatment-as-unit is effective and to what extent it is so. A transfer into a different administrative environment and subsequent modifications to fit the requirements of that environment may drastically alter the treatment's effectiveness, if the elements changed are among the more important within the treatment-as-unit. Indeed, since the translation of a proposed program into an enacted program always requires modification to fit the administrative environment into which it is placed, as well as to the political acceptability constraints of the policymakers, it is important to be able to point out what are the essential and nonessential components of a proposed program.

MODELING INTERVENING PROCESSES

The main points made with respect to the modeling of the treatment processes and components of delivered treatments apply as well as to the modeling of intervening processes. Indeed any model of the treatment process necessarily includes modeling intervening processes. From some viewpoints it hardly makes any sense to distinguish intervening processes except that, for programs that may be expected to have very long time effects, whether or not intervening processes occur may be the first sign of whether or not a program is working. For example, if a manpower training program is to be installed to increase the earning power of participants over the long run, it may be useful as a first step to specify what has to change in the short run in order that the long-range effects of the desired sort may be eventually captured. Thus, if a training program does not increase the job-relevant skills of participants, it seems unlikely that long-run wages will also increase. In short, the specification of intervening processes provides the opportunities for the more sensitive testing of the effectiveness of programs and also for their redesign in the unhappy eventuality that postulated intervening steps do not occur.

RESPONSE FUNCTIONAL FORMS

Another issue in modeling centers around the functional forms that relate program variables to each other and to outcome variables. Recursive models postulate one-way relations among variables and nonrecursive models postulate that at least some of the relations involve reciprocal effects.

In program evaluation it is possible to find that modeling causal processes of the intervention requires postulating reciprocal relations among outcome variables and/or between the outcome variables and the intervening variables. For example, an educational program might affect students' test scores and self-esteem. However, if existing knowledge suggests that a reciprocal process exists between test scores and self-esteem, then a nonrecursive model should be proposed for evaluating this program.

Interactive effects may also be postulated in which treatment variables are differentially effective among subgroups of targets.

Interactions are sufficiently well known that evaluators routinely look for them, but the search for interactions should not be a matter of systematically testing out all possible interactions—a strategy that maximizes Type I errors—but one which looks for those interactions that one has a good a priori reason to suspect exist.

Finally, linear additive models of response effects are popular because they are both easy to compute and simple to interpret when found. But in some cases there may be good reason to suspect that polynomial models may be more appropriate. For example, increasing the amount of treatments may lead one to expect a point of maximum effect per unit of treatment with lower rates of return for points above and below the maximum. Thus transfer payments that are too small may not affect labor force supply at all, while transfer payments that are very large may not affect labor force supply any more than modest transfer payments, as the diminishing marginal returns formulation suggests.

Implementation Modeling

Implementation systems traditionally have not been given the amount of attention they fully deserve in evaluation research. As pointed out earlier, experimental evaluations of prospective programs involve setting up arrangements for delivering programs (or treatments); hence even programs set up for testing purposes by researchers involve implementation systems. Even more important is the fact that a program once enacted must be carried out through an implementation system that includes administrative rules and regulations, bureaucratic structures, and personnel who have been given the responsibility to administer the program in question.

An understanding of program implementation is important in program evaluation, since successful implementation is also a necessary condition in assessing program theory success. Only when treatment variables are implemented successfully, at least to some extent, can we test whether or not the treatment variables have had any impact upon outcome variables.

In the evaluation literature there has been no dearth of interest in implementation, but too much of the attention has been given to worrying about whether programs have been delivered as intended, and not enough attention has been given to understanding the process of

implementation. Thus Levine (1972) stated that the main problem of the War on Poverty was the failure of programs to be implemented in the field. Gramlich and Koshel (1975) found that the performance contracting experiments failed to the extent that they were not implemented (or implemented incorrectly) in the field.

Part of the problem of integrating a concern for implementation process into evaluation stems from the fact that evaluators tend to be specialists in the disciplines relevant to treatment processes. Thus an evaluator concerned with the outcome of educational programs usually knows a great deal about educational processes, but may know very little about theories of organization; hence the organizational contexts of the program may be neglected or unspecified.

In Figure 1 we have designated an implementation system as the organizational arrangement that is either specially designed to deliver treatments (or programs) or given the responsibility to do so. We do not mean to imply that this box represents a simple system. Indeed at least six subsystems have been identified in the existing literature (e.g., Van Meter and Van Horn, 1975; Williams and Elmore, 1976; Scheirer, 1981), and these are detailed below.

Implementing organization. An agency of some sort, either newly created for the purpose or already existing, is usually given the mandate to administer a program. Its characteristics, such as the particular type of authority structure, the composition of personnel, existing standard operating procedures, and the system of incentives employed to achieve coordination of activities among personnel and departments may all affect how much and what specific forms of a given treatment are delivered. For example, schools are considered to be loosely coupled systems in which component personnel (e.g., teachers) are not linked together into an extensive division of labor in which the work of one member is closely dependent in time on the work of another member. Hence the activities of teachers in their classrooms are notoriously difficult to control (and affect). In contrast, a public welfare agency in which caseworkers each handle only part of the treatment of a case may be more easily changed, since it is easier to detect caseworkers who are not performing according to plan.

Organizational theory is not the best developed of social sciences and tends to be heavily dominated by theories that were developed in connection with the study of industrial and business enterprises. The study of public sector organizations that process people rather than material objects has been relatively neglected.

Target groups. Every program defines a target population consisting of some human units—persons, households, communities, and so on—in which its effects are to be manifested in the form of specified changes. Target groups affect the implementation of programs to the extent that such implementation implies the cooperation, compliance, or participation of the groups in question. If targets, for whatever reasons, refuse to accept delivery of a program's treatment, clearly the program can have no effect. Participation rates, therefore, are extremely important characteristics of programs. For example, the fact that *Sesame Street* achieved so large a penetration of its intended audience of nursery-school-age children in poor families contributed greatly to its overall success, even though the effects of viewing on each child may have been relatively slight. In contrast, the failure of *Feeling Good* (Mielke and Swinehart, 1976) was largely caused by its inability to reach more than a very small proportion of its intended target of poor adults. Understanding the conditions under which targets of various sorts will or will not participate in programs may involve knowing how subgroups within the population receive information, subcultural beliefs concerning participation in similar programs, and so on.

Environmental context. Implementation takes place within an environment containing other organizations, competing activities and programs, political structures, and so on. All these exogenous contextual processes can affect whether or not a program can be effectively implemented. Thus the Community Action Program of the Office of Economic Opportunity was eventually handed over to local political control after mayors protested against the federal government's setting up independent political entities in their domains (Moynihan, 1969). Also, the fact that the health education television program *Feeling Good* was broadcast at prime viewing times means that it had to compete with extremely popular programs for the attention of its intended audiences (Mielke and Swinehart, 1976).

Characteristics of treatments. Some treatments are intrinsically difficult and others much easier to deliver. Perhaps the critical element is the extent to which the treatment is "operator-robust"—capable of being delivered relatively intact as intended, regardless of the activities of the persons in whose hands responsibility for delivery is given. Thus, at one extreme, transfer payments are relatively robust treatments, since there are limited numbers of ways in which checks can be delivered to

persons through the mail. At the other extreme, treatments that involve tailoring interventions to the characteristics of targets usually involve allowing considerable discretion to the frontline implementer, a circumstance that may considerably distort program intentions. Indeed for most human services programs the dilemma is what is the optimum level of discretion to be allowed? If too little discretion is allowed, inappropriate treatments may be administered to clients. If too much discretion is allowed, it may become very difficult to determine precisely what was delivered, as is the case with many educational programs designed to alter the teaching practices in the classroom.

Another characteristic of treatments that bears attention is the matter of dosage. Thus a transfer payment that amounts to $100 per week is simply worth a lot more than 100 times a transfer payment of $1 per week. Or, just becuase we know that three hours of counseling per week may help a client does not mean that one hour per week will simply do one-third less. The amount of an intervention, especially as actually delivered, ought to be an important concern of evaluators.

Resources. Obviously a program requires sufficient resources to enable it to accomplish the delivery of treatment. Funds are used to hire persons, physical facilities, and so on. An underfunded program simply will not be able to deliver the treatments as prescribed.

Interorganizational transactions. An implementing organization may have to deal with other organizations in order to be able to deliver treatments. For example, treatments that call for the cooperation of other organizations may not be able to function if cooperation is withheld; or an organization may be under the jurisdiction of a more superordinate organization whose command superiority may either interfere with or facilitate the implementation of the program.

All of the above characteristics of implementation systems may need to be taken into account in developing a model of implementation in particular cases. We cannot pretend that the construction of implementation models will be easy at this stage in our understanding of the public sector human services organizations. All that is clear to us is that neglect of understanding implementation has made it ambiguous in many cases of evaluation researches whether the program or the implementation system or both were at fault in a demonstrated failure to achieve outcomes.

CONCLUSIONS

This article presents a set of arguments for a new appraisal of the dominant experimental paradigm as applied to evaluations. A central feature of that paradigm as elaborated has been to emphasize black box randomized experiments and quasi-experiments. We have argued for a paradigm that accepts experiments and quasi-experiments as dominant research designs, but that emphasizes that these devices should be used in conjunction with a priori knowledge and theory to build models of the treatment process and implementation system to produce evaluations that are more efficient and that yield more information about how to achieve desired effects.

It should also be emphasized that this article does not argue for postponing evaluations until the most adequate theory and knowledge have been constructed. It argues, rather, that we make do with what we have, at least for the time being, drawing upon existing stocks of theory and knowledge to the extent relevant. We also make a special plea for more intensive attention to developing knowledge and theory concerning how human services organizations work, so that our general understanding of implementation systems will be advanced

In sum, we hope that what we have to say here will inspire evaluators to spend more effort on understanding how programs work than on the effort to find out whether or not they actually work in some specific and nongeneralizable instance.

NOTES

1. It has become increasingly clear with experience in social experimentation (as opposed to short-term laboratory experiments) that the beneficial effects of randomization can be undermined seriously by exogenous factors operating on the delivery of treatments, as exemplified in nonrandom attrition in treatment and control groups in the income maintenance experiments (Watts and Rees, 1976; Rossi and Lyall, 1976).

2. Treatment variables include treatment characteristics *as delivered.* In order to make the discussion in this article flow more smoothly, we will simply refer to "treatment factors" with the understanding that such terms refer to treatments as delivered.

3. A similar argument against black box experiments was made recently in a commentary on experiments on the effectiveness of seeding clouds with silver iodide crystals in order to produce rainfall (Kerr, 1982a, 1982b). The author made the point that without good knowledge of the processes that take place within clouds that ordinarily lead

to rainfall, the very expensive experiments were not powerful enough to detect treatment effects. A strong argument was then advanced against any additional black box experiments on the effects of cloud seeding.

4. Since prisoners did not know that their working would affect their eligibility for and amount of benefits (the legislation had been enacted at the time of imprisonment), they could not have worked in prison because they anticipated postrelease benefits. Of course, for subsequent cohorts of prisoners, the possibility of prison work being affected by anticipated benefits will have to be taken into account.

REFERENCES

BARNOW, B., G. G. CAIN, and A. GOLDBERGER (1980) "Issues in the analysis of selectivity bias," in E. W. Stormsdorfer and G. Farkas (eds.) Evaluation Studies Review Annual, Vol. 5. Beverly Hills, CA: Sage.

BERK, R. and S. C. RAY (1982) "Selection biases in sociological data." Social Science Research 11, 4: 352-398.

CAMPBELL, D. T. and J. C. STANLEY (1966) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

CHEN, H. and P. H. ROSSI (1980) "Multi-goal theory-driven approach: a model linking basic and applied social science." Social Forces 51, 1: 106-122.

CICIRELLI, V. G. and Associates (1969) The Impact of Head Start. Athens, OH: Westinghouse Learning Corporation and Ohio University.

COOK, T. D. and D. T. CAMPBELL (1979) Quasi-Experimentation: Design and Analysis Issues for Field Setting. Chicago: Rand McNally.

DEUTSCHER, I. (1977) "Toward avoiding the goal trap in evaluation research," in F. G. Caro (ed.) Readings in Evaluation Research. New York: Russell Sage.

GRAMLICH, E. M. and P. P. KOSHEL (1975) "Is real-world experimentation possible? The case of educational performance contracting." Policy Analysis 1 (Summer): 511-530.

GUBA, E. G. and Y. LINCOLN (1981) Effective Evaluation. San Francisco: Jossey-Bass.

KERR, R. A. (1982a) "Test fails to confirm cloud seeding effect." Science 217, 4557.

——— (1982b) "Cloud seeding: one success in 35 years." Science 217, 4559.

LEVINE, R. A. (1972) Public Planning: Failure and Redirections. New York: Basic Books.

MIELKE, K. W. and J. W. SWINEHART (1976) Evaluation of the Feeling Good Television Series. New York: Children's Television Workshop.

MORRIS, L. L., C. T. FITZ-GIBBON, and M. E. HENERSON (1978) Program Evaluation Kit. Beverly Hills, CA: Sage.

MOYNIHAN, D. P. (1969) Maximum Feasible Misunderstanding. New York: Free Press.

RAUMA, D. and R. BERK (1982) "Crime and poverty in California." Social Science Research 11, 4: 318-351.

RIECKEN, H. W. and R. F. BORUCH (1974) Social Experimentation—A Method for Planning and Evaluating Social Intervention. New York: Academic.

ROSSI, P. H. and K. LYALL (1976) Reforming Social Welfare. New York: Russell Sage.
ROSSI, P. H., R. A. BERK, and K. J. LENIHAN (1980) Money, Work and Crime. New York: Academic.
SCHEIRER, M. A. (1981) Program Implementation: The Organizational Context. Beverly Hills, CA: Sage.
SCRIVEN, M. (1972) "Pros and cons about goal-free evaluation." Evaluation Comment 3: 1-4.
SUCHMAN, E. A. (1969) "Evaluating educational programs." Urban Review 3, 4: 15-17.
VAN METER, D. S. and C. E. VAN HORN (1975) "The policy implementation process: a conceptual framework." Administration and Society 6, 4: 445-488.
WATTS, H. and A. REES (1976) The New Jersey Income Maintenance Experiment, Vols. 2 and 3. New York: Academic.
WHOLEY, J. S., J. N. NAY, and R. E. SCHMIDT (1975) "Evaluation: where is it really needed?" Evaluation Magazine 2, 2: 89-93.
WILLIAMS, W. (1976) "Implementation analysis and assessment," in W. Williams and R. F. Elmore (eds.) Social Program Implementation. New York: Academic.
——— and R. F. ELMORE [eds.] (1976) Social Program Implementation. New York: Academic.
WRIGHT, J. D., P. H. ROSSI, and K. DALY (1983) Under the Gun. Hawthorne, NY: Aldine.

17

The Expanding Scope of Alcoholism Treatment Evaluation

Rudolf H. Moos and John W. Finney

Recent advances in alcoholism treatment evaluation research have sparked new perspectives on the nature of alcoholism and the role of treatment and extratreatment factors in the recovery–relapse process. Combined with current trends in behavioral medicine and evaluation research, these perspectives are guiding the development of a conceptually based approach to treatment evaluation. This approach can help to integrate seemingly disparate facts about alcohol abuse, to improve clinical services by contributing to program development, and ultimately, to formulate a clearer understanding of the biopsychosocial nature of alcohol problems. In short, evaluation research is beginning to fulfill some of the promise it derives from its unique location at the interface of basic research on alcoholism and applied concerns with the development and delivery of treatment programs.

Alcohol abuse is often seen as an inexorably progressive endogenous disorder, with treatment outcome being predicted most adequately from predisposing factors (such as sociodemographic, biogenetic, and prior drinking characteristics) inherent in the individual. Investigators who adopt this conceptual approach typically find that these predisposing factors account for only a small proportion (usually less than 20%) of the variance in drinking and drinking-related outcome criteria (Moos, Cronkite, & Finney, 1982). Moreover, many individuals apparently "mature out" of serious problem drinking (Hyman, 1976), while others fluctuate between periods of alcohol abuse and nonabuse (Polich, Armor, & Braiker, 1981). Although predisposing factors can predict problem drinking following treatment, the most notable finding is that they are rather poor predictors. Such facts support the optimistic perspective that alcohol abuse can be treated successfully.

Recent findings from alcoholism program evaluations are consistent with this perspective. A substantial proportion of alcohol abusers improve after participating in behavioral treatments (Miller, Taylor, & West, 1980), medically based aversion-conditioning programs (Neubuerger, Hasha, Matarazzo, Schmitz, & Pratt, 1981), milieu-oriented and other comprehensive residential programs (Bromet, Moos, Bliss, & Wuthmann, 1977; Costello, Baillargean, Biever, & Bennett, 1980), outpatient and day treatment centers (Armor, Polich, & Stambul, 1978; McLachlan & Stein, 1982; Polich et al., 1981), and in the personal support networks provided by Alcoholics Anonymous (AA; Alford, 1980). Some successfully treated alcoholic persons show stable patterns of long-term recovery and are able to resume essentially normal patterns of functioning (Kurtines, Ball, & Wood, 1978; Moos, Finney, & Chan, 1981). Given these findings, it is not surprising that alcoholism treatment can be cost effective (Cicchinelli, Binner, & Halpern, 1978).

Many persons also recover from alcohol abuse without formal treatment. Estimates of the proportion of individuals in minimal-treatment control groups who improve range from 32% using relatively stringent criteria (Polich et al., 1981) to 53% using more relaxed criteria (Armor et al., 1978). In addition, about 10%–20% of problem drinkers who have not attempted to enter treatment recover "spontaneously" (Imber, Schultz, Funderburk, Allen, & Flamer, 1976), but such estimates depend on

From Rudolf H. Moos and John W. Finney, "The Expanding Scope of Alcoholism Treatment Evaluation," *American Psychologist*, 1983 38(10), 1036-1044.

the length of the follow-up period and criteria used to evaluate problem drinking and improvement.

These findings show that the personal and social problems that foster alcohol abuse need not lead to permanent deficits, that the stigma of alcoholism can be overcome, and that beneficial influences from treatment and extratreatment contexts can help some alcohol abusers to resume essentially normal lives. Formal treatment is neither necessary nor sufficient to effect long-term improvement (Mulford, 1977). However, treatment apparently facilitates the recovery process in that treated individuals show higher rates of improvement in many studies than do minimally treated or untreated comparison groups (Emrick, 1975; Polich et al., 1981; Armor, Note 1).

Another set of more pessimistic findings demands equal attention, however. Evaluations of alcoholism programs also have shown that a considerable number of patients drop out of treatment prematurely and, more importantly, that many patients are not helped by a single exposure to current treatments—Relapse rates during the year after the completion of treatment may be as high as 60% or more (Costello, 1975a). Moreover, researchers have not been very successful in identifying superior treatment methods or in finding treatment approaches that are particularly effective for specific types of patients. Even the idea that more treatment (longer treatment of greater intensity) is better than less treatment has not received much support (e.g., Edwards et al., 1977). Finally, a large number of persons do not recover "spontaneously," but continue to drink heavily and to incur substantial personal and social costs by doing so.

Such apparently divergent findings indicate that intervention programs and life-context factors can have a powerful impact on the course of alcoholism. By suggesting that this impact can be for better or for worse, they highlight a set of important issues: Why do some alcohol abusers respond positively to an intervention while others show little or no response and quickly resume problem drinking? In what ways do the characteristics of an individual's life context foster or inhibit the recovery process? How do patient, intervention, and life-context factors interrelate to affect recovery and relapse? These issues are beginning to be addressed within the context of a conceptually based, process-oriented evaluation framework.

A Process-Oriented Model for Alcoholism Evaluation Research

A process-oriented framework for alcoholism treatment evaluation draws on current trends in evaluation research and behavioral medicine. Until recently, most evaluation researchers were guided by an idealized paradigm in which individuals were assessed, assigned to treatment (or control) conditions, and then reevaluated at follow-up to identify treatment-related changes in their functioning and behavior. This "summative" paradigm is being expanded in several ways. Since intervention programs typically are neither implemented as planned nor delivered to recipients in a fixed, standard manner, one area of development is an emphasis on treatment implementation. Given that treatment often varies across patients, researchers are beginning to develop a more differentiated view of treatment processes and to explore the relationship between treatment factors and outcome. Evaluators are also realizing that powerful extratreatment or life-context factors can affect the relative benefits of intervention programs. Thus, there is growing recognition that evaluation research is more than just a technical enterprise; such research can help to formulate conceptual issues and produces its greatest yield when it is grounded within a conceptual framework (Cronbach, 1982).

New trends in behavioral medicine involve the development of a systems orientation and an interdisciplinary biopsychosocial perspective (Schwartz, 1982). Building on an ecological perspective in psychosomatic medicine, these trends emphasize the need to consider links between biological, psychological, and environmental factors in assessing an individual's health and making recommendations for treatment. This contextual, multicausal approach provides a framework that enables clinicians and evaluators to consider an individual's overall life situation in planning and evaluating an intervention. It can also help to overcome the unsatisfactory mind–body dualism that besets alcoholism research and to foster the eventual integration of biological and psychosocial perspectives.

The framework for alcoholism evaluation research shown in Figure 1 embodies these trends in two ways. First, it reflects an emphasis on a better understanding of treatment, that is, on documenting the implementation and delivery of treatments and on assessing the quality of treatment processes. Second, it considers life-context or extratreatment factors as additional determinants of treatment entry, duration, and outcome. This approach acknowledges the fact that the traditional summative evaluation model does not adequately capture the complexity of the alcoholism treatment–rehabilitation

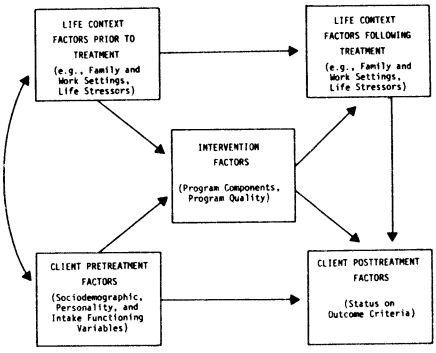


Figure 1
A Process-Oriented Framework for Evaluation of Alcoholism Treatment

process and that an intervention program is only one among many sets of factors that influence subsequent adaptation (Cronkite & Moos, 1980; Moos et al., 1982). The framework can be used to examine the effects of treatments that range from pharmacological interventions (such as disulfiram) to community "treatment" (such as AA), as well as to identify factors involved in recovery without treatment. In the following sections, we will use this paradigm to illustrate advances in alcoholism evaluations in (a) documenting treatment implementation, (b) analyzing treatment processes, and (c) examining the role of extratreatment factors.

Treatment Implementation

The intervention program lies at the heart of any study of alcoholism treatment. Information about the treatment as actually delivered enables evaluators to identify aspects of interventions that have been only partially implemented or operationalized in a form different from that intended. Examination of treatment implementation, or what Suchman (1967) refers to as the evaluation of "effort," focuses on the quantity and quality of treatment activity. This assessment can be accomplished either by documenting what was done by treatment providers or by demonstrating that treatment produced in clients intermediate changes presumed to foster the ultimate positive outcome. In either case, to estimate the degree of treatment implementation, evaluators must determine the congruence between the intervention as actually conducted or responded to and the intervention as it was intended to be applied or experienced. Relevant standards for assessing treatment quantity and quality have been developed from (a) information about conditions in other programs, and (b) theoretical analysis and expert judgment (Sechrest, West, Phillips, Redner, & Yeaton, 1979).

Normative Comparisons

Some investigators have compared indexes of treatment activity, such as the length of treatment and the "delivery" of specific treatment components, to normative standards in successful programs. For example, we found that the average length of patient stay in three residential alcoholism programs varied from about four weeks in one program to more than three months in the other two (Finney, Moos, & Chan, 1981). This latter duration of treatment compared favorably with the median of six to eight weeks found in two groups of programs in which patients exhibited relatively high rates of positive treatment outcome at one- and two-year follow-ups (Costello, 1975a, 1975b). However, a long stay does not necessarily imply that patients are "treated" more intensively. Clients in a milieu-oriented program we studied participated in more therapy sessions, AA meetings, and lectures on alcoholism during their average 30-day stay than did clients in a halfway house and a Salvation Army program during an average stay of over three months. Lack of intensity of treatment may explain why length of stay is not related to outcome in some programs (Finney et al., 1981).

Whereas information on treatment components taps the quantity of treatment activities, "treatment quality" refers to the manner in which such activities are conducted. One useful indicator of treatment quality is provided by data on the "social climate" of treatment settings, such as is supplied by two conceptually parallel scales that assess hospital-based and community-based treatment programs. The Ward Atmosphere Scale (WAS) and the Community-Oriented Programs Environment Scale (COPES) describe treatment settings along 10 dimensions that focus on the quality of interpersonal relationships, the emphasis on such treatment goals as enhancing autonomy and self-understanding, and the degree to which the setting maintains stability and is open to change. Clients and staff judge the quality of their program on these dimensions, and these judgments can be tracked over time and compared with information obtained from normative samples of other treatment programs (Bromet, Moos, & Bliss, 1976; Moos, 1974).

Theory and Expert Judgments

Standards for evaluating treatment implementation can also be formulated by consulting theory or using expert judgments about the implications of a particular therapeutic approach. Learning theory (broadly defined) has provided implementation

standards for several evaluations of alcoholism treatment. For example, Elkins (1980) reported that covert sensitization (verbal aversion) actually resulted in conditioned nausea in less than half of the patients to whom it was applied. Conditioned individuals were less likely to have relapsed at follow-up than were nonconditioned persons.

Sanchez-Craig and Walker (1982) found no differences at 6-, 12-, or 18-month follow-ups between halfway house residents taught a five-step problem-solving process and two other groups (covert sensitization and discussion only). During treatment, problem-solving skills were taught to criterion so that there was no variation in implementation across individuals. One month after training, however, only 2 of 15 residents (13.3%) could recall each of the five steps; another 3 residents were able to repeat four of the five steps. The authors conclude that adequate implementation of this treatment approach will require taking into account the cognitive deficits experienced by many alcoholic persons.

The use of expert judgment can be illustrated with data from a social climate measure. After reviewing the relevant literature, Price and Moos (1975) concluded that one of the six profiles identified in a cluster analysis of treatment program climates constituted a "therapeutic community" (TC) type. Subsequently, as part of their evaluations, Steiner, Haldipur, and Stack (1982) and Bell (Note 2) carried out implementation assessments by examining the goodness of fit between the social-climate profiles of their treatment units and the TC type identified by Price and Moos. One potential consequence of inadequate treatment implementation is illustrated in the study by Bell (Note 2) in which the program whose profile failed to measure up most consistently to the TC type had the highest dropout rate.

Value of Implementation Assessment

Implementation assessment is important no matter what the form of treatment. For example, what could be a more direct treatment than disulfiram (Antabuse) when it is subcutaneously implanted? Nevertheless, when Malcolm, Madden, and Williams (1974) checked blood samples for evidence of disulfiram, they found that only 8 of 31 samples taken within one week of implantation were positive. Moreover, Bergström, Öhlin, Lindblom, and Wadstein (1982) found no evidence in blood samples of the expected aldehyde dehydrogenase inhibition and acetaldehyde increase 30 minutes after ethanol was ingested by 11 chronic alcoholics who had received implants. Thus, few (if any) treatments are so direct or obvious that their implementation can be taken for granted.

An evaluation of treatment implementation should consider the strength as well as the integrity of treatment. Yeaton and Sechrest (1981) have defined treatment integrity as "the degree to which treatment is delivered as intended" (p. 160), whereas treatment strength is "the a priori likelihood that the treatment could have the intended outcome" (p. 156). Strength and integrity are related to the extent that a treatment with low integrity is not likely to be very strong. Both concepts reflect attempts to gauge patient prognosis on the basis of information about the treatment in addition to information about the patient. These ideas expose logical fallacies in the standard pessimistic interpretations of evaluation results in the alcoholism field. Although a few sessions of outpatient treatment may have a positive effect on an alcoholic's functioning after six months or one year, they cannot reasonably be expected to have a measurable impact four years later (Polich et al., 1981). Such considerations argue for matching the intensity of treatment to the severity of clients' drinking problems, and suggest that "strong" treatments may be effective for "low bottom" chronic alcoholics (Sheehan, Wieman, & Bechtel, 1981).

Process Analyses: Linking Program Components to Treatment Outcome

Current standards that can be used to evaluate the implementation of alcoholism treatment programs are still largely a priori. A firmer footing for such standards is being developed in the search for effective treatment elements via "treatment process" analyses. Treatment process analyses focus on the entire causal chain between treatment and outcome (Judd & Kenny, 1981). One overarching linkage in that chain is the relationship between treatment as actually implemented and outcome. Such relationships have been examined by meta-analyses in which the presence or absence of treatment elements is related to aggregate outcome, by studies pitting a standard treatment regimen against one augmented with a new treatment component, and by analyses in which individual involvement in specific aspects of treatment is associated with outcome.

Treatment Components and Treatment Quantity

Costello (1980; Costello et al., 1980) conducted an integrated set of studies that illustrate the utility of meta-process-analysis for program planning and validating treatment implementation standards. Costello (1975a; 1975b) first examined treatment outcome in 58 studies that provided one-year and 23 studies that provided two-year follow-ups. Programs with the best outcome had an active, intensive milieu orientation, a moderate length of inpatient stay (median 6–8 weeks), considerable use of disulfiram,

behavior therapy with feedback, outpatient aftercare following hospital discharge, and active involvement of relatives and employers in the treatment process. The findings were used to plan an intensive inpatient alcoholism program that included these "strong" components (Costello et al., 1980). One- and two-year follow-ups of patients discharged from this program showed that about 40% could be considered treatment successes. This proportion was comparable to the best outcome standard identified among prior studies for similar patient populations and thus met normative expectations.

An experimental approach to process analysis can be illustrated by the work of Chaney, O'Leary, and Marlatt (1978) on short-term skills training during inpatient treatment. Adequate treatment integrity was shown when the experimental group performed better on a role-playing measure of skills to cope with relapse-inducing situations in comparison with discussion and no-additional-treatment control groups. At a one-year follow-up, the skills training group exhibited shorter and less severe relapse episodes than did patients in the control groups.

We conducted within-program analyses of the effects of length of stay (LOS) and three treatment components (therapy sessions, AA meetings, and films and lectures on alcoholism). Although the associations of individual program components and LOS with outcome measures were weak to moderate, the four treatment-experience variables together accounted for a significant proportion of the variance in outcome after patients' background characteristics and intake symptoms were controlled (Finney et al., 1981).

Treatment Quality

The quality of the client–counselor relationship and of the treatment milieu plays an important role in determining the effectiveness of an intervention. For example, in a recent evaluation of focused versus broad spectrum behavior therapy for problem drinkers, the degree of therapist empathy proved to be a good predictor of treatment outcome (Miller et al., 1980). Counselors with higher levels of interpersonal functioning achieved better treatment outcomes among their alcoholic clients in another study (Valle, 1981).

At the broader, milieu level, our studies indicate that the perceived quality of alcoholism treatment programs is one of the important predictors of six-month outcome, relative to patient factors (demographic characteristics and intake symptoms) and other treatment factors (type of program and participation in treatment components). Patients who saw their program as more involving, cohesive, well organized, and oriented toward independence and self-understanding tended to do better on drinking-related outcome criteria even after considering their prior drinking history and other personal factors. Since the links between the treatment environment perceptions and posttreatment functioning were independent of those for treatment components, the findings show that the quantity and quality of treatment can have independent effects on outcome (Cronkite & Moos, 1978). Treatment quality also may contribute to better outcome by increasing the intensity of client involvement in treatment and aftercare services.

Extratreatment Factors

The evaluation paradigm we have described explicitly assumes that treatment is part of an "open system." An intervention program is but one (indeed, a temporary one) of the multiple environmental microsystems or specific settings in which a client is involved. During treatment, and even more directly afterward, a client is exposed to a myriad of influences emanating from other, more enduring microsystems such as family and work environments.

Family and Work Settings and Treatment Outcome

Research on the relationship between family milieu and treatment outcome has indicated that the more cohesive and supportive the family, the better the patient's prognosis. For example, Orford, Oppenheimer, Egert, Hensman, and Guthrie (1976) noted that information about marital cohesion obtained at intake to treatment predicted posttreatment status on drinking indexes assessed 12 months later. We found that patients located in families characterized by more cohesion and less conflict tended to function better after treatment. These relationships held when family functioning dimensions assessed six months after treatment were used to predict patient adaptation at a two-year follow-up (Finney, Moos, & Mewborn, 1980). However, female patients more often than men report poor marital relationships and have spouses who encourage them to drink and who drink heavily themselves (Cronkite & Moos, in press). These studies provide substantial support for broadening the provision of family treatment.

An individual's work environment can also affect the outcome of an intervention. Pretreatment job satisfaction has been positively related to outcome among alcoholic patients assigned to reality therapy or self-awareness therapy (Ward, Bendel, & Lange, 1982). In our alcoholism treatment study, we found little or no connection between the characteristics of work settings and follow-up functioning among patients who returned to families after treatment. Among individuals not living in families,

however, those who saw their work milieu as more involving and cohesive, and their supervisor as more supportive, experienced better treatment outcome even after demographic factors and intake symptoms were considered. Location in family settings may help to cushion the adverse impact of stressful conditions in the workplace (Bromet & Moos, 1977).

Life Stressors and Relapse Episodes

Recent studies of the precipitants of relapse episodes have highlighted the importance of stressful life situations. Marlatt and Gordon (1979) point out that a high proportion of relapses occur within the first 90 days after treatment and that many of these relapses stem from exposure to such situations as being confronted with social pressure to drink or interpersonal conflicts that generate feelings of frustration and anger. In this regard, we found that negative life events (such as the death of a friend and economic or legal problems) were more prevalent among relapsed alcoholics than among recovered patients or demographically matched community controls (Moos et al., 1981). Moreover, the number of negative events that occurred during the first six months after treatment predicted treatment outcome on two criteria (complaints of physical symptoms and depression) at a two-year follow-up even after sociodemographic and intake functioning factors were controlled (Finney et al., 1980).

Strength and Generality of Contextual Factors

Although comparatively little is known about the impact of extratreatment or life-context factors on alcohol abuse, current evidence indicates that this is a highly promising area. In this regard, we formulated a conceptual model that considers the domains of extratreatment factors just described (family and work settings and stressful events) in conjunction with patient and treatment factors (Cronkite & Moos, 1980). The extratreatment factors accounted for an increment of between 7% and 27% of the variance in treatment outcome (depending on the specific criterion), compared with between 4% and 20% accounted for by patient-related and treatment-related factors. In short, the inclusion of extratreatment factors in the model more than doubled the explained variance in treatment outcome. These findings suggest that alcoholism treatment may be more effective when oriented toward patients' ongoing life circumstances.

The life-context factors we have described can affect the outcome of "informal" interventions such as AA as well as the likelihood of "spontaneous" remission. Part of the success of AA is due to the provision of extensive personal support networks and to norms that foster a less stressful life-style, both of which reduce the risk of relapse. Positive changes in family and occupational factors, such as marrying a spouse who exerts effective social controls on alcohol intake or entering an occupation with a low risk for problem drinking, can foster the process of recovery from alcoholism among untreated individuals (Saunders & Kershaw, 1979; Tuchfeld, 1981).

The suggestion of a continuity in the beneficial aspects of treatment and extratreatment factors points to the value of thinking of "treatment" in broad terms and attempting to integrate formal and informal treatment resources. In this regard, evidence from studies of both treatment and community settings indicates that moderately cohesive, well-organized environments that emphasize one or more areas of personal growth tend to have beneficial impact for relatively well-adjusted persons. For more disturbed individuals, however, somewhat more structured and less pressured, less expression-oriented settings may be most helpful (Cronkite, Moos, & Finney, in press). The identification of convergent cross-setting effects could significantly enhance the formulation of more integrated and powerful intervention programs.

Implications for Research and Practice

We have highlighted the trends in evaluation of alcoholism programs to include documentation of treatment implementation, process analyses that link program components to outcome, and analyses of the role of extratreatment factors. Respectively, these trends reflect the descriptive, evaluative, and model-building activities that Rossi (1983) argues are involved in virtually all social and behavioral science research. These advances have important implications for the utilization of evaluation research, both "instrumentally" in developing and improving intervention programs and "conceptually" in changing the way people think about the underlying disorder or problem to which treatment is directed.

Monitoring and Improving Intervention Programs

The evaluation paradigm we have described emphasizes the value of ongoing feedback of evaluation findings in the development and reformulation of intervention programs. By monitoring the intensity and quality of treatment, evaluators can provide clinical staff with important information about the adequacy of program implementation. If the treatment is being delivered and received as intended, the relationship between specific treatment components and (especially immediate) outcomes can be explored. At this point the evaluator can help clinicians to reorient the program and concentrate its resources on those components associated with better outcome. This perspective assumes that treat-

ment evaluation research is an integral component of the provision of effective clinical services.

Information about the quality of the treatment setting can serve a monitoring or quality control function. For instance, the WAS and COPES can be used to describe and compare alcoholism programs, to identify client–staff discrepancies in program perceptions, to highlight differences between actual and preferred treatment settings, and to trace the evolution and function of a program over time (Ryan, Bell, & Metcalf, 1982). When the intensity or quality of a program does not meet acceptable standards, changes are likely to be suggested. Evaluators can monitor and provide information about the effects of changes planned by staff, or they can take a more active part in the process of initiating and facilitating change. Such feedback–change procedures have been used to reduce the stressfulness of staff work environments (Koran, Moos, Moos, & Zasslow, 1983), as well as to improve the quality of treatment settings for patients (Moos, 1974, 1979). This process of data-based feedback and change is the essence of formative program evaluation.

Extratreatment Factors and the Reformulation of Treatment

Information about the causal mechanisms through which a treatment exerts its effects—including the extratreatment factors that enhance or impede positive outcome—can help to generate new and potentially more effective intervention strategies. Recognizing that stressful or relapse-inducing life situations inevitably occur, researchers have begun to identify coping resources that clients can acquire to help them deal with these situations more effectively. Litman, Eiser, Rawson, and Oppenheim (1979) found that patients who did not relapse used a varied set of cognitive coping styles and possessed a flexibility that enabled them to handle a variety of difficult situations. Marlatt (1982) is using a cognitive social-learning model to develop individualized intervention strategies that reduce the probability of relapse episodes and help patients to handle them effectively when they occur. In another approach, Azrin (1976) has successfully employed a "community reinforcement" program to restructure alcoholic patients' marital, occupational, and community resources. These developments are part of a general trend to consider psychiatric patients' life situations in planning interventions (Ryback, Longabaugh, & Fowler, 1981). They underscore the inadequacy of the model of intensive residential treatment during which the endogenous "disease" is treated, followed by occasional aftercare or "check-up" visits. In fact, Litman (1980) has suggested a restructuring of treatment in which hospitalization, if necessary, is followed by intensive outpatient treatment (not "aftercare") geared specifically toward the prevention of relapse.

Matching Patients With Treatments

There is an implicit "uniformity" assumption in much of the research on treatment of alcoholism. Homogeneity of patients is implied in the quest for the best treatment method, or in referring to the treatment or recovery–relapse process. As noted earlier, however, there is considerable heterogeneity in patients' response to treatment. Thus, enthusiasm has developed for matching different types of alcoholic patients with the most appropriate form of treatment. The evaluation framework outlined here has implications for such matching efforts and for research to evaluate their effectiveness.

Scattered studies of the matching hypothesis have been carried out during the past 25 years. One of the few conceptually guided efforts was that by McLachlan (1974) who used conceptual systems theory to identify relevant cognitive capabilities and styles of alcoholic patients and to characterize the different degrees of "structure" to which they were exposed during inpatient treatment and aftercare. No main effects were found for patient conceptual level (CL) or treatment structure; however, there was an important interaction effect. Of patients matched with both inpatient and aftercare services (that is, low CL patients with structured treatments; high CL patients with less structured treatments), 77% were rated as "recovered" at a 12–16 month follow-up, whereas for mismatched patients, only 38% were in the "recovered" category. The expanded evaluation framework we have described suggests that capitalizing on cross-setting effects by modifying the "structure" of clients' extratreatment environments might produce even better results. For example, the organization of low CL clients' life situations could be increased using procedures such as those employed by Azrin (1976) in his "community reinforcement" program.

Program Evaluation and Conceptual (Re)formulations

Evaluations can help to shape or reorient the thinking of policymakers and program developers about the problem or behavior that required the intervention in the first place (Cronbach, 1982). As an example, the Rand Corporation's evaluation of alcoholism treatment (Polich et al., 1981) provided new descriptive information on the natural course of alcoholism. Consistent with earlier findings on problem drinkers, it was discovered that the drinking behavior of many severely symptomatic alcoholics fluctuated widely over time between high and low levels of alcohol consumption. An evaluation grounded in an appropriate conceptual framework

could offer an explanation of within-individual variability in drinking behavior by linking these behavioral changes with intraindividual variation in exposure to such extratreatment factors as social pressure to drink, social controls over drinking, environmental stress, and social support.

Future Directions

Some individuals who suffer from alcohol abuse recover and attain adequate or normal levels of functioning. Such recoveries may be due to the effects of psychiatric or other types of formal treatment, of organized personal support networks such as are provided by AA, of stressors and resources that exist in individuals' everyday life contexts, or (most likely) of some combination of these sets of factors. A process-oriented evaluation framework affords the opportunity to identify more effective intervention processes, to understand the mechanisms through which extratreatment factors contribute to recovery and relapse, and to develop an enriched data base with which to plan more powerful treatment programs oriented toward clients' normal life situations.

Alcoholism treatment evaluation is being integrated into the mainstream of basic research on alcoholism and alcohol abuse. Analogous conceptual developments are taking place in related areas such as smoking (Shiffman, 1982) and drug abuse (Krueger, 1981). As the resulting new knowledge is organized within a biopsychosocial perspective, evaluation research will make more substantive contributions to an understanding of the underlying nature of alcohol problems. Hopefully, these developments will lead to the formulation of more effective pharmacological and psychosocial interventions and, ultimately, to knowledge that can be applied in primary as well as secondary and tertiary prevention.

REFERENCE NOTES

1. Armor, D. J. *A comparison of outcomes for treated and untreated alcoholics.* Paper presented at the meeting of the American Public Health Association, Los Angeles, November 1981.
2. Bell, M. *The perceived environments of three therapeutic communities with the same treatment model for drug abusers.* Unpublished research report, Veterans Administration Medical Center, Psychology Service, West Haven, Connecticut, 1979.

REFERENCES

Alford, G. S. Alcoholics Anonymous: An empirical outcome study. *Addictive Behaviors,* 1980, *5,* 359–370.

Armor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment.* New York: Wiley, 1978.

Azrin, W. H. Improvements in the community-reinforcement approach to alcoholism. *Behaviour Research and Therapy,* 1976, *14,* 339–348.

Bergström, B., Öhlin, H., Lindblom, P. E., & Wadstein, J. Is disulfiram implantation effective? *Lancet,* 1982, *1,* 49–50.

Bromet, E., & Moos, R. H. Environmental resources and the posttreatment functioning of alcoholic patients. *Journal of Health and Social Behavior,* 1977, *18,* 326–338.

Bromet, E. J., Moos, R. H., & Bliss, F. The social climate of alcoholism treatment programs. *Archives of General Psychiatry,* 1976, *33,* 910–916.

Bromet, E., Moos, R., Bliss, F., & Wuthmann, C. Posttreatment functioning of alcoholic patients: Its relation to program participation. *Journal of Consulting and Clinical Psychology,* 1977, *45,* 829–842.

Chaney, E. F., O'Leary, M. R., & Marlatt, G. A. Skill training with alcoholics. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 1092–1104.

Cicchinelli, L. F., Binner, P. R., & Halpern, J. Output value analysis of an alcoholism treatment program. *Journal of Studies on Alcohol,* 1978, *39,* 435–447.

Costello, R. M. Alcoholism treatment and evaluation: In search of methods. *International Journal of the Addictions,* 1975, *10,* 251–275. (a)

Costello, R. M. Alcoholism treatment and evaluation: In search of methods: II. Collation of two-year follow-up studies. *International Journal of the Addictions,* 1975, *10,* 251–275. (b)

Costello, R. M. Alcoholism treatment effectiveness: Slicing the outcome variance pie. In G. Edwards & M. Grant (Eds.), *Alcoholism treatment in transition.* Baltimore, Md.: University Park Press, 1980.

Costello, R. M., Baillargeon, J. G., Biever, P., & Bennett, R. Therapeutic community treatment for alcohol abusers: A one-year multivariate outcome evaluation. *International Journal of the Addictions,* 1980, *15,* 215–232.

Cronbach, L. J. *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass, 1982.

Cronkite, R. C., & Moos, R. H. Evaluating alcoholism treatment programs: An integrated approach. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 1105–1119.

Cronkite, R. C., & Moos, R. H. Determinants of the posttreatment functioning of alcoholic patients: A conceptual framework. *Journal of Consulting and Clinical Psychology,* 1980, *48,* 305–316.

Cronkite, R. C., & Moos, R. H. Sex and marital status in relation to the treatment and outcome of alcoholic patients. *Sex Roles,* in press.

Cronkite, R. C., Moos, R. H., & Finney, J. W. Contexts of adaptation: An integrative perspective on community and treatment settings. In W. A. O'Connor & B. Lubin (Eds.), *Ecological models in clinical and community mental health.* New York: Wiley, in press.

Edwards, G., Orford, J., Egert, S., Guthrie, S., Hawker, A., Hensman, C., Mitcheson, M., Oppenheimer, E., & Taylor, C. Alcoholism: A controlled trial of treatment versus "advice." *Journal of Studies on Alcohol,* 1977, *38,* 1004–1031.

Elkins, R. L. Covert sensitization treatment of alcoholism: Contributions of successful conditioning to subsequent abstinence maintenance. *Addictive Behaviors,* 1980, *5,* 67–89.

Emrick, C. A review of psychologically oriented treatment of alcoholism: II. The relative effectiveness of different treatment approaches and the effectiveness of treatment versus no treatment. *Journal of Studies on Alcohol,* 1975, *36,* 88–108.

Finney, J. W., Moos, R. H., & Chan, D. A. Length of stay and program component effects in the treatment of alcoholism: A comparison of two techniques for process analyses. *Journal of Consulting and Clinical Psychology,* 1981, *49,* 120–131.

Finney, J. W., Moos, R. H., & Mewborn, C. R. Posttreatment experiences and treatment outcome of alcoholic patients six months and two years after hospitalization. *Journal of Consulting and Clinical Psychology,* 1980, *48,* 17–29.

Hyman, M. M. Alcoholics 15 years later. *Annals of the New York Academy of Sciences,* 1976, *273,* 613–623.

Imber, S., Schultz, E., Funderburk, F., Allen, R., & Flamer, R. The fate of the untreated alcoholic. *Journal of Nervous and Mental Disease,* 1976, *162,* 238–247.

Judd, C. M., & Kenny, D. A. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review,* 1981, *5,* 602–619.

Koran, L. M., Moos, R. H., Moos, B., & Zasslow, M. Changing hospital work environments: An example of a burn unit. *General Hospital Psychiatry,* 1983, *5,* 7–13.

Krueger, D. W. Stressful life events and the return to heroin use. *Journal of Human Stress,* 1981, *7,* 3–8.

Kurtines, W. M., Ball, L. R., & Wood, G. H. Personality characteristics of long-term recovered alcoholics: A comparative analysis. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 971–977.

Litman, G. K. Relapse in alcoholism: Traditional and current approaches. In G. Edwards & M. Grant (Eds.), *Alcohol treatment in transition.* Baltimore, Md.: University Park Press, 1980.

Litman, G., Eiser, J., Rawson, N., & Oppenheim, A. Differences in relapse precipitants and coping behavior between alcoholic relapsers and survivors. *Behavior Research and Therapy,* 1979, *17,* 89–94.

Malcolm, M. T., Madden, J. S., & Williams, A. E. Disulfiram implantation critically evaluated. *British Journal of Psychiatry,* 1974, *125,* 485–489.

Marlatt, G. A. Relapse prevention: A self-control program for the treatment of addictive behaviors. In R. B. Stuart (Ed.), *Adherence, compliance and generalization in behavioral medicine.* New York: Brunner/Mazel, 1982.

Marlatt, G. A., & Gordon, J. R. Determinants of relapse: Implications for the maintenance of behavior change. In P. Davidson (Ed.), *Behavioral medicine: Changing health life styles.* New York: Brunner/Mazel, 1979.

McLachlan, J. Therapy strategies, personality orientation, and recovery from alcoholism. *Canadian Psychiatric Association Journal,* 1974, *19,* 25–30.

McLachlan, J. F. C., & Stein, R. L. Evaluation of a day clinic for alcoholics. *Journal of Studies on Alcohol,* 1982, *43,* 261–272.

Miller, W. R., Taylor, C. A., & West, J. C. Focused versus broad-spectrum behavior therapy for problem drinkers. *Journal of Consulting and Clinical Psychology,* 1980, *48,* 590–601.

Moos, R. H. *Evaluating treatment environments.* New York: Wiley, 1974.

Moos, R. H. Improving social settings by social climate measurement and feedback. In R. F. Munoz, L. R. Snowden, & J. G. Kelly (Eds.), *Social and psychological research in community settings.* San Francisco: Jossey-Bass, 1979.

Moos, R. H., Cronkite, R. C., & Finney, J. W. A conceptual framework for alcoholism treatment evaluation. In E. M. Pattison & E. Kaufman (Eds.), *Encyclopedic handbook of alcoholism.* New York: Gardner, 1982.

Moos, R. H., Finney, J. W., & Chan, D. A. The process of recovery from alcoholism: I. Comparing alcoholic patients with matched community controls. *Journal of Studies on Alcohol,* 1981, *42,* 383–402.

Mulford, H. Stages in the alcoholic process: Toward a cumulative nonsequential index. *Journal of Studies on Alcohol,* 1977, *38,* 563–583.

Neubuerger, O. W., Hasha, N., Matarazzo, J. D., Schmitz, R. E., & Pratt, H. H. Behavioral chemical treatment of alcoholism: An outcome replication. *Journal of Studies on Alcohol,* 1981, *42,* 806–810.

Orford, J., Oppenheimer, E., Egert, S., Hensman, C., & Guthrie, S. The cohesiveness of alcoholism-complicated marriages and its influence on treatment outcome. *British Journal of Psychiatry,* 1976, *128,* 318–339.

Polich, J. M., Armor, D. J., & Braiker, H. B. *The course of alcoholism: Four years after treatment.* New York: Wiley, 1981.

Price, R. H., & Moos, R. H. Toward a taxonomy of inpatient treatment environments. *Journal of Abnormal Psychology,* 1975, *84,* 181–188.

Rossi, P. H. Pussycats, weasels or percherons? Current prospects for social science under the Reagan regime. *Evaluation News,* 1983, *4,* 12–27.

Ryan, E., Bell, M., & Metcalf, J. The development of a rehabilitation psychology program for schizophrenics: Changes in the treatment environment. *Journal of Rehabilitative Psychology,* 1982, *27,* 67–85.

Ryback, R., Longabaugh, R., & Fowler, D. R. *The Problem Oriented Record in psychiatry and mental health.* New York: Grune & Stratton, 1981.

Sanchez-Craig, M., & Walker, K. Teaching coping skills to chronic alcoholics in a coeducational halfway house: I. Assessment of programme effects. *British Journal of Addiction,* 1982, *77,* 35–50.

Saunders, W. M., & Kershaw, P. W. Spontaneous remission from alcoholism—A community study. *British Journal of Addiction,* 1979, *74,* 251–265.

Schwartz, G. E. Testing the biopsychosocial model: The ultimate challenge facing behavioral medicine. *Journal of Consulting and Clinical Psychology,* 1982, *50,* 1040–1053.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4). Beverly Hills, Calif.: Sage, 1979.

Sheehan, J. J., Wieman, R. J., & Bechtel, J. E. Follow-up of a twelve-month treatment program for chronic alcoholics. *International Journal of the Addictions,* 1981, *16,* 233–241.

Shiffman, S. Relapse following smoking cessation: A situational analysis. *Journal of Consulting and Clinical Psychology,* 1982, *50,* 71–86.

Steiner, H., Haldipur, C. V., & Stack, L. C. The acute admission ward as a therapeutic community. *American Journal of Psychiatry,* 1982, *139,* 897–901.

Suchman, E. A. *Evaluative research.* New York: Russell Sage, 1967.

Tuchfeld, B. S. Spontaneous remission in alcoholics: Empirical observations and theoretical implications. *Journal of Studies on Alcohol,* 1981, *42,* 626–641.

Valle, S. K. Interpersonal functioning of alcoholism counselors and treatment outcome. *Journal of Studies on Alcohol,* 1981, *42,* 783–790.

Ward, D. A., Bendel, R. B., & Lange, D. A reconsideration of environmental resources and the posttreatment functioning of alcoholic patients. *Journal of Health and Social Behavior,* 1982, *23,* 310–317.

Yeaton, W. H., & Sechrest, L. Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology,* 1981, *49,* 156–167.

18

Measuring Implementation and Multiple Outcomes in a Child Parent Center Compensatory Education Program

Kendon J. Conrad and Maurice J. Eash

This paper presents the results of a 2-year evaluation of the Child Parent Center Compensatory Education Program in Chicago, Illinois. Child Parent Centers were established in certain disadvantaged areas to provide systematic educational experiences for preschool children as young as 3 years old. The objectives of this study were (a) to document implementation of the program in the Child Parent Center (CPC) and its latest adaptation, the Child Parent Expansion (CPX) Program; and (b) to evaluate the effectiveness of CPCs in improving academic achievement, locus of control, and home concern and

---

The authors are grateful to Robert F. Boruch, Geneva D. Haertel, Ernest T. Pascarella, Herbert J. Walberg, and the reviewers of *AERJ* for their thoughtful comments on earlier drafts of this paper, and to Jane Coffey and Cheryl Priller for manuscript assistance. The study would not have been possible without the dedication of Henry Springs, Debora Gordon, and the principals, teachers, and school and community representatives of the Westside Chicago Public Schools.

From Kendon J. Conrad and Maurice J. Eash, "Measuring Implementation and Multiple Outcomes in a Child Parent Center Compensatory Education Program," *American Educational Research Journal,* 1983, 20, 221-236.

support for academic achievement. The study was designed to contribute to a better understanding of the practice of at least one phase of early childhood compensatory education, the CPC. The strategy of employing rigorous methodology in evaluating local compensatory programs has been suggested as a preferred alternative to massive, expensive evaluations that could show no effect (House, Glass, McLean, & Walker, 1978).

RATIONALE AND DESCRIPTION

Child Parent Centers, funded under ESEA Title I, were begun in Chicago in 1967. The rationale underlying the CPC is fundamentally similar to that of other early childhood compensatory education programs, most notably Head Start (Westinghouse Learning Corporation/Ohio University, 1969) and Follow Through (Anderson, St. Pierre, Proper, & Stebbins, 1978; Haney, 1977; House et al., 1978). Like these two massive programs, the CPC was intended to break the cycle of poverty through an early (age 3 through 8), intense, systematic, and continuous educational intervention because it was believed that ability was more susceptible to change during early years (Bloom, 1964). Other studies suggested that the warmth of the relationship between parent and child had significant effects on children's acceptance of self-responsibility (Becker, 1964) and that parental behaviors were related to the behaviors of children (Baumrind, 1971). Generally, it is an accepted tenet that parental influences are crucial and pervasive in child development and that it is almost impossible to discuss any aspect of this field without considering its relationship to parent attitudes and behavior (Ausubel & Sullivan, 1957; Bronfenbrenner, 1975).

The distinguishing characteristics of CPCs, therefore, are their emphasis on direct parental involvement in the schools and a structured language/basic skills program. The CPCs also provide a wide range of educational materials, a reduced class size, attention to improved nutrition and health care, half-day classes of 17 students, a teacher aide, and a full complement of secretarial and other support staff.

A special adaptation of the CPCs, the CPX, was established in six selected sponsor schools in September 1978. This expansion involved the immediate extension of CPC treatment to include children ages 6, 7, and 8. The four sponsor schools included in the present study were involved in the immediate expansion program. Probably the most notable difference between the CPC and the CPX is that class size is 17 in the CPC and 25 in the CPX. Otherwise, services and resources are designed to remain at the same level. From this point in the paper, CPCs and CPXs will both be referred to as CPCs unless it is necessary to distinguish between them.

Evaluation of Child Parent Centers

A number of evaluations of the educational effects of the CPCs have been

conducted (e.g., Fuerst, 1977; Stenner & Mueller, 1973; Eash & Rasher, Note 1). Due to the limitations of time, resources, and purpose, however, the existing evaluations share a number of weaknesses: (a) they have been cross-sectional, one-shot studies; (b) they have often lacked comparison groups; (c) they have analyzed achievement outcomes almost exclusively; (d) generally they have failed to document the nature or degree of implementation of the experimental treatment versus the conventional treatment; and (e) they have been conducted or sponsored by the same agencies that fund and direct the program rather than by independent agencies.

To improve on past efforts, a 2-year evaluation was designed. In this evaluation, measures were sought on academic achievement, home support for academic achievement, program implementation of the instructional treatment, locus of control (Crandall, Katkofsky, & Crandall, 1965), and several status variables (e.g., attendance, family size). The study employed a quasi-experimental, nonequivalent control group design (Cook & Campbell, 1979), which used multiple regression procedures in an attempt to adjust statistically for initial group differences.

The most obvious threat to internal validity in this study was the threat of differential selection of students in the treatment program on the basis of ability. However, it was not clear in which direction, if any, the results would be biased—in favor of the CPC children or in favor of the children in regular classrooms. In past evaluations of compensatory education programs, it was found that the bias favored regular programs because the compensatory program selected children who were disadvantaged (Campbell & Erlebacher, 1970). In the case of the CPC, however, some have theorized that bias would favor the CPC because only the most caring parents would enroll their children. The latter argument is mitigated by the reality that the poor people served by CPCs have little choice of where they live and little awareness of educational opportunities. In fact, CPCs have to recruit students and parents into the program. Combined with the fact that CPCs have no real means to compel the parents' attendance, it is doubtful that self-selection would play a strong role in biasing the results in favor of the CPCs. On these grounds, the investigators had no strong theoretical reasons for expecting differences between treatment and comparison groups.

METHOD

The 2-year evaluation design involved the comparison of two populations (treatment group vs. comparison group) for each of two age cohorts (5-year-olds and 8-year-olds). The evaluative measures are listed in Table I. In addition, CPC and non-CPC classrooms were observed to gauge differential implementation of the intended CPC program characteristics. First, 5-year-olds enrolled in CPCs with 1 to 3 years of CPC experience ($N = 227$) were compared to similar children, that is, living in the same sort of disadvantaged

TABLE I

Measures Obtained, Age Groups, and Years in Which Data Were Gathered

Age 5–6 cohort		Age 7–8 cohort	
5-year-olds (1979)	*6-year-olds (1980)*	*7-year-olds (1978)*	*8-year-olds (1979)*
Comprehensive Tests of Basic Skills	*Comprehensive Tests of Basic Skills* (excluded from analyses because of over 60% missing data/attrition)	*Iowa Tests of Basic Skills*	*Iowa Tests of Basic Skills*
Sex	"Home Concern and Support" from a random sample of parents (interview, Dolan, 1978)		Sex
First born			First born
Number of siblings in family			Number of siblings in family
Days absent			Locus of control (Crandall et al., 1965)
"Home Concern and Support" from a random sample of parents (interview, Dolan, 1978)			Days absent
			"Home Concern and Support" from a random sample of parents (interview, Dolan, 1978)

community within the same school district in regular classrooms ($N = 304$). Measures of achievement (*Comprehensive Tests of Basic Skills*), attendance, number of siblings, birth order, and home support were collected during the first year of the evaluation.

The second comparison was of students (8-year-olds) at the third grade level who had had prior CPC experience to students at the same level who had attended regular classrooms. The 8-year-olds with prior CPC experience ($N = 54$) were enrolled in expansion (CPX) classrooms, while the other 8-year-olds in the sample were in regular classrooms ($N = 425$). The purpose of this comparison was to help assess the long-range effects of participation in CPCs and CPXs on students and to assess the effectiveness of the relatively diluted CPX program. Measures collected for the 8-year-old sample were the same as for the 5-year-olds, with the addition of data on locus of control. In addition to the measures just described, a sample of classroom observations was conducted in CPC, CPX, and regular classrooms for both age groups during the first year of the study to document the instructional treatment of CPC/CPX classrooms versus regular classrooms.

Organization

The sample of children included in the study was drawn from four elementary sponsor schools plus associated CPCs and two additional comparison elementary schools, all in similar neighborhoods in close proximity. All the schools and CPCs served predominantly black, low-income populations. Achievement in this area, as measured by standardized tests, was below national norms.

All student measures reported here were group administered or were obtained from school records. Home interviews were administered to a random sample of experimental and control parents ($N = 121$) by school and community representatives on an individual basis. Each interview took approximately 1 hour to complete.

Program Implementation Measures

Data for the program implementation component were collected through observations of 43 randomly selected classrooms from CPCs and comparison schools by experienced observers. A single, 1-hour training session was held for the 22 experienced observers in which items were read and examples of situations that would apply to particular items were presented. Forty-three observations were made: 25 in conventional classrooms and 18 in CPC classrooms. The instrument used was a modified version in the Classroom Observation Rating Scale (CORS) developed by Walberg and Thomas (1974).[1]

[1] Walberg and Thomas report internal consistencies of the total scale to be above .95.

In an effort to delineate separate processes of the CPC program, four scales were composed a priori from the CORS (see Table II). The scales were (a) child centeredness, (b) presence of evaluation of student achievement, (c) enriched environment, and (d) parent involvement (five items composed specifically for the CPC study, which were substituted for original CORS items).

An additional measure of parent involvement in the school program was obtained by using parent self-reports from parent interviews given to a random sample of 83 treatment and comparison group parents, 34 parents of children in CPCs (coded 1), and 49 parents of children in conventional classrooms (coded 0). Two questions were used as variables in this study:

1. How often do you come to (name of school or CPC)?
2. Are you a member of a school Advisory Council, Parent Council, PTA, or other school-related organization?

Question 1 was coded 1 (seldom or never) to 5 (almost every day) and question 2 was coded 1 (yes) and 0 (no). This parent involvement measure

TABLE II

The Four A Priori Scales of the Classroom Observation Rating Scale and Their Item Representation

Sample items	*N* of items	Measures implementation of
1. Child-centeredness: Children work individually and in small groups at various activities. Children are expected to do their own work without getting help from other children (R).[a] Teacher bases instruction on each individual child and his or her interaction with materials and equipment.	21	Reduced class size, individualization, humaneness
2. Evaluation: Teacher gives children tests to find out what they know. Teacher views evaluation as information to guide instruction and provisioning for the classroom.	6	Structured basic skills program and evaluation
3. Enriched environment: Materials are readily accessible to children. Books are supplied in diversity and profusion (including reference, children's literature). Manipulative materials are supplied in great diversity and range, with little replication.	11	Provisioning for abundant materials
4. Parent involvement: There are parents in the classroom. The environment includes materials for parents to read or use.	5	Parent involvement in the classroom

[a] The "R" means that coding was reversed when this item was scored.

was necessary because the CORS "parent involvement" scale measured only parent classroom participation, whereas the interviews assessed general parent involvement in the school.

Dependent Variables

Outcome measures used were the *Comprehensive Tests of Basic Skills* (CTBS/McGraw-Hill, 1973), the *Iowa Test of Basic Skills* (ITBS) (Lindquist et al., 1972), a measure of locus of control called the Intellectual Achievement Responsibility Questionnaire (IAR) (Crandall et al., 1965), and an educational home support interview by Dolan (1978).

The IAR questionnaire, which measures the degree to which students accept responsibility for their own intellectual progress, was administered to 8-year-olds also. The measure is considered an assessment of locus of control with regard to academic performance. Locus of Control Total,[2] which reflects the extent to which students assume responsibility for both positive and negative events was the score used in these analyses. The items were read, as recommended by the authors, to children below grade five.

A modified version of Dolan's (1978) home support interview was administered to a random sample of 120 parents in the first year and 83 parents who were followed in the second year.[3] These data were used to assess the level of parent support for academic achievement. The interviews were regarded as especially valuable because they were conducted by school and community representatives (SCRs). The SCRs were known to parents because most of them lived in the community. It was felt that if anyone could get reliable information from parents it was the SCRs. A 100 percent sample return on parent interviews was achieved during the first year and a 70 percent response of the original sample was achieved during the second year.

Independent Variables

The chief independent variable of interest in the present study was the treatment, that is, the CPC experience. Students without previous CPC experience served as the comparison group. In addition to the treatment, however, sex, firstborn, days absent, and total number of siblings were included as independent variables in some of the analyses.

RESULTS OF THE PROGRAM IMPLEMENTATION STUDY

Measures of CPC program processes were obtained from the modified CORS (Walberg & Thomas, 1974) and questions asked of parents regarding

---

[2] The authors report test-retest reliability coefficients for third and fifth grades of .69 for this score.

[3] Dolan reports a test-retest reliability of .74 and internal consistency coefficients ranging from .68 to .79.

TABLE III

Group Means, Standard Deviations, Mean Item Values, F-Ratios, and Statistical Significance of the Four A Priori Scales of the CORS

Scale	Regular mean	CPC/CPX mean	Regular *SD*	CPC/CPX *SD*	Mean item values[a] (4-pt. scale) CPC	Non-CPC	*df*	*F*	*p*-value
Child-centeredness	41.9	47.1	8.67	6.64	2.24	2.0	41	4.426	.04
Evaluation	14.2	18.4	4.71	5.02	3.1	2.4	41	7.9	.008
Enriched environment	27.6	30.1	5.16	4.98	2.73	2.5	41	2.356	.13
Parent involvement	7.0	8.3	2.78	3.16	1.7	1.4	41	2.143	.15

[a] To obtain comparability of measurement on these scales the group means were divided by the number of items. This gives an indication of relative magnitude as well as absolute magnitude on a four point scale.

their attendance at school and membership in organizations. First, it was found that the CPC program scored significantly higher than conventional programs on two aspects, "child-centeredness" and "evaluation of student activities." This finding was consistent with the belief that the program was indeed more intensive than conventional programs for the disadvantaged in these two ways (Table III). Statistically significant results were not obtained, however, for the other two scales, "enriched environment" and "parent involvement in the classroom." The finding that CPCs did not have a significantly richer material environment was blunted by the concurrent finding that the mean item values for CPCs (2.73) and non-CPCs (2.5) were both average or better, indicating the presence of a fairly rich environment for both. In contrast, neither CPCs (1.7) nor non-CPCs (1.4) showed evidence of noticeable parent involvement in the classroom.

On the other hand, in the questions asked of a random sample of CPC and non-CPC parents regarding "attendance at school" and "membership in school organizations," it was found that CPC parents had significantly greater school attendance ($p = .009$) and school organization membership ($p = .05$). These findings support the belief that the CPCs were indeed involving parents in the life of the school to a greater extent than were conventional schools, although this involvement did not extend visibly into classrooms.

In summary, this study of CPC program implementation concluded that the CPC program was being implemented as intended in its major aspects, that is, child-centeredness, evaluation of student activities, enriched environment, and parental attendance at school and school organization membership. The one negative finding provided evidence to support the belief that parents were not becoming involved in the program at the classroom level.

ACHIEVEMENT, LOCUS OF CONTROL, AND PARENT INTERVIEW RESULTS

Multiple regression analysis of covariance was the major analytical procedure used. For the 5-year-old sample each reading and mathematics achievement measure was regressed on student sex, birth order, family size, and days absent (treated as covariates) and a dummy coded variable representing CPC versus comparison group membership. Days absent was regarded as a covariate because it was thought to control for nonequivalence in the receipt of treatment. Standardized partial regression coefficients (or beta weights) were employed to represent the contribution of individual variables to prediction with the influence of all other variables in the equation held constant statistically.

The analysis of the data from the 8-year-old sample proceeded in essentially the same manner. The one exception was the addition of locus of control as a covariate in the prediction of each achievement measure. Locus of control was thought to be a useful covariate because it has been associated

with increased achievement (Lefcourt, 1976). This theory holds that children with higher internal locus of control will have higher achievement (de Charms, 1972; Messer, 1972; Reimanis, Note 2). Therefore, in the covariance analysis it was regarded as a pretreatment measure which would statistically improve equivalence of groups. In doing this kind of analysis, of course, one risks removing some of the treatment effects. If effects would persist, then, they would seem robust.

Conversely, locus of control also was regressed on each of the other covariates and the dummy treatment variable to determine if it was significantly and positively associated with CPC exposure. In this analysis locus of control was regarded as a dependent variable, an outcome associated with the CPC treatment. Because home support data were not available for most students, the measure of home support was not included in the regressions.

Results: Age 8

For all subtests of the ITBS, the regression coefficient for expansion versus control was greater than zero, indicating a positive effect of CPX classroom participation (see Table IV). The beta weights for Vocabulary, Spelling, and Word Analysis were positive but not statistically significant, whereas those for Reading Comprehension, Mathematics Concepts, Mathematics Problems, and Mathematics Total were statistically significant at the .05 level. These findings indicated significant differences not in the areas where recall is the dominant cognitive function (Vocabulary and Spelling), but, rather, in those areas where intepretation and application are most important (Reading Comprehension and Mathematics Concepts). Although the relationships were small and the sample large, the results were considered significant because locus of control was used as a covariate in an effort to ensure greater group equivalence. Without the covariate, results favor CPCs more strongly (see e.g., Tables V and VI).

It was also found that expansion students demonstrated a significantly higher internal locus of control ($p < .01$) when compared to students in regular schools. Further, it was found that higher internal locus of control was strongly related to increased achievement. This was demonstrated by significant beta weights for locus of control in the regressions for all ITBS subscales. These findings clearly supported the findings of other studies relating locus of control to achievement (de Charms, 1972; Messer, 1972; Reimanis, Note 2). A related finding was that larger family size had an inverse relationship with greater locus of control, possibly indicating that the different level of interaction of adults and child was responsible in some degree for a higher sense of internal control.

Results: Ages 5 and 7

In the 5-year-old sample (Table V), the relationship between CPC treat-

TABLE IV

Total R^2 and Beta Weights for 8-year-old Achievement Measures and Locus of Control

	Vocabulary	Reading Comprehension	Spelling	Word Analysis	Mathematics Concepts	Mathematics Problems	Mathematics Total	Locus of Control
R^2 Total	.25**	.28**	.18**	.19**	.31**	.27**	.31**	.06**
Sex	.09	.06	.22	.10	.08	.12*	.11*	.01
Previous CPC experience	.05	.07	.03	.05	.15*	.10	.13	.06
First born	−.01	−.04	−.14	.003	.04	.04	.04	.05
Number of siblings in family	.02	−.01	−.05	−.05	−.04	.01	−.01	−.17**
Locus of control	.37**	.37**	.24**	.24**	.41**	.42**	.44**	—
Days absent	−.30**	−.27**	−.11	−.25**	−.28**	−.20**	−.25**	−.03
Expansion of CPC vs. control	.05	.18*	.09	.16	.18*	.15*	.17*	.21**
N of cases	289	288	166	207	289	289	289	302

Note. Achievement measures were taken from the *Iowa Test of Basic Skills*; locus of control was coded 1 = CPC, 0 = control.

* $p < .05$

** $p < .01$

TABLE V

Total R^2 and Beta Weights for 5-year-old Achievement Measures

	Letter Forms	Letter Names	Alphabet Skills	Listening for Information	Letter Sounds	Visual Discrimination	Sound Matching	Total Discriminating	Language	Mathematics
R^2 Total	.08**	.07**	.09**	.04**	.07**	.13**	.07**	.12**	.007	.10**
Sex	.003	−.03	−.01	.06	.05	−.12*	−.02	−.07	−.004	.01
First born	−.09	−.07	−.09	−.01	.004	−.07	.01	−.03	.04	−.01
Number of siblings in family	.05	.04	.06	−.06	.03	.07	.12	.12	−.06	−.02
Days absent	−.25**	−.23**	−.26**	−.18**	−.22**	−.24**	−.12**	−.19**	−.03	−.24**
CPC vs. control[a]	.12*	.12*	.13*	.07	.14*	.23**	.20**	.25**	.06	.21**
N of cases	276	277	276	278	279	283	285	283	281	282

Note. Achievement measures were taken from the *Comprehensive Tests of Basic Skills.*

[a] Coded 1 = CPC, 0 = control

* $p < .05$

** $p < .01$

TABLE VI

The Effect of CPC Treatment on 1978 ITBS (7-year olds) Achievement Scales Controlling for Days Absent

Scale	N^a	Adjusted Means Non-CPC	CPC	p-value	R^2
Vocabulary	250/48	21.75	25.23	.01	.091
Reading Comprehension	238/45	24.00	27.59	.01	.064
Spelling	207/34	27.58	34.81	.001	.095
Word Analysis	246/48	20.73	25.68	.001	.106
Math Concepts	250/48	21.69	25.18	.003	.09
Math Problems	249/48	20.87	25.67	.001	.066
Math Total	249/48	21.29	25.31	.001	.085

[a] Number of subjects in comparison group over number of subjects in treatment group.

ment and achievement was even more pronounced. As was the case for the 8-year-old sample, regression coefficients for all subscales were positive. On the scales involving composite scores, Total Discriminating and Mathematics, the effect of CPC participation was significant at the .01 level.

When measuring the effects of treatment on the 1978 ITBS scales (age 7) controlling for days absent, it was found that CPC treatment was positively associated with increased achievement ($p < .01$) on all scales (Table VI). This analysis lacks the covariates used in the analysis of the 1979 ITBS data, but relies instead on the theoretical equivalence of groups implied by the original quasi-experimental design.

Parent Results

A strong relationship was found between being a CPC parent and home support for academic achievement. For the age 5 cohort, CPC parents scored higher on the home control and support interview ($t = 1.56$, $df = 59$, $p = .13$), but without a statistically significant difference. However, the parents of the age 8 cohort from CPCs did show a statistically significant increase over non-CPC parents on this same measure ($t = 2.25$, $df = 59$, $p = .05$). This, coupled with the finding of higher parental attendance at school, supports the belief that the CPC parent program is having the intended effects of increasing the parents' willingness and ability to support the academic achievement of their children. It appears from these results that when parents participate in the CPC parent program, they tend to enrich their home environments in ways that are supportive of enhanced school achievement. It also appears that increased parental attendance at school is a mediating factor in increasing parental ambitions for their children's academic achievement.

DISCUSSION

Rather than evaluate a mass of programs superficially, this study focused in depth on one compensatory education program using rigorous methodology including multiple outcome measures and an implementation measure. Therefore this study examined multiple facets of program impact, thus providing decisionmakers with numerous policy-relevant variables on which to judge the program.

On achievement measures comparing CPC students vs. non-CPC students, the CPC students scored significantly higher even when locus of control was used as a covariate. This was considered an indication that the results were quite robust.

As expected, locus of control was found to correlate highly with achievement, and CPCs were again found to be superior on this variable. Additionally, an inverse relationship between increased locus of control and number of siblings supported the belief that increased access to parents promotes increased feelings of responsibility for achievement in children. This indication supports the strategy of parent involvement employed by the CPCs.

In addition to improvement in student measures, this study found that the CPC program did promote increased parental involvement in school activities. This may be a factor in the finding that CPC families had a home environment that was more supportive of academic achievement. These findings would be consistent with Bronfenbrenner's (1975) belief that programs should affect the home environment and family attitudes if they are to have long-term effects. Finally, classroom observations and parent interviews support causal inferences relating the effectiveness of the program to components such as child-centeredness, the presence of classroom evaluation, and parent involvement in the school but not in the classroom.

Taken as a whole, this 2-year study gives substantial support to the usefulness of early intervention through formal education beginning at age 3. It is limited insofar as it is unable to call clear causal inferences from the results. Many questions remain unanswered, such as "which is chiefly responsible for increased achievement and locus of control—parent involvement or a structured classroom with reduced class size?" However, the study does at least lend strong correlational support to the belief that when a total program is developed that involves parents as well as children, it does produce increased achievement and the effect persists as seen when students are tested as much as 4 years later (grade 3). Therefore, this study contributes substantial evidence to the growing body of research supporting compensatory education programs, especially those having parent involvement (e.g., Consortium for Longitudinal Studies, 1980; Weikart, Bond, & McNeil, 1978).

In a section of the population where low achievement has had grave consequences for limiting job opportunities, the effectiveness of the early

intervention strategy in causing improvement in achievement takes on added importance. Continued ineffective education feeds those attendant social problems of unemployment, low social mobility, crime, and dependence on welfare. To interdict this vicious cycle should be of high priority in national goals. Effective early intervention as documented in this study appears to be one approach that holds promise of making long-range impacts on children's achievement and expanding the opportunities for them to enter more effectively into the mainstream of society. This evidence is contrary to a national policy which has been, over the last decade, directed to reducing formal schooling opportunities for young children.

In addition, it is our conclusion that this strategy of evaluating local programs rigorously and in depth will contribute not only to program improvement locally but will build a body of cases to direct national policy. This strategy, we are convinced, is more promising and more economical than expending vast sums on superficial, and often unmanageable, national evaluations built from questionable data bases.

REFERENCE NOTES

1. EASH, M. J., & RASHER, S. P. *Longitudinal evaluation of Child-Parent Centers.* Unpublished manuscript, University of Illinois at Chicago Circle, 1976.
2. REIMANIS, G. *Effects of experimental IE modification techniques and home environment variables of IE.* Paper presented at the American Psychological Association Convention, Washington, D.C. 1971.

REFERENCES

ANDERSON, R. B., ST. PIERRE, R. G., PROPER, E. C., STEBBINS, L. B. Pardon us, but what was the question again?: A response to the critique of the Follow Through evaluation. *Harvard Educational Review,* 1978, *48,* 161–170.

AUSUBEL, D. P., & SULLIVAN, E. V. *Theory and problems of child development.* New York: Grune & Stratton, 1957.

BAUMRIND, D. Current patterns of parental authority. *Developmental Psychology Monographs,* 1971, *4,* 1–103.

BECKER, W. C. Consequences of different kinds of parental discipline. In M. L. Hoffman & L. W. Hoffman (Eds.), *Review of child development research.* New York: Russell Sage Foundation, 1964.

BLOOM, B. *Stability and change in human characteristics.* New York: Wiley, 1964.

BRONFENBRENNER, U. Is early intervention effective? In M. Guttentag & E. Streuning (Eds.), *Handbook of evaluation research, 2.* Beverly Hills, Calif.: Sage, 1975.

CAMPBELL, D. T., & ERLEBACHER, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (ED.), *Compensatory Education: A national debate,* vol. 3 of *Disadvantaged child.* New York: Bruner-Mazel, 1970.

Consortium for Longitudinal Studies. *Persistance of preschool effects* (Grant No. 90-C-1311 [03]). Washington, D.C.: Administration for Children, Youth, and Families, Office of Human Development Services, Dept. of Health, Education, and Welfare, October 1980.

COOK, T., & CAMPBELL, D. *Quasi-experimentation.* Chicago: Rand McNally, 1979.

CRANDALL, V. C., KATKOFSKY, W., & CRANDALL, V. J. Children's beliefs in their own control of reinforcement in intellectual-academic achievement situations. *Child Development*, 1965, *36*, 91–109.

CTBS/McGraw-Hill. *Comprehensive tests of basic skills*, Level A, Form 5 (Expanded ed.). Monterey, Calif.: Author, 1973.

DE CHARMS, R. Personal causation training in the schools. *Journal of Applied Social Psychology*, 1972, *2*, 95–113.

DOLAN, L. The affective consequences of home support, instructional quality and achievement. *Urban Education*, 1978, *13*, 323–344.

FUERST, J. S. Child Parent Centers: An evaluation. *Integrated Education*, 1977, *15*, 17–20.

HANEY, W. *The Follow Through planned variation experiment Volumes 5: A technical history of the national Follow Through evaluation.* Cambridge, MA: The Huron Institute, 1977.

HOUSE, E. R., GLASS, G. V, MCLEAN, L. D., & WALKER, D. F. No simple answer: Critique of the Follow Through evaluation. *Harvard Educational Review*, 1978, *48*, 128–160.

LEFCOURT, H. M. *Locus of control: Current trends in theory and research.* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

LINDQUIST, E. F., HIERONYMUS, A. N., HOOVER, H. D., PETERSON, J., WHITNEY, M., LEWIS, T., NECKERE, E., STRAYER, F., FRY, M. MONROE, V., HUMPHREY, K., BILLINGTON, R., & COHEN, A. *Iowa Tests of Basic Skills.* Boston: Houghton Mifflin, 1972.

MESSER, S. B. The relation of internal-external control to academic performance. *Child Development*, 1972, *43*, 1456–1462.

STENNER, A. J., & MUELLER, S. G. A successful compensatory education model. *Phi Delta Kappan*, 1973, *55*, 246–248.

WALBERG, H. J., & THOMAS, S. C. Defining open education. *Journal of Research and Development in Education*, 1974, *8*, 4–13.

WEIKART, D., BOND, J., & MCNEIL, J. *The Ypsilanti Perry Preschool Project—Preschool years and longitudinal results through fourth grade.* Ypsilanti, Mich.: High/Scope Educational Research Foundation, 1978.

Westinghouse Learning Corporation/Ohio University. *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development.* Washington, D.C.: Office of Economic Opportunity, 1969. (Distributed by Department of Commerce, Clearinghouse for Federal Scientific and Technical Information, Springfield, Va.)

19

Measuring the Degree of Program Implementation
A Methodological Review

Mary Ann Scheirer and Eva Lantos Rezmovic

The concept "degree of implementation" is critical in order to derive valid conclusions from both outcome and process studies of social and technological innovations. To correctly attribute the observed outcomes of a social program to the intervention, the researcher should have empirical evidence on the extent to which program components were implemented. Without such evidence, researchers may erroneously conclude that an intervention was ineffective when, in fact, treatment implementation was inadequate to afford a valid test of the program. Further, even successful interventions may not be replicable without a knowledge of program components. Evaluation

---

AUTHORS' NOTE: *This article was prepared with support from the National Science Foundation, Grant No. PRA-8022612. The views expressed above are those of the authors and do not necessarily reflect the politics of the National Science Foundation. A*

From Mary Ann Scheirer and Eva Lantos Rezmovic, "Measuring the Degree of Program Implementation: A Methodological Review," *Evaluation Review*, 1983, 7, 599–633.

researchers now emphasize the importance of assessing program implementation prior to analyzing program effectiveness (Boruch and Gomez, 1979; Cook and Poole, 1982; Sechrest and Rednor, 1979; Rossi et al., 1979).

In innovation process studies, the interest is in examining the processes by which social and technological innovations are diffused to, adopted by, and implemented in new locations (Berman and McLaughlin, 1978; Beyer and Trice, 1978; Tornatzky et al., 1979; Scheirer, 1981, Yin 1979.) This growing body of research addresses such issues as, why do some potential users adopt and implement an innovation more quickly and more easily than others? What organizational supports or individual changes are necessary for the full implementation of a major technical innovation? In such process studies, the extent or degree of implementation is logically the key dependent variable, to which variability in processes over time or across locations is related.

In spite of the critical importance of the degree of implementation to both evaluations of program outcomes and examinations of innovation processes, applied researchers have not developed standard methodological paradigms for constructing implementation measures. Instead, they have tended to create ad hoc implementation indicators consistent with their intuitions or their research budgets. Often, they appear to have given little attention to prior conceptual or empirical work. Because research using the concept of "implementation" has expanded so rapidly in recent years, and because it has been conducted by investigators from many disciplinary bases, there is little consensus among researchers on the appropriate conceptualization or measurement of the degree of implementation.

This article reports conceptual and methodological issues derived from a review of 74 studies that have included measures of the concept "degree of implementation." We describe the types of implementation measures used in several content fields, then examine the extent to which these measures satisfy a number of basic methodological criteria. Using both quantitative and qualitative information, we explore the consistency of results obtained from the use of various measurement techniques. The objective of this article is thus to assess whether

---

preliminary version was presented at the Evaluation Research Society Annual Meeting, Baltimore Maryland, October 1982.

measurement of the degree of implementation has achieved the scientific underpinning adequate to support its major role in applied research.

Perhaps some definitions will orient the reader to the concepts used here. "The innovation" is defined as whatever material invention, social technology, policy change, or legislative mandate was being described in each document reviewed. This definition was deliberately left quite broad in order to compare the methods used to assess the implementation of "hard" technologies with those applied to social programs or policy changes.

In contrast, we maintain a careful distinction between the terms "degree of implementation" and "implementation processes." *Degree of implementation* is the extent of change that has occurred at some particular time toward full, appropriate use of the target innovation. Some frequently used synonyms are the extent of "utilization", or simply "use." *Implementation processes* are the sequences of organizational changes and support mechanisms that account for the degree of implementation found at a given time. Implementation processes would include training staff members to use the innovation, obtaining supervisory and administrative support for the innovation, perhaps forming an interdisciplinary committee to coordinate the change process, and so forth. These processes occur after the implementing entity has decided to adopt the innovation. Adoption is viewed analytically as a separate phase in the total technology transfer process (see Scheirer, 1983). We emphasize these definitional distinctions in order to make clear that the focus here is solely on the degree of implementation.

This article does not address the important question of a relationship between findings about the degree of implementation and the research techniques used in each study. Unfortunately, information about the overall extent of implementation was extracted from only 23 studies (31%). These were nearly evenly divided in the degree of implementation found: 7 studies showed a "high" degree of implementation, 9 studies found "mixed or moderate" results for this variable, and 7 studies showed a "low" degree of implementation. Many other studies could not be coded for the overall extent of implementation for various reasons: (1) they examined multiple innovations or had varying degrees of implementation in separate sites, (2) they examined implementation processes without explicit descriptive data about the extent of implementation achieved, or (3) they provided separate data concerning

several components of an innovation without integrating these data into an overall assessment of the degree of implementation. Consequently, not enough studies could be coded on this dimension for the overall findings on level of implementation to be analyzed in relation to other variables reported here.

METHODS

Literature examining implementation is widely scattered among various disciplines; it is usually not indexed under any common terms in standard abstracts, and is frequently in unpublished reports and documents. Consequently, there is not a readily identifiable pool of implementation studies available for review. An extensive search for relevant studies [1] yielded about 140 studies that appeared to contain an empirical examination of the extent of implementation. Upon closer examination, however, those that contained no measure at all of implementation were eliminated from the sample for review.

The final sample encompassed 74 studies in the nine content areas shown in Table 1. As shown, we could find few studies that examined the implementation of innovations in health, labor or employment, or public administration. For further analysis, therefore, these content areas are combined under the label "other public services." In contrast, more studies were available in education and criminal justice than we had time to review thoroughly. We attempted to include from these fields at least the more widely known, large-scale, and recent examinations of federally-funded innnovations. A complete list of the projects reviewed appears in Appendix A. The comparative recency of this research topic is indicated by the fact that 69% of the documents were issued in 1978 or later.

The differential success of the search process itself reveals some significant insights about the status of implementation research. This topic seems to have received its major empirical development within the "softer" technologies of education, criminal justice, mental health, and management science and information systems. Perhaps because of the extensive recent efforts to evaluate program effectiveness in these areas, coupled with the frequent absence of a material invention whose operation can be easily seen, these researchers or their funding agents have felt more need for implementation assessment. In contrast, researchers examining industrial technologies have appeared to put

TABLE 1
Percentage Distribution of Implementation Studies Reviewed, by Content Area

CONTENT AREA	*NUMBER*	*PERCENT*
Criminal Justice	9	12%
Education	19	26%
Health	1	1%
Human Services	9	12%
Information Systems (and Operations Research)	10	14%
Labor/Employment	4	5%
Mental Health	11	15%
Public Administration	3	4%
Industrial Technology (includes Energy/Environment)	8	11%
TOTAL	74	100%

more emphasis on benefit-cost assessments of the results of new technologies (e.g., Gold et al., 1980). The examination of health innovations has emphasized the development, diffusion and adoption of innovations (e.g., Kaluzny and Veney, 1973; Office of Technology Assessment, 1982; Russell, 1979) rather than their implementation. It is still puzzling, however, whether implementation has been less studied in these content areas because it is less problematic; that is, innovations are more easily implemented, or whether there has simply been less rigorous assessment of outcomes for these innovations, perhaps concealing implementation failures.

Review of the 74 studies in our final sample used a 22-page analysis guide, incorporating both precoded quantitative ratings and space for qualitative comments. The ratings relevant to this article will be discussed in connection with those specific findings. The two authors were the major document coders, with minor assistance from a third research associate. All coders had graduate training in evaluation research methodology, as well as intensive introduction to the issues being investigated.[2]

TYPES OF IMPLEMENTATION MEASURES

The measurement tools used to assess implementation were classified into nine techniques, shown in Table 2. They are roughly ordered from

TABLE 2
Percentage of Studies Using Each of Nine Techniques for Measuring Implementation, Ordered by Objectivity of Technique

Extent of Human Judgement Required	*Incorporation of "Meaning"*	*Technique*	*Number of Studies*	*Percent**
Low ↓ High	Low ↓ High	Technical measure of equipment performance**	2	3%
		Technical measure of output**	4	5%
		Unobtrusive indicators	2	3%
		Behavioral observation	19	26%
		Institutional records**	29	39%
		Interviews (telephone or in person)**	51	69%
		Questionnaires**	31	42%
		Ethnographic observations	28	38%

*Percents total more than 100% because multiple measuring techniques could be tabulated for each study; N of studies = 74.
**The order of techniques in brackets is alphabetical, rather than based on a judgement of their comparative objectivity.

the more objective (not requiring a human judgement) to the more subjective (dependent on a human observer or respondent). The converse of this objectivity dimension is, of course, the extent to which the measure is intended to capture and preserve the meaning of the data recorded, particularly the meaning experienced by the participants. The low percentage of studies using more objective measuring techniques is likely to be related to the low proportion of studies that examined a physical product (only 13 or 18% of the studies), as well as to the exploratory nature of much implementation research.

The infrequent use of the more objective methods may signal a need for research movement in that direction. As Nunnally (1976) points out, the development of standardized, independently verifiable measures is often a key step in opening a topic for scientific research. Objective

measures facilitate agreement about the observation of empirical events, permit accurate communication among researchers, and often save considerable time and money in the collection of data. The heavy reliance on subjective methodologies for measuring implementation may be one reason for the ambiguous conceptualization and frequently contradictory results on this topic. The use of each measurement technique will be briefly described, before we return to comparisons among them.

Technical measures. Implementation assessment by technical measures of equipment performance or of output was infrequent. Technical measure of equipment performance was defined as a measurement taken directly from a piece of equipment to indicate whether it is operating correctly, for example, a speed indicator, or a temperature reading for overheating. Technical measure of output was defined as the measurement of a product resulting from equipment functioning, such as the quality of air released by emissions control equipment. All six studies which employed such technical measures were within the information systems and industrial technology areas, fields in which relatively few, only 18, empirical studies of implementation could be located.

Unobstrusive indicators. A third type of relatively objective measure is made by unobstrusive indicators, such as the wide variety of nonconventional measures discussed in detail in Webb, Campbell, Schwartz, and Secrest's provocative book, *Unobtrusive Measures* (1966). Other than institutional records, discussed separately below, only two studies used unobtrusive indicators. It is unclear whether researchers used these techniques infrequently because unobtrusive indicators were judged to be inappropriate for the particular innovations under study, or whether these techniques are simply outside the standard repertoire of measures considered by implementation researchers.

Behavioral observation. The use of behavioral observation was moderately frequent; it was found in 26% of all studies. Behavioral observations were defined as observer-collected data using a prespecified set of categories, each with an operational definition of target behaviors to be recorded. Such observations are usually collected in explicit units of analysis, such as time periods, or blocks of work space. The procedures are used extensively by behavioral psychologists and have a

long tradition in industrial "time-and-motion" studies. These techniques yield data incorporating as little human judgement as possible, even though they require human observations of the target activities.

Institutional records. Institutional records, interviews, and questionnaires were all quite frequent in these implementation studies. All require a moderate to high degree of human judgement in the recording of data: by the institution's record producers as well as sometimes by record coders for the implementation study, by the interviewer in classifying and coding interviewee responses, and by questionnaire respondents in deciding which precoded alternative best captures their experiences or feelings.

Institutional or archival records systems included documents such as police and court records, therapy plans for mental patients, and individualized instruction plans. In most cases (24 of 29 studies), the researchers used existing records rather than creating a new record-keeping format. While this data source does prevent bias introduced from the researcher's judgements, the record keepers in the target organization may have little incentive for maintaining accurate, consistent records. Several studies reviewed here (e.g., Murray et al., 1978) mentioned that their archival data were maintained inconsistently, or conflicted with information from other data sources.

Interviews and questionnaires. The most common technique to assess implementation has been to ask people in the implementing organizations: over two-thirds of all these studies employed personal or telephone interviews, while about 40% used questionnaires. There were two major respondent types. Users, the staff who were the hands-on operators, or delivered the innovative service to clients, were questioned in 75% of the interview studies and 87% of the questionnaire studies. Managers, the supervisors administering the use of an innovation, were contacted in 65% of the interview studies, but in only 29% of studies using questionnaires. Perhaps surprisingly, researchers seldom asked the recipients of an innovation what service or products were delivered to them; only 18% of interview studies and 6% of questionnaires employed clients as an implementation data source. Questionnaires also tended to be confined to a single respondent type, users, using prestructured questions, while interviews ranged more widely by tapping several types of respondents with an unstructured format.

Ethnographic observations. The use of unstructured interviews can shade into the final classification for implementation measurement. This is labeled ethnographic observation, and is also called case study method or naturalistic observation. We defined this method as data collection that proceeds without prespecified observational categories, with an emphasis on understanding the totality of the situation observed. It might include a mixture of several other methods, such as document review, informal interviews, or personal observation. As indicated in Table 2, more than a third of the studies reviewed used this general technique.

COMPARISON OF FINDINGS FROM MULTIPLE MEASURES

Nearly three-quarters of the studies used more than one technique for assessing the degree of implementation. This provided the opportunity to examine the degree to which different measurement techniques yield comparable findings. Would the frequently used self-report measures yield more positive conclusions about the extent of implementation than would less subjective measures? A tendency for such "positive perceptions" to occur has been noted and explicated for outcome evaluation studies (Scheirer, 1978), but has not previously been examined for implementation studies.

Only 21 of the 55 studies (38%) with multiple measures of implementation did compare findings from the different measures. Even these comparisons were as often qualitative and judgmental as they were quantitative. Consequently, rigorous assessment for method-based biases was impossible. Further, biases were not necessarily reduced by the use of multiple measurement techniques, when the same respondents were used for both types. For example, in a study of a primary prevention school program (Moskowitz et al., 1980), both interview and questionnaire data from teachers revealed infrequent use of two innovation components. However, the same biases may have been operating in both measurement methods, thereby inflating the probability of congruent findings.

Findings on the consistency of results across measurement types were mixed, and no general conclusions emerge from them. For example, Scheirer's (1981) study of a goal planning system in a mental health facility found that self-reports were a highly inflated indicator of the actual behavior of therapy aides: the self-reported number of goal plans written was nearly three times the number of goal plans in the

organizational records. Using a case study approach involving interviews, institutional records, and ethnographic observations, Britan (1981) drew a similar conclusion about various indicators of the Experimental Technology Incentive Program of the U.S. Department of Commerce; program objectives, as stated in formal documents, did not match actual program operations, as observed.

Such negative findings on intermethod consistency of measurement are balanced by positive findings. Schaps et al.'s (1980) study of a classroom management program found that mean ratings of teacher behavior were consistent across behavioral observations, interviews, and questionnaires. Similarly, Anno (1977-1981) found that institutional records and questionnaire responses of both jail officials and clients converged regarding extent of health services delivered in jails. The one study (Schaps et al., 1980) used a rigorous multitrait-multimethod analysis found that only three of eleven skill areas included in a drug prevention program converged in the analysis of the empirical data; these three, rather than all eleven, were then formed into an implementation index. These results suggest that the "degree of implementation" for complex innovations is not a unidimensional phenomenon, but instead may require independent measures for distinct components.

Several studies recognized this need with different measures intended to tap separate components of the innovation. For example, an evaluation of a family intervention program for disadvantaged children (Nauta, 1981) found diverging levels of implementation shown by separate techniques about different program components: parent attendance at program meetings was quite low, but the number of home visits by agency staff was near target, although observations of the interactions during the home visits revealed some deviations from the intended activities. In this case, different measurement techniques yielded diverse findings, but do not shed any light on the question of potential methodological bias.

The reasons varied for the failure to compare the findings from diverse techniques. For several studies, the metric for judging level of implementation differed across techniques, preventing meaningful comparisons among measures. For other cases, data from several measurement types were simply presented descriptively as data from diverse sources, with no attempt to draw an overall picture of the extent of implementation. Sometimes, conclusions about an aggregate level of implementation were not tied at all to explicit data, so the reader was not able to determine whether the authors' conclusions were or were not based on their data.

Finally, some investigators apparently did not check for convergence among different measurement types either because such methodological examinations were deemed tangential to the main purpose of the study, or because resource and time limitations did not permit such investigations. However, given the great expense for collecting data from multiple sources, it would seem to be cost-effective to pay greater attention to comparative analyses.

In sum, based on the available data, we cannot draw firm conclusions about the extent of convergence among data obtained by different measurement techniques. In our sample of 74 implementation studies, empirical cross-technique comparisons were relatively infrequent. Although some studies did find evidence of a positive bias from the use of self-reports (Scheirer, 1981; Gersten et al. 1982), others as cited above found convergent results from self-reports and other methods. The cross-study comparisons suggest two conditions for self-report data that accurately reflect behavior. The first condition is that data collection be close in time to the implementation experience. The second is constructing interview or questionnaire items that ask about the specific actions defined by the researcher as components of the innovation, rather than general questions about the extent of use of or "satisfaction" with the innovation.

QUALITY CRITERIA FOR IMPLEMENTATION MEASURES

The major question in this discussion is how adequate are the measuring tools being used to assess degree of implementation? The underlying issue is the confidence one can have that the information they supply accurately reflects the extent of implementation taking place. In this section, we will examine five measurement criteria as they applied to these 74 studies:

(1) the use of multiple measurement techniques,
(2) the presence of an operational definition,
(3) the examination of reliability,
(4) the assessment of validity, and
(5) the use of sampling.

As each criterion is discussed, its use in the several content areas will be described in order to assess whether measurement adequacy differs by

content area. Table 3 summarizes the findings for all five criteria, showing the percentage of studies in each content area that met each criterion.[3]

MULTIPLE MEASUREMENT TECHNIQUES

When examining a complex construct such as degree of implementation, the use of multiple techniques is desirable to avoid the method-specific biases associated with each individual technique. The larger the number and variety of implementation measures used, the greater the likelihood that method-specific biases will be detectable, that multiple components of the innovation will be examined, and that implementation, therefore, can be validity assessed. This feature was retained as a criterion for desirable measurement, even though few studies used multiple techniques to systematically assess biases, as discussed in the previous section.

Examples of studies using multiple measuring techniques were plentiful. Some studies used the different techniques to assess separate parts of a multiple component program. For example, a community-based crime prevention program (Rasmussen et al., 1979) used police records to gauge the extent of physical premise surveys (performed by police), conducted interviews with residents to measure the extent and types of block club participation, and used case study observations to record progress in reconstruction efforts intended to increase street lighting. A few studies used the multiple measures for cross-validation of findings, such as detailed behavioral observations of classrooms, structured interviews with the same teachers, plus supervisors' rating forms; all used to determine the degree of implementation of a Direct Instruction elementary education program (Gersten et al., 1982).

The majority of studies reviewed in all content areas did use multiple measuring techniques, about 74% in total, but there was variation across content areas, shown in Table 3. All the studies within criminal justice used more than one technique to obtain implementation data, while only slightly more than half of mental health studies had done so. In addition, a substantial minority of studies (23%) used four or more different techniques, showing a considerable investment in data collection. Yet, as discussed before, data from the diverse sources were seldom compared, nor were multi-attribute indexes often constructed to provide a summary indicator of the degree of implementation. Frequently, data from different sources were simply described in different

TABLE 3
Percentage of Studies Fulfilling Measurement Quality Criteria, by Content Area

Content Area		*Criminal Justice*	*Education*	*Mental Health*	*Other Public Service**	*Information Systems*	*Industrial Technology*	*ALL AREAS*
Used Multiple Measurement Techniques	%	100	73	54	76	70	74	74
Operational Definition Present	%	44	68	100	47	70	25	64
Assessed Reliability	%	33	47	55	47	10	13	38
Assessed Validity	%	11	47	36	18	30	0	27
Used Random or Full Census Sampling	%	33	32	55	18	20	13	27
Number of Studies in Content Area		9	19	11	17	10	8	74

*This category includes studies of innovations in health, human services, labor or employment, and public administration.

sections of a report, with no attempt to distill from them an overall picture of the extent of implementation.

OPERATIONAL DEFINITION PRESENT

"An operational definition assigns meaning to a construct or variable by specifying the activities or 'operations' necessary to measure it" (Kerlinger, 1973:31). An operational definition connects the empirical observations to the underlying logic of the construct being measured, and thus clarifies the conceptualization of key concepts. In the absence of an operational definition, a gap remains between the level of theory and the level of observation. Then the researcher lacks the necessary prescriptions for what to observe in order to measure the construct of interest. Further, in this case, the reader is not provided with an explicit statement of the researcher's logical links between the intended concept, the implementation of the specific innovation, and the particular observations collected. Under conditions of explicit operationalization of implementation, then, both measurement and communication of results are likely to proceed more efficiently, logically, and with greater clarity of purpose.

Table 3 shows that 64% of studies across all content areas did specify degree of implementation in operational terms. Studies of innovations in education, information systems, and especially in mental health were more likely than other content areas to include an operational definition.

The type of measuring technique was somewhat related to the presence of an operational definition, with a significant negative association between this criterion and the use of ethnographic methods ($\chi^2 = 4.99$; df = 1, $p < .05$). This finding for ethnography is likely to reflect the usual purpose of ethnographic methods in implementation research: to provide an intensive, detailed analysis of implementation processes in the context of an environment. Because its emphasis is on understanding the totality of the situation observed, the procedure may be less reliant on an operational explication of implementation (with its associated measurement specifications) than are more structured data collection methods. However, when coding such studies, it was often difficult to determine just what was meant by "implementation" in that situation, and how much progress toward full implementation had actually occurred.

As the pattern of frequencies in Table 3 suggests, the presence of an operational definition is negatively related to the number of measurement techniques used (biserial correlation of $-.22$, $p < .05$). In some studies using multiple types of data collection but no operational definition, data from various sources were simply presented separately, without a clear explanation of which innovation components, or which implementation processes each data type was intended to describe. The reader was required to synthesize these extensive data to determine how much implementation had in fact occurred. An explicit operational definition for the extent of implementation of these complex innovations might have focused both the researcher's and the reader's attention on specifying, and thus understanding the component activities, even if such specification could only be derived ex post facto from the data resulting from multiple measures.

RELIABILITY

The reliability of a measuring instrument can be defined in terms of a relative lack of errors of measurement:

> Reliability concerns the extent to which measurements are repeatable—by the same individual using different measures of the same attribute or by different persons using the same measure of an attribute . . . To the extent to which measurement error is slight, a measure is said to be reliable [Nunnally and Durham, 1975: 311].

A reliable measure is particularly important for assessing the degree of implementation in order to distinguish true change from variability due to measurement error.

Among the 74 implementation studies reviewed, the researchers examined the reliability of measurement of implementation in only 28 studies (38%). In nearly two-thirds, this fundamental measurement criterion was not even addressed, let alone established. As indicated in Table 3, the content field of the study had a slight association with the tendency to assess reliability. Studies examining educational, mental health, or other public service innovations were somewhat more likely to include reliability estimates than were reports in other content fields. Even in these content areas, only about half the studies addressed reliability.

The major type of reliability assessment was inter-rater or inter-observer agreement—23 studies used this type. Seven studies assessed measurement stability over time, while three examined the internal consistency of a set of items. A few studies had multiple types of reliability evidence.

The review revealed that there is little basis for assuming that implementation measures have adequate reliability. This is shown by the coders' judgements concerning the overall extent to which reliability was established for the implementation measures, detailed in Table 4. Coders judged probable reliability both from the explicit reliability findings, and from any other information in the document that provided clues about likely errors in measurement. As shown in Table 4, even with this liberal basis for a reliability rating, no inference about the adequacy of measurement reliability was possible for 51% of the studies. For only 13 studies was reliability well established. Another 8 studies could be given a positive rating, even though doubts or problems with reliability remained. Thus, only 58% of the studies that contained reliability findings were judged to have a positive rating on this criterion.

These ratings indicate that reliability of data is a real problem for implementation assessments. Its absence in many other studies cannot be justified by any likelihood that this criterion will be automatically satisfied. Sources of data unreliability noted by coders ranged across the whole spectrum of measurement error: inconsistent data in institutional records, ambiguous definition and specification of innovation components, erratically changing behaviors of implementors from one day to the next, and inadequately trained observers. A coder summarized the methodological problems in one assessment of a human service delivery program by stating it contained "unreliable measures used by inadequately trained observers to assess the implementation of unarticulated program components."

VALIDITY

The validity of a measuring instrument is less easily established than is reliability, for validity addresses the question: "Are we measuring what we think we are measuring?" (Kerlinger, 1973). For implementation assessment, this is a central question involving the conceptualization of the construct "implementation" as well as the definition and specification of the innovation. This section will address only the more specific aspects of validity as a psychometric property: did the authors of

TABLE 4
Ratings of Reliability and Validity of Implementation Measures

Coder's ratings of reliability for measure(s) of implementation:

	NUMBER OF STUDIES	*PERCENT*
1. Major flaws; reliability not established.	3	4%
2. Questionable: some weaknesses present; (often little information given on which to base judgment).	12	16%
3. Positive rating, but minor problem or questions	8	11%
4. Reliability well established	13	18%
5. No information; or rating missing	38	51%
TOTAL	74	100%

Coder's ratings of validity for measure(s) of implementation:

	NUMBER OF STUDIES	*PERCENT*
1. Major flaws; validity not established.	5	7%
2. Questionable; some weaknesses present; (often, little information given on which to base judgment).	23	31%
3. Positive rating, but minor problem or questions.	4	5%
4. Validity well established	11	15%
5. No information; or rating missing	31	42%
TOTAL	74	100%

implementation studies include any explicit assessment of validity in their reports, and what were the coders' overall judgements concerning the extent to which validity was established?

As shown in Table 3, validity was assessed in these documents even less often than was reliability: only 20 (27%) of the studies mentioned it. Like reliability, validity was explicitly examined more frequently by the studies in mental health and education. Again, coders were requested to form a judgement concerning the validity of the implementation measures used. As shown in Table 4, coders rated validity as well established in only 11 (15%) of the 74 studies, with a positive rating in another 4 studies. For most studies, either no information at all was present, or so little positive data were available that the coder rated validity as questionable.

No one type of validity assessment predominated among the few implementation studies that examined this criterion, although all validity types are potentially relevant to this complex topic. Face validity—the extent to which users or respondents believe a measure "looks like" the intended concept—was addressed in only 7 studies. Future implementation researchers could usefully check on face validity by asking innovation users, developers, or clients if the intended measure captures the components they believe are essential to that innovation.

Assessing the content validity of an implementation measure requires comparison with a content plan or list of components developed for each innovation, preferably with assistance from experts on that innovation, such as its designer. Only 7 studies addressed content validity, in spite of the seemingly obvious necessity of specifying what an innovation consists of before attempting to measure its implementation.

Construct validity uses statistical procedures such as factor analysis or a multitrait-multimethod matrix (Campbell and Fiske, 1959) to establish an empirical test for a hypothesized underlying theory or construct. Although implementation analysts seem to agree that "extent of implementation" is a complex, multifaceted construct, only 10 of these studies examined the construct validity of their measures. Finally, criterion or predictive validity was noted in only 5 studies; much fuller use of this type could be relevant if the presence of a specific component or activity early in the implementation process were found to be a predictor of later overall implementation.

A number of studies simply used without further explanation or justification, a measure of implementation that appeared to have questionable face validity. For example, one project defined use of a technical innovation as the point at which it was first sold commercially. This definition overlooked the fact that the technology might be sold a few times, but not implemented in actuality, and then turn out to be a commercial failure. In another study examining the use of computer assisted instruction, "full utilization" was defined as the presence of one or more computer terminals for each group of 2,500 students. This definition failed to consider whether 2,500 students could possibly receive any educational benefit from one computer terminal, let alone whether the mere presence of a terminal meant that it was fully used.

Several studies measured implementation through indicators of user satisfaction or goal achievement. In these studies, degree of implementation was assessed by such questions as, "How satisfied are you with the operation of (the innovation)?", or "What proportion of the project's goals have been achieved by now?" Such items are inherently ambiguous. Respondents may be thinking about extent of implementation in giving their responses, but they may also be giving their perceptions of the innovation's effectiveness, or their satisfaction with their own job role in relation to the innovation, or even whether their personal or professional interests are served by the adoption of the innovation. Thus, the wording or construction of items to assess the degree of implementation may have a great impact on the resulting data. It demands much more careful consideration in future implementation studies.

SAMPLING OF RESPONDENTS

Although an appropriate sampling strategy is not usually considered to be a measurement criterion, it was included in this review because it influences so strongly the adequacy of data collected by any measuring technique. In the absence of a representative sample, it is impossible to estimate the effects of sampling errors on the conclusions reached. Implementation researchers have not, in general, developed explicit guidelines for determining which elements or units should be included in the data collection plan.

Few implementation studies have systematically sampled the units for data collection from some larger population. As shown in Table 5,

TABLE 5
Distribution of Sampling Strategies
Used in Implementation Studies

TYPE OF SAMPLING	*NUMBER*	*PERCENT*
Nonrepresentative	26	35%
Representative of larger population (random, etc.)	7	9%
Full census (or attempted full census with deletions)	13	18%
Multiple types	7	9%
Unknown, missing	21	28%
TOTAL	74	99%

only seven studies (9%) used any type of representative sampling, whether random or stratified. Another group of 13 studies (18%) stated that the population measured was a full census (or attempted full census)[4] of the units applicable for that innovation. This occurred particularly when the innovation was being studied on an experimental basis with a very limited number of individuals or organizations, all of whom were assessed for implementation. A large group of studies had a nonrepresentative sample (26 or 35%); usually a sample of convenience to the researcher. A few other studies had multiple types of samples for various types of data.

The number of sites examined in these implementation studies ranged widely. Fourteen studies (19%) used only one site, although several of these collected data from multiple individuals or other units within the site. Perhaps wisely, given the state-of-the-craft for implementation research, few studies attempted a large scale investigation, with only nine studies examining more than 60 sites.

Multiple data sources were frequently used within each site Thus, even with fewer than 60 sites, the volume of data collected was often very large, even overwhelming. Yet, given the scarcity of representative or full census methods, there are real questions whether the sampling

strategies were adequately representative of the individuals or other units within each site.

Investigation of implementation may necessitate multiple units of analysis, of, for example, both individuals within an agency and the agency as a whole. Any single organizational informant may be an inaccurate reporter about the extent of implementation achieved by each individual within that organization. Therefore, sampling strategies for measuring the degree of implementation should consider both the accurate assessment of the level of implementation within a site and the generalizability of the findings across sites.

To use adequacy of sampling strategy as a methodological criterion across content fields, this multi-category nominal variable was dichotomized into use of either representative sampling or a full census, versus all other categories listed in Table 5. The results, shown in Table 3 along with the other criteria, indicate that the content area did not have a strong association with the use of adequate sampling, although mental health was again somewhat higher than other areas. Overall, only 20 studies (27%) were rated positively on this criterion, which reiterates the findings from other criteria of a serious lack of attention to the quality of the data produced.

The relative neglect of sampling criteria in studies of implementation is understandable, given the recency of research on this topic and the conceptual ambiguity that still engulfs the field. Most researchers appear to place greatest emphasis on attempting to conceptualize and measure implementation processes in one or a few locations before tackling a larger scale study with a sampling strategy that would permit generalization to a larger population. Given the complexity of the topic, the costs of multi-site research, the difficulty of operationalizing the implementation of many innovations, and the nonrandom adoption of many innovations, the low emphasis on generalizability may be necessary.

Even with a nonrepresentative sample of sites, however, implementation researchers need to carefully consider the sampling of respondents within each research location in relation to the most appropriate unit(s) of analysis for the target innovation. A new set of efforts will then be required to translate the methods into those that can be used for larger scale, generalizable studies and to synthesize the findings that are derived from the more intensive, exploratory small-scale studies.

CONCLUSIONS ABOUT MEASUREMENT QUALITY IN IMPLEMENTATION STUDIES

This review of measurement criteria in application to measures of the degree of implementation has revealed serious gaps in their adequacy. The most positive finding was that multiple measures were used in about three-quarters of the studies, although the potential value of such multiple measures remained unexploited, due to the scarcity of comparative analysis, or index construction. Nearly two-thirds of the studies did present an operational definition, thus making explicit their conceptualization of the construct, "degree of implementation." However, the extent of operationalization differed considerably by content area. Reliability and validity were examined in less than half of these implementation studies, and coders rated reliability and validity as established in fewer than one-fifth. Finally, random or full census sampling was used in only about one-quarter of the studies, indicating serious deficiency in the generalizability of findings from the studies. A consequence of the numerous flaws in measurement methodology is the unknown meaning of the findings. A scientific basis for assessing the degree of implementation has not been established in most of the available studies.

The type of measuring technique used for assessing implementation, whether behavioral observation, institutional records, interviews, questionnaires, ethnography, or other measures, was not, in general, associated with the presence or absence of reliability or validity assessment; nor with coders' ratings concerning the adequacy of either reliability or validity. Nor were sampling strategies related to the type of measuring technique used. While the presence of an operational definition was somewhat negatively related to the use of ethnographic methods, as discussed above, this association is likely to reflect the usually exploratory purpose of ethnographic methods. Thus, the variation in methodological quality of implementation studies does not appear to be a consequence of the measuring technique chosen.

The deficiencies noted for previous measures of the degree of implementation suggest the central problem in the construction of such measures: that of defining what the innovation is and what is meant by "full implementation." These definitions are frequently not fully specifiable in advance, particularly for a social program or piece of legislation that has no material component. For policy innovations, the definition of full implementation may be essentially a political decision;

ideally a definition put into researchable terms after consultation with legislators, policy administrators, and local service deliverers.

For an example of methods relevant to this task, see the literature on "evaluability assessment," by Wholey, 1979. For example, the degree of implementation of an income transfer program might be defined as the extent to which checks are delivered to eligible beneficiaries who apply for the benefits. Or it might be defined as the proportion of eligible beneficiaries who become aware of the program and apply for these benefits. Or both of these criteria might be used as components of "full implementation."[5] The essential task for a researcher assessing the implementation of the transfer program is to construct and report one or more working definitions of implementation. Then, readers can know what definitions were used in the analyses and conclusions of the report. That implementation *is* defined may be more critical than *how* it is defined.

IMPLEMENTATION MEASURES AND THE NATURE OF THE INNOVATION

As we have seen, the measuring instruments used in past implementation studies fall considerably short of meeting standard methodological criteria. We have noted some differences in quality factors associated with the content area of the innovation, but content area (and the likely concomitant differences in the disciplinary affiliations of the investigators) do not appear to be major determinants of good versus poor quality among these studies. Likewise, the use of particular measuring techniques is not frequently related to quality factors. Therefore, what does account for the variation in methodological quality?

A tentative answer emerges from our detailed examination of the characteristics of each innovation being implemented.[6] One innovation characteristic was related to several quality criteria, and in turn was interrelated with other characteristics of the innovation. This dimension designates the ultimate users of the innovation: single individuals versus a group of workers requiring coordinated efforts. Further consideration of this dimension also leads to suggestions for improving the assessment of the degree of implementation.

The distinction labeled "individual versus work group" as users refers primarily to the extent of coordinated division of job roles necessary

among the hands-on users of the innovation. In the documents reviewed, sixteen innovations were used by individual workers, whose extent of implementation could vary independently from other parallel workers in the same organization. Some examples are a teacher using a curricular innovation within a single classroom, an employment service worker using a computerized job matching system, or a mental health therapist delivering one-on-one therapy.

Other innovations cannot be implemented by any single individual, but require a number of different, coordinated roles from various staff members, called here a "work group." For example, in order to implement a community-based residential Lodge for deinstitutionalized mental patients, staff members had to coordinate their efforts on a variety of tasks (Fairweather et al., 1974; Tornatzky et al, 1980). These tasks ranged from developing a number of potential residents into a socially functioning group, to purchasing or renting a building for the Lodge, and helping to find appropriate employment for the residents. Other examples occurred in the different types of work duties performed by the group of steel factory workers using a basic-oxygen furnace (Gold et al., 1980), or the redesign of job tasks performed by a variety of hospital workers including nurses, doctors, pharmacists, and medical records technicians, in order to replace handwritten records with a computerized hospital information system (Gall et al., 1975). A total of 31 innovations in our sample examined work groups.[7]

The distinction between individuals and work groups as users was associated with other characteristics of the innovation. Innovations requiring work groups also tended to be rated as of higher complexity ($r = .59$, $n = 45$), less divisible ($r = -.44$, $n = 41$), of higher initial cost ($r = .67$, $n = 19$), and somewhat less well specified in terms of subtask requirements ($r = -.18$, $n = 37$). In spite of the higher complexity and coordinated work roles required for innovations used by work groups, these implementation efforts were less likely to have provided training for the users ($r = -.32$, $n = 47$). Thus, a pattern of interrelated characteristics did emerge from the innovations examined in these documents, a pattern that appears to be usefully typified by the distinction between an individual and a work group as user.

Methodological criteria previously discussed are also associated with the type of user. As shown in Table 6, studies of innovations whose users were individuals were more likely to have developed an operational definition for degree of implementation, to have examined reliability and validity, and were much more likely to have a representative or full

TABLE 6
Quality Criteria for Implementation Measures, by Type of User

QUALITY CRITERIA	*USERS*	
	INDIVIDUALS (*n* = 16)	*COORDINATED WORK GROUP* (*n* = 31)
1) Operational Definition Present	88%	55%
2) Used Multiple Measuring Techniques	68%	80%
3) Reliability Examined	50%	36%
4) Validity Examined	50%	16%
5) Sampling — Representative or Full Census	64%	26%

census sample. In contrast, studies of innovations that required work groups were somewhat more likely to have used multiple measuring techniques, which is a result consistent with the higher complexity ratings for these innovations. These findings suggest that previously unrecognized differences in the characteristics of innovations, specifically, characteristics associated with an individual versus a work group as user, are important reasons for variation in methodological quality.

CONSTRUCTING IMPLEMENTATION MEASURES BY DIAGNOSING THE INNOVATION

It appears that researchers have more easily conceptualized and measured the implementation of innovations used by individuals, but have faltered when attempting to measure the behavior of a group. Needed improvement in the quality of implementation measures might come about with greater attention to a diagnosis of the nature of the innovation. This diagnosis should start with an explicit description of the components of the innovative equipment, program or policy. Who will do their work differently under conditions of full implementation? What new activities will be done by each type of worker? A separate issue, and desirable to keep distinct if possible, is what supporting changes in the organization are likely to be necessary to permit the

components of the innovation to be implemented? This issue enters the realm of implementation processes, which are outside the scope of this paper, but have been addressed by Scheirer, 1981 as well as by several others.

With a full description of the components of the innovation, measuring tools can be constructed that most adequately match each component. By planning for measuring the extent of implementation before the innovative effort begins, it is more likely that measures can tap into target behaviors as they occur rather then depending on ex post facto recollections having dubious reliability. Several sources of ongoing implementation monitoring data can help to avoid the potential bias from any single source, particularly from interviews with a single organizational informant. Depending on the activities required for each component, the implementation measure might be as simple as a use activity log, records of consumable supplies used, or a periodic "walk-through" by a trained observer.

With explicit measures well defined for each innovation component, construction of an aggregate implementation index should not present difficult problems. Empirical experience with the measures will probably be necessary to determine whether the innovation's components really do come together as a whole, or conversely, whether only a portion of the components are ever implemented as intended. In the latter case, the "innovation" may not be viable either as a researchable construct, or as a real entity capable of producing an intended effect.

The construction of more adequate measures for degree of implementation thus requires both an analysis of the innovation into its component parts, and synthesis of the separate measures into an overall index. Along with such systematic attention to the content of a measure, which will help to establish its validity, the researcher should carefully assess the repeatability of the data collection procedures, across time and across interviewers or coders, to establish reliability. Sampling considerations should include both the extent to which the adopting organizations are equally represented, and the randomness of sampled units (individuals or work groups) *within* adopting sites.

This review has revealed the limited scientific basis underlying the young field of implementation research. Researchers have charged ahead to claim "findings" before establishing a solid foundation in adequate measurement of the most basic construct, degree of implementation. Nevertheless, enough examples of good practice have been noted across various disciplines using a variety of measurement

techniques to attest to the possibility of improvement. With increased attention to the match between the components of each innovation and the measures collected to assess their degree of implementation, much more useful data should result. Adequate measures of the degree of implementation can become the treatment documentation essential for outcome evaluations, as well as a basis for fostering a cumulative body of replicable findings about innovation processes.

APPENDIX: LIST OF IMPLEMENTATION PROJECTS REVIEWED* (Grouped by Content Area of Innovation)

CRIMINAL JUSTICE PROJECTS

Anno, B. *American Medical Association's program to improve health care in jails.* (11 reports) Washington, D.C.: Blackstone Assoc. Inc, 1977-1981.

Greenfield, L. *High impact anti-crime program* (3 reports) McLean, VA: MITRE Corp., 1975.

Haapanen, R. *Youth service bureau: An evaluation of nine California youth service bureaus.* Sacramento, CA: California Youth Authority, 1980.

Larson, R. *Police AVM (automatic vehicle monitoring system).* (2 reports). Washington, D.C.: U.S. Dept. of Justice, 1977-1978.

Murray, C., Thomson D., & Israel, C. *UDIS: Deinstitutionalizing the chronic juvenile offender.* Washington, D.C.: American Institutes for Research, 1978.

Rasmussen, M., W. Muggli, & C. M. Crabill. *Evaluation of the Minneapolis crime prevention demonstration.* St. Paul, MN: Crime Control Planning Board, December, 1979.

Rezmovic, E. L. Program implementation and evaluation results: A reexamination of type III error in a field experiment. *Program Planning & Evaluation,* in press.

Schwartz, A. & Clarren, S. *Cincinnati (OH) team policing experiment—a summary report.* Washington, D.C.: Urban Institute, 1977.

System Sciences, Inc. *Evaluation of the treatment alternatives to street crime national evaluation program—phase 2 report.* Washington, D.C.: U.S. Dept. of Justice, January 1979.

(continued)

*Several projects are reported in multiple reports or papers. Only an overall project title is included in this list.

EDUCATION PROJECTS

Berman, P. & McLaughlin, M. W., and others. *Federal programs supporting educational change.* (8 volumes). Santa Monica, CA: Rand Corp., 1974-1977.

Crandall, D. *"Understanding change in school practice: Preliminary findings from a study of innovative implementation in local schools."* (Unpublished paper) Andover, MA: The NETWORK, Inc., June, 1981.

Darnell, C. D. Evaluation Report for the extent of implementation of the Jefferson Co., schools revised science program. Jefferson County, CO: Program Evaluation Dept., Jefferson Co. Schools, March 1979.

Education Turnkey Systems. *Case study of the implementation of PL 94-142.* Washington, D.C.: Education Turnkey Systems, Inc., May 1979.

Gersten, R. and others; Emrick, J. A. and others. *Implementation study of Direct Instruction.* (5 reports) Oregon: University of Oregon and J. A. Emrick and Associates, 1980 and 1981.

Gross, N., Giacquinta, J. B. & Bernstein, M. *Implementing organizational innovations: A sociological analysis of planned educational change.* New York: Basic Books, 1971.

Goodlad, J., Klein, M. F., et al. *Behind the classroom door.* Worthington, Ohio: Jones, 1970.

Hall, G. E. & Loucks, S. F. *Innovation configurations: Analyzing the adaptation of innovations.* R&D Center for Teacher Education. Univ. of Texas at Austin, 1978.

Hall, G. et al. Making change happen: A case study of school district implementation. (A symposium of 4 papers delivered at AERA Annual Meeting, Boston, 1980). Austin, TX: R& D Center for Teacher Education, Univ. of Texas at Austin, 1980.

Johnston, J. *An evaluation of Freestyle: A television series to reduce sex-role stereotypes.* Ann Arbor, MI: Institute for Social Research, Univ. of Michigan, 1980.

Leithwood, K. & Montgomery, D. Evaluating program implementation. *Evaluation Review* 4, 1980, 193-214.

Leinhardt, G. "Modeling and measuring educational treatment in evaluation" *Review of Educational Research* 50, Fall: 393-420, 1980; plus 6 other papers by Leinhardt.

Loucks, S., Newlove, B. & Hall, G. *Measuring levels of use of the Innovation: A manual for trainers, interviewers and raters (plus several papers on use of LoU methods).* Austin, TX: The Univ. of Texas, 1975.

Louis, K. et al. *Linking R & D with schools.* (6 papers). Cambridge, MA.: Abt Associates, 1979-1981.

Owens, T. R. & Haenn, J. F. "Assessing the level of implementation of new programs." Presented at AERA, New York City, 1977. Portland, OR: Northwest Regional Educational Laboratory.

Stallings, J. & Kaskowitz, D. *Follow-Through classroom observation evaluation, 1972-73.* Menlo Park, CA.: Stanford Research Institute, 1974.

Stearns, M. S., Greene, D. & David, J. L. *Local implementation of PL 94-142.* (2 reports). Menlo Park, CA.: SRI, April 1980.

St. Pierre, R. et al. *An evaluation of the Nebraska nutrition education and training program: Nebraska's experience nutrition curriculum.* Cambridge, MA.: Abt Associates, 1981.

Weatherley, R. & Lipsky, M. (1977) Street level bureaucrats and institutional innovation: implementing special-education reform. *Harvard Educational Review* 47, (May): 171-197.

Zigarmi, P., The implementation of a new approach to discipline in a junior high school: A case study of intervention during the process of change (unpublished paper), and Rutherford, W. and Loucks, S. F., Examination of the implementation of a junior high school's new approach to discipline by longitudinal analysis of changes in teachers' Stages of Concern and Levels of Use (unpublished). Both Austin, TX: University of Texas at Austin, R & D Center for Teacher Education, 1979.

HEALTH

Delbecq, A. & Pierce, J. (1978) Innovation in professional organizations. *Administration in Social Work* 2: 411-424.

ENERGY/ENVIRONMENT

Lewis, C. *Implementation of a microcomputer-modified electrical aerosol analyzer.* (Report prepared for Environmental Sciences Research Lab.) Research Triangle Park, N.C.: Environmental Science Research Lab., August 1979.

Maddalone, R. & Goarner, N. *Process measurement procedures: H_2SO_4 emissions.* (Report prepared for EPA). Redondo Beach, CA: TRW Defense & Space Systems Group, July 1979.

Marcus, A. *Promise and performance. Choosing and implementing an environmental policy.* Westport, Ct: Greenwood Press, 1980.

HUMAN SERVICES

Benton, B., Field, T. & Millar, R. *Social services: Federal legislation vs. state implementation.* Washington, D.C.: The Urban Institute, 1978.

Chesterfield, B. *An evaluation of the Head Start bilingual bicultural curriculum development project.* (Report prepared for HEW). Los Angeles, CA: Jua'rez and Associates, December 1979.

Collignon, F. & Shea, S. *Implementing the Rehabilitation Act of 1973: The VR program response.* (Report prepared for the Office of the Assistant Secretary for Planning & Evaluation). Berkeley, CA: Berkeley Planning Associates, February 1978.

Deloria, D., & Fellenz, P.; and Love, J., & Ruopp. *Home Start evaluation study.* (4 reports prepared for HEW). Ypsilanti, MI: High/Scope Educ. Research Foundation, Cambridge, Mass: Abt Associates, 1973 and 1979.

Love, J., Granvile, A., & Smith, A. *A process evaluation of project Developmental Continuity.* (8 reports prepared for HEW). Ypsilanti, MI: High/Scope Educational Research Foundation, April 1978.

Lukas, C. & Wohllieb, C. *Implementation of Head Start planned variation: 1970-1971.* (2 reports prepared for HEW). Cambridge, MA: Huron Institute, December 1972.

Macias, J., Levine, H., & Hays, W. *Field plan and implementation package for the ethnographic component of the child and family mental health project evaluation.* (Report prepared for HHS) San Francisco, CA.: The Urban Institute for Human Services, May 1981.

Monaghan, A. C. *An exploratory study of the match between classroom practice and educational theory (in Head Start).* (Report prepared for HEW). Cambridge, MA: Huron Institute, 1973.

Nauta, M. *Evaluation of the child and family resource program.* (4 reports). Cambridge, MA: Abt Associates, March 1981.

INFORMATION SERVICES

Alter, S. and Ginzberg, M. (1978) "Managing uncertainty in MIS implementation," *Sloan Management Review*, Fall: 23-31.

Bean, A. S., Neal, R. D., Radnor, M., & Tansik, D. A. Structural and behavioral correlates of implementation in U.S. business organizations. In R. L. Schultz and D. P. Slevin, Eds. Implementing Operations Research/ Management Science. New York: American Elsevier, 1975.

Eveland, J. D., Rogers, E. & Klepper, C. *The innovative process in public organizations.* Ann Arbor: University of Michigan, Dept. of Journalism, 1977.

Gall, J., M. Cook, J. Fleming, D. Norwood, R. Rydell & R. Watson *Demonstration & evaluation of a total hospital information system.* (2 reports prepared for HEW). Mountain View, CA.: El Camino Hospital, 1975 & 1980.

Ginzberg, J. J. A study of the implementation process. In R. Doktor, R. L. Schultz & D. P. Slevin, Eds., *The Implementation of Management Sciences* (TIMS studies in the Management Sciences, Vol. 13). Amsterdam, NY: North-Holland Publ. Co., 1979.

Lucas, H. C., Jr. The implementation of an operations research model in the brokerage industry. In R. Doktor, R. L. Schultz, & D. P. Slevin, Eds., *The Implementation of Management Sciences* (TIMS studies in the Management Sciences, vol. 13). Amsterdam, NY: North-Holland Publ. Co., 1979.

Narasimhan, R. & Schroeder, R. G. An empirical investigation of implementation as a change process. In R. Doktor, R. L. Schultz, and D. P. Slevin, Eds., *The Implementation of Management Sciences* (TIMS studies in the Management Sciences, Vol. 13). Amsterdam, NY: North-Holland Publ. Co., 1979.

Rubin, M., & Hunter, B., & Knetsch, M. *Evaluation of the experimental CAI network (1973-1975) of the Lister Hill National Center for Biomedical Communications, National Library of Medicine.* (Report prepared for the National Library of Medicine). Alexandria, VA.: HumRRO, January 1975.

Vertinsky, I., Barth, R. T. & Mitchell, V. F. A study of OR/ MS implementation as a social change process. In R. L. Schultz & D.P. Slevin, Eds., *Implementation Operations Research/ Management Science.* New York: American Elsevier, 1975.

Weimer, D. (1980) "CMIS implementation: a demonstration of predictive analysis". *Public Administrative Review* 40, May-June: 231-240.

LABOR/EMPLOYMENT

Lennox, K. & Feder, J. *Employer's Implementation of Dual Choice.* Washington, D.C.: Urban Institute, 1979.

Mitchell, J. *Implementing welfare-employment programs: an institutional analysis of the work incentive (WIN) program.* Washington, D.C.: U.S. Dept. of Labor, R & D Monograph 78, 1979.

Rohrbaugh, R. & Quinn, J. *Perspectives on change: a study of automated matching & local office performance (2 reports).* Rennsselaerville, NY: The Institute on Man and Science, July 1979 & 1980.

U.S. Dept. of Labor, Employment and Training Division. *CETA in Eastern Massachusetts and the implementation of CETA in Boston, 1974-1977.* R & D Monograph 57, 1978.

MENTAL HEALTH

Bass, R. *A method for measuring continuity of care in a community mental health center.* DHEW Publication No. (ADM) 76-377. 1972; reprinted 1976.

Beyer, J. & Trice, M. *Implementing change: alcoholism policies in work organizations.* New York: The Free Press, 1978.

Fabry, P. & Reid, D. Teaching foster grandparents to train severely handicapped persons. *Journal of Applied Behavior Analysis* 11, 1978, 111-123.

Fairweather, G., Sanders, D., Tornatzky L. & Harris, R. *Creating change in mental health organizations.* New York: Pergamon Press, 1974.

Greene, B., Willis, B., Levy, R. & Bailey, J. (1978) Measuring client gains from staff-implemented programs. *Journal of Applied Behavior Analysis,* 11 3: 395-412.

Moskowitz, J., Schaps, E., & Malvin, J. *A process and outcome evaluation of a magic circle primary prevention program* (report submitted to the Prevention Bureau, NIDA) Napa, CA.: Pacific Institute for Research and Evaluation, August 1980.

Nelson, G. & John, D. (1979) Multiple-baseline analysis of a token economy for psychiatric in-patients. *Journal of Applied Behavior Analysis* 12: 255-271.

Schaps, E., Moskowitz, J., Condon, J., & Malvin, J. *A process and outcome evaluation of an effective teacher training primary prevention program.* (Report to Prevention Branch, NIDA) Napa, CA.: Pacific Institute for Research and Evaluation, November 1980.

Scheirer, M. *Program implementation: The organizational context.* Beverly Hjlls, CA.: Sage, 1981.

Tornatzky, L. et al. *Innovation and social process.* New York: Pergamon, 1980.

PUBLIC ADMINISTRATION

Britain, G. M. *Bureaucracy and Innovation: An Ethnography of Policy Change*, Beverly Hills, CA: Sage, 1981.

Pressman, J. & Wildavsky, A. *Implementation.* Berkeley, CA.: Univ. of California Press, 1973.

Yin, R. *Changing urban bureaucracies: How new practices become routinized.* Lexington, MA.: Lexington Books, 1979.

TECHNOLOGY (hardware)

Burger, R., & Massaglia, M. *RANN utilization experience (Case studies 22 through 31).* Vol. II: Appendices containing case studies. (Report prepared for the National Science Foundation). Research Triangle Park, NC.: Research Triangle Institute, August 1976.

Gold, B., Rosegger, G. & Boylan, M. *Evaluating technological innovations: methods, expectations and findings.* Lexington, MA: Lexington Books, 1980. (Reviews done for 2 innovations).

Innovative systems research. *Program to conduct ongoing observations of federally premeditated actions to accelerate utilization of civilian oriented research and development.* (Report prepared for National Science Foundation). Pennsauken, NJ.: Innovative Systems Research, Inc., May 1979.

Ryan, T. *Blocks to effective technology transfer in construction.* Champaign, IL.: Construction Engineering Research Laboratory, December, 1978.

NOTES

1. This article draws upon a more extensive review of the measurement of degree of implementation (Scheirer and Rezmovic, 1982). Methods for locating relevant studies included: (1) use of abstracts available from reference services, such as the Smithsonian Science Information Exchange and the National Technical Information Service; (2) search of abstracts in Index Medicus, Business Periodicals Index and Sociological Abstracts; (3) custom searches by the National Criminal Justice Reference Center and the National Clearinghouse for Alcohol Information; and (4) personal contacts with a network of more than 40 individuals, primarily in government agencies.

2. Intercoder reliability was checked by duplicate, independent reviews of four documents. Using a conservative estimate, the percentage of exact agreement, the lower limits of inter-rater reliability ranged between 58% and 69%. Given the ambiguity or the incompleteness of information in the original reports, these reliability findings were judged to be adequate for this descriptive review.

3. The reader should keep in mind that the number of studies in most content areas is quite small; therefore the percentage distributions are likely to be unstable. They are presented here strictly to describe the results in the studies reviewed, rather than to be necessarily generalizable to any larger population of studies in each content area.

4. An "attempted full census" occurred when the researcher sought, but was unable to secure data from each unit in the target population, due to individual refusals, data missing in official records, and other problems.

5. This example was suggested by a perceptive reviewer.

6. Space precludes a full presentation of the findings about the characteristics of the innovations, which are detailed in Scheirer and Rezmovic, 1982. Coding of innovation characteristics was based only on the description of the innovation provided in each document reviewed. Frequently, the descriptions given were fragmentary, and the coders had to reach judgmental ratings for many characteristics. The ratings used a 4-point scale for the continuous dimensions, such as complexity and divisibility, and simple classification for other characteristics.

7. A number of studies could not be coded on this dimension either because information was missing or because both coordinated work group efforts and parallel individual delivery were required for separate components of a complex innovation.

REFERENCES

ANNO, B. (1977-1981) American Meidcal Association's Program to Improve Health Care in Jails. (5 Vols.). Washington, DC: Blackstone Assoc., Inc.

BERMAN, P. and M. W. McLAUGHLIN (1978) Federal Programs Supporting Educational Change: Implementing and Sustaining Innovations. (Vol. VIII) (Report prepared for the U.S. Office of Education, R-1589/8-HEW). Santa Monica, CA: RAND Corp.

BEYER, J. M. and H. M. TRICE (1978) Implementing Change: Alcoholism Policies in Work Organizations. New York: The Free Press.

BORUCH, R. F. and H. GOMEZ (1979) "Measuring impact: Power theory in social program evaluation," pp. 139-170 in L. Datta and R. Perloff (eds.) Improving Evaluations. Beverly Hills, CA: Sage.

BRITAN, G. M. (1981) Bureaucracy and Innovation: An Ethnography of Policy Change. Beverly Hills, CA: Sage.

CAMPBELL, D. T. and D. W. FISKE (1959) "Convergent and discriminant validity by the multitrait-multimethod matrix." Psychological Bulletin 56: 81-105.

COOK, T. J. and W. K. POOLE (1982) "Treatment implementation and statistical power: A research note." Evaluation Review 6: 425-430.

FAIRWEATHER, G. W., D. SANDERS, L. G. TORNATZKY, and R. N. HARRIS, Jr. (1974) Creating Change in Mental Health Organizations. New York: Pergamon Press.

GALL, J., M. COOK, J. FLEMING, D. NORWOOD, R. RYDELL and R. WATSON (1975) Demonstration and Evaluation of a Total Hospital Information System. Mountain View, CA: El Camino Hospital. (NTIS Document PB 262-106).

GERSTEN, R. M., D. W. CARNINE, and P. B. WILLIAMS (1982) "Measuring implementation of a structured educational model in an urban school district: An observational approach." Educ. Evaluation and Policy Analysis 4: 67-80.

GOLD, B., G. ROSEGGER, and M. G. BOYLAN, JR. (1980) Evaluating Technological Innovations: Methods, Expectations and Findings. Lexington, MA: Lexington Books.

KERLINGER, F. N. (1973) Foundations of Behavioral Research (2nd edition). New York: Holt, Reinhart & Winston.

KALUZNY, A. D. and J. E. VENEY (1973) "Attributes of health services as factors in program implementation." J. of Health and Social Behavior 14: 124-133.

MOSKOWITZ, J., F. SCHAPS, and J. MALVIN (1980) A Process and Outcome Evaluation of a Magic Circle Primary Prevention Program, Report submitted to the National Institute for Drug Abuse, Prevention Bureau. Napa, CA: Pacific Institute for Research and Evaluation.

MURRAY, C., D. THOMPSON, and C. ISRAEL (1978) UDIS: Deinstitutionalizing the Chronic Juvenile Offender. Washington, DC: Amer. Institutes of Research.

NAUTA, M. (1981) Evaluation of the Child and Family Resource Program. Cambridge, MA: Abt Associates.

NUNNALLY, J. C. (1967) Psychometric Theory. New York: McGraw-Hill.

NUNNALLY, J. C. and R. L. DURHAM (1975) "Validity, reliability and special problems of measurement in evaluation research," in E. Struening and M. Guttentag (eds.) Handbook of Evaluation Research (Vol. 1). Beverly Hills, CA: Sage.

OFFICE OF TECHNOLOGY ASSESSMENT (1982) Strategies for Medical Technology Assessment. Washington, DC: Congress of the United States, Office of Technology Assessment.

RASMUSSEN, M., W. MUGGLI, and C. M. CRABILL (1979) Evaluation of the Minneapolis Crime Prevention Demonstration. St. Paul, MN: Crime Control Planning Board.

ROSSI, P. H., H. E. FREEMAN, and S. R. WRIGHT (1979) Evaluation: A Systematic Approach. Beverly Hills, CA: Sage.

RUSSELL, L. B. (1979) Technology in Hospitals: Medical Advances and Their Diffusion. Washington, DC: Brookings Institute.

SCHEIRER, M. A. (1983) "Approaches to the study of implementation." IEEE Transactions on Engineering Management 30: 76-82.

SCHEIRER, M. A. and E. L. REZMOVIC (1982) Measuring the Implementation of Innovations, Report to the National Science Foundation, Grant No. PRA-8022612. Annandale, VA: Amer. Research Institute.

——— (1981) Program Implementation: The Organizational Context. Beverly Hills, CA: Sage.

——— (1978) Program participants' positive perceptions: Psychological conflicts of interest in social program evaluation." Evaluation Q. 2: 53-60.

SCHAPS, E., J. M. MOSKOWITZ, J. W. CONDON and J. MALVIN (1980) A Process and Outcome Evaluation of an Effective Teacher Training Primary Prevention Program, Report submitted to the National Training Institute for Drug Abuse, Prevention Branch. Napa, CA: Pacific Institute for Research and Evaluation.

SECHREST, L. and R. REDNOR (1979) "Strength and integrity of treatments in evaluation studies," in How Well Does it Work? Review of Criminal Justice Evaluation, 1978. Washington, DC: U.S. Department of Justice, National Criminal Justice Reference Service.

TORNATZKY, L. G., D. ROITMAN, M. BOYLAN, J. CARPENTER, J. D. EVELAND, W. H. HETZNER, and J. SCHNEIDER (1979) Innovation Processes and Their Management: A Conceptual Empirical and Policy Review of Innovation Process Research (A Strategic Planning Document of the Innovation Processes and Their Management Working Group). Washington, DC: National Sci. Foundation.

TORNATZKY, L. G., E. O. FERGUS, J. W. AVELLAR, G. W. FAIRWEATHER and M. FLEISCHER (1980) Innovation and Social Process: A National Experiment in Implementing Social Technology. New York: Pergamon.

WEBB, E. J., D. T. CAMPBELL, R. D. SCHWARTZ, and L. SECHREST (1966) Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally.

WHOLEY, JOSEPH S. (1979) Evaluation: Promise and Performance. Washington, DC: The Urban Institute.

YIN, R. K. (1979) Changing Urban Bureaucracies: How New Practices Become Routinized. Lexington, MA: Lexington Books.

V

VALIDITY CONSIDERATIONS IN DESIGNING EVALUATIONS

With this part of the *Annual*, we begin our consideration of papers that focus on various parts of the process of actually conducting an evaluation study. In previous sections, we have reviewed papers that address more general concerns of thinking about the evaluation enterprise, of selecting general evaluation approaches, or of the context in which evaluation is undertaken. In this section our attention shifts to the considerations an evaluator faces when he or she is in the first phase of designing an evaluation study. These considerations include the critical issue of the validity of the proposed study. Whether the central issue in the evaluation is one of need, process, or outcome, the evaluator wants a study design that will give accurate, valid information.

The goal of all scientific research is to uncover valid truths. Evaluation, as part of the scientific enterprise, also aspires to this goal, one that is easy to state but difficult to achieve. Perhaps because evaluation began as a blending of basic and applied research, evaluation researchers have always been centrally concerned with questions of validity, both internally in a study and externally toward other similar situations. Campbell and Stanley's (1966) classic evaluation work on experimental and quasi-experimental designs focuses on just such validity issues.

The more evaluation theorists have looked into the issue of validity, the more complex and multifaceted it has become. For example, in an update of the Campbell and Stanley classic by Cook and Campbell (1979), the number of validity types increased, as did the possible threats to each type. In addition, it became clear that researchers could not achieve high levels of all types of validity at the same time. If the internal validity of a study is high, for instance, the external validity tends to be low. Thus, the conscientious evaluation researcher, faced with the job of designing an evaluation study that has an acceptable mix of internal and external validity, confronts difficult and complex issues and choices.

The five papers in this section focus on different aspects of the validity issue. The first paper by Emrick and Hansen provides excellent examples of the types of confusion that can occur if validity issues are not thoroughly considered. Emrick and Hansen review the reasons for the wide disparity in the success rates attributed to different alcoholism treatment programs, some of which report abstinence rates among clients as high as 90 percent. The authors demonstrate a variety of external and internal validity problems with

alcoholism treatment studies. In the external validity domain, they discuss differences in the composition of the study samples, in the definition and measurement of outcome criteria, in the time of the evaluation measures, in the handling of treatment dropouts and of patients lost to follow-up, and in data analysis procedures. In the internal validity domain, Emrick and Hansen review such issues as investigator bias, demand characteristics, and the use of multiple sources of data. In the final section of their paper, Emrick and Hansen present a list of "core indices" that all alcoholism treatment program evaluations should use in order to increase the internal and external validity of these studies.

The next three papers in this section focus specifically on the issue of external validity. In articles by Lynch, by Calder, Phillips, and Tybout, and by McGrath and Brinberg, the question of the need for external validity in basic and applied research is addressed from several perspectives. While these authors are writing from the perspective of marketing and consumer research, the comments they make are equally pertinent for evaluation research, where we frequently are working on program evaluations that have basic (that is, theoretically derived) and applied aspects. Indeed, as McGrath and Brinberg state in their synthesis and extension of the Lynch and Calder et al. viewpoints, basic and applied research are interdependent. Consequently the field as a whole, whether it be consumer research or evaluation research, must be concerned with all types of validity if we are to have any confidence in the knowledge that the field produces.

The first two papers in this set are the most recent statements of these authors in a debate that began several years ago. Lynch's paper is a response to an earlier paper by Calder et al. (For the curious reader, citations to the earlier works are included in the bibliographies of each paper.) The papers presented here provide a sufficient introduction to the critical differences in opinion. Essentially, Lynch believes that theoretical research must concern itself with external validity. He advocates two main methods of assuring that some attention is given to external validity in any study: deliberate sampling for heterogeneity and the "selective approach" to choosing a small number of background factors to be varied orthogonally to the main treatments.

Calder, Phillips, and Tybout, on the other hand, do not see a need for researchers to be concerned with external validity during the basic research stage. They believe that the scientific process, and in particular the falsification principle of this process, are sufficient to develop applicable theories. When a researcher's interest moves to applying existing theories (as is sometimes the case in evaluation research), then external validity becomes important.

The final article in this set by McGrath and Brinberg presents a broader schema for analysis of validity and the research process generally. McGrath and Brinberg list several points on which Lynch and Calder et al. agree, then go on to describe their validity network schema. The complete research process consists of three stages in the McGrath and Brinberg formulation: (1) a

planning stage involving the development, clarification, and selection of elements and relations in the conceptual, methodological, and substantive domains; (2) the research study itself, which involves the combination and use of elements and relations from each of the three domains along one of three different paths (experimental, theoretical, or empirical); and (3) a testing stage involving the verification, extension, and delimiting of the findings from the previous stage. Validity is involved in each of these stages but takes on different meanings, from "validity as value" in Stage 1, through 12 different types of validity in Stage 2, to "validity as robustness" in Stage 3. The McGrath and Brinberg formulation raises important considerations for evaluation researchers no matter what the mix of basic and applied research aspects is in their work.

The final paper in this section is a critique of evaluation researchers' obsession with external validity and a defense of "external invalidity." Mook argues that when we are testing generalizations, as opposed to making generalizations, we do not need to concern ourselves with external validity. He discusses several well-known psychology studies that would flunk the external validity test but have produced useful, important knowledge. To determine the extent to which external validity ought to be a major concern, Mook provides a series of sample questions that researchers need to address prior to conducting their studies.

The papers in this section all argue for more careful attention by researchers to the expected purposes and uses of their study findings. One risk of the Campbell and Stanley (1966) listing of validity types, threats, and solutions was that researchers would make adjustments in and additions to their designs without giving careful thought to the need or purpose for these changes. All of the authors whose works appear in this section would urge evaluation researchers to think very carefully about the actual threats to validity that confront them in any particular study, then to develop solutions that are meaningfully responsive.

REFERENCES

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Design and analysis of quasi-experiments for field settings.* Chicago: Rand McNally.

20

Assertions Regarding Effectiveness of Treatment for Alcoholism

Fact or Fantasy?

Chad D. Emrick and Joel Hansen

Alcohol abuse and alcohol dependence are major national problems that affect an estimated 9–13 million Americans directly and up to 45 million others indirectly (e.g., family members, victims of automobile accidents; Brandsma, Maultsby, & Welsh, 1980). Alcohol-related problems cost our society tens of millions of dollars annually through lost work time, damage to property, and utilization of social welfare and medical services (Brandsma et al., 1980). Despite the vastness of the problem, it is estimated that less than 10% of those who are dependent on alcohol ever receive formal treatment for their dependency (Brandsma et al., 1980). Because there are no clearly established baseline data with which to characterize these millions of untreated people, no solid estimate can be made of how effective treatment is (if it is effective at all) in reducing alcohol dependence compared to no treatment. Nevertheless, there is considerable interest in knowing how alcohol treatment affects the lives of the apparently small percentage of alcohol-dependent people who do seek help. Vastly different impressions are obtained, depending on the source of information. National publications contain highly optimistic assertions of effectiveness. For example, the February 1983 issue of *Alcoholism, The National Magazine* contains advertisements that claim exceptional success for certain treatment programs. One ad reads, "The successes here have set the standard for addictive disease programs across the nation." Another states, "Our approach to treating the problems related to alcoholism and drug addiction is credited by most professionals for our remarkable success rate." Still another asserts, "No other treatment program for alcohol addiction can equal [our] record of effectiveness." Hunter (1982), in describing the program at Peachford Hospital's Addictive Disease Unit, Atlanta, Georgia (one of the programs that advertises in *Alcoholism, The National Magazine*), states that "for the patients who complete the program and continue in the aftercare groups we have achieved a better than 90 percent recovery rate" (p. 406). Inasmuch as a 90% recovery rate is an exceptionally high figure, the reader might infer that the Peachford Hospital's program is uniquely effective. Hunter proudly proclaims that

> the recognition of this treatment center, its goals, philosophy and success have spread nationally and internationally. This has resulted in the creation of a division within a national corporation to further expand and develop additional treatment centers based on the same medical model of treatment modality in different locations throughout the United States and abroad. (p. 407)

Although this program has sparked corporate growth, it remains to be proven that the treatment offered is *exceptionally* effective.

The results of the Rand report (Polich, Armor, & Braiker, 1980a, 1980b) contrast with claims of highly effective treatment. In their sample of alco-

From Chad D. Emrick and Joel Hansen, "Assertions Regarding Effectiveness of Treatment for Alcoholism: Fact or Fantasy?" *American Psychologist*, 1983, 38(10), 1078-1088.

hol-dependent patients ($N = 474$), only 7% were found to be abstinent following treatment throughout 4.5 years of evaluation. A similarly nonoptimistic evaluation of treatment effectiveness is found in a recent major review of alcohol treatment by Miller and Hester (1980). They state that for treated alcohol-dependent individuals, only 26% remain abstinent or at least improved in drinking behavior one year after treatment.

What can we make of such varying assertions about treatment of alcoholism? Are programs so different that some generate a 7% abstinence rate while others produce a 90% abstinence rate? Or is such variance a function of variables other than treatment per se? How are we to use these highly variant rates to chart our course in referring alcohol-dependent individuals to treatment and in creating public policy for funding alcohol-treatment efforts?

Generalizability of Data

The reasoned interpretation of any outcome figure requires an assessment of many factors outside the direct effects of treatment. Patient characteristics, sample selection and attrition, patient experiences outside of and after treatment, time of evaluation in relation to treatment, type and definition of criterion variables, measurement of criterion variables, and the analysis and interpretation of data all influence the size of treatment-outcome rates. We will discuss these variables here, but for a more thorough discussion we refer the reader to a number of excellent methodological reviews that pertain to alcohol-treatment evaluation (Baekeland, 1977; Blane, 1977; Costello, 1975; Crawford & Chalupsky, 1977; Hill & Blane, 1967; Jeffrey, 1975; Ludwig, 1973; Mandell, 1979; May & Kuller, 1975; Nathan & Lansky, 1978; Schuckit & Cahalan, 1976; Sobell, 1978; Sobell & Sobell, 1982; Solomon, 1981; Tuchfeld & Marcus, 1982; Voegtlin & Lemere, 1942).

Sample Composition

Sociopsychological characteristics and alcohol use variables have been shown to have a sizable impact on treatment-outcome rates. For example, Costello, Baillargeon, Biever, and Bennett (1980) found a large negative correlation (−.70) between scores on the Treatment Difficulty Scale (TDS) (Costello & Baillargeon, 1981) and overall successful outcome on six dimensions 12 months after treatment intake. The TDS is a measure of biosociopsychological and alcohol-use variables. Neubuerger, Hasha, Matarazzo, Schmitz, and Pratt (1981) followed up hospitalized patients 12 months after admission. Drinking behavior following treatment was observed to have a strong relationship with patient characteristics. For patients who were married and employed at admission, 73% had been either totally abstinent or had had only one or two drinks over the year of evaluation. On the other hand, patients who were on Medicare, under 62 years of age, and disabled had a 33% improvement rate using the same criterion.

McGuire (1982) conducted a nonrandomized study of drinking drivers who were referred by the courts to six different programs for treatment with the goal of improving drinking-and-driving habits or reducing the abuse of alcohol. The programs varied considerably in scope and intensity, ranging from a patient's receiving nine letters dealing with drinking and driving over a 13-week period to his or her receiving intense small-group and individualized medical and psychological treatment as well as education. Subjects were evaluated 24 months after starting treatment with respect to a number of drinking and driving criterion variables. "Heavy" drinking subjects were defined as those who possessed a blood alcohol level (BAL) of .17% or higher at the time of arrest and who had had one or more prior convictions for an alcohol-related violation. "Light" drinkers either had one or neither of these characteristics. Overall, light drinkers responded well to treatment *no matter which of the six programs they entered.* By contrast, heavy drinkers responded poorly on the criterion measures, again *no matter which of the six programs they entered.*

Similar findings, testifying to the power of patient variables on treatment response, are found in a study by McLellan, Luborsky, O'Brien, Woody, and Druley (1983). Male, veteran, alcohol-dependent patients were evaluated six months after admission to one of six programs. These programs varied considerably in scope, location, and intensity, ranging from a variable-length outpatient alcohol program to a 60-day therapeutic community oriented toward Alcoholics Anonymous to a methadone maintenance clinic with psychiatric and social work counseling. Patients who were rated high in severity of "psychiatric" disturbance at admission showed no improvement overall *in any of the six programs.* By contrast, patients who were rated low in "psychiatric" disturbance tended to respond well to treatment in *every* program.

In a review of alcoholism-treatment-outcome literature published from 1952 to 1971, Emrick (Note 1) found that the following patient characteristics predicted a positive response to treatment whenever a statistically significant relationship was observed: higher social class, employed, married, socially active, financially well situated, good work adjustment, good marital and family relationships, good social relationships, good residential adjust-

ment, good "general situation," no or minimal pretreatment arrest history, good physical condition, higher intelligence, good psychological insight, at least moderate self-acceptance, good motivation, previous outpatient treatment, diagnosed "normal," being cooperative during treatment, drinking none or a little during treatment, and having the spouse involved in treatment. Patient characteristics that more often than not predicted a negative response to treatment (whenever statistically significant relationships were observed) included having had previous inpatient treatment, being aggressive, having had suicide attempts, having an organic brain syndrome, and having a "sociopathic" personality disorder (Emrick, Note 1). Although many patient characteristics have been found to predict treatment outcome, none have been found to relate significantly and in the same direction every time they have been analyzed (Gibbs & Flanagan, 1977; Emrick, Note 1). Such lack of generalizability does not, however, negate the importance of the influence of patient variables on treatment outcome.

Whether patient characteristics are a more potent determinant of treatment outcome than are treatment variables is uncertain. Some investigators have yielded results suggesting that patient characteristics are the more powerful (e.g., Armor, Polich, & Stambul, 1978; Costello, 1980), but these findings have been criticized as more reflective of the statistical procedures used to arrive at the estimated proportional variance accounted for than of "true" relative predictive strength (Moos, Cronkite, & Finney, 1982). Path analyses have been used by some investigators to estimate more precisely the direct, indirect, and joint influences of patient and treatment variables on treatment outcome (Costello, 1980; Moos et al., 1982). Consistent with the search for indirect influences, some investigators have obtained data suggesting that patient characteristics have an indirect effect on treatment response by influencing the type of program or modality of treatment received (Cronkite & Moos, 1978; Obitz, 1978; Pattison, 1982).

Whatever the mechanism of influence of patient characteristics on treatment response, the fact that they are important determinants renders it necessary to describe samples in detail. At a minimum, patients should be described by age, sex, race, ethnicity, religion, marital status, family and social relationships, psychological functioning, legal status, nature of referral to treatment (i.e., whether court ordered or not), physical health status, education, employment and financial status, occupational level, and alcohol and other substance use. Shyrock, Siegel, and their associates (1976); Kaplan and Van Valey (1980); and Hansen (Note 2) present resources for developing valid and standardized approaches to collecting demographic data. The Personality Assessment Survey (PAS; see Wanberg, 1983), the Alcohol Use Inventory (AUI; see Wanberg, 1983), and background and current situation scales (see Wanberg, 1983) are recommended instruments for collecting self-report data on many dimensions of background characteristics, personality, current situation, and alcohol and other drug use patterns. These instruments are particularly appropriate inasmuch as they have been standardized on alcohol-dependent patient populations. Another instrument, the Addiction Severity Index (ASI; McLellan, Luborsky, Woody, & O'Brien, 1980; Parente, 1980), also developed particularly for alcohol and other drug-dependent patient populations, yields valid and reliable data with respect to medical condition, employment and financial situation, alcohol and other drug use, legal situation, family and social relationships, and psychological health. Evaluators should have sufficient data to provide a comprehensive description of patient samples if they collect standardized demographic data as well as administer the ASI or, in lieu of the ASI, the PAS, AUI, and background and current situation scales. Whatever methods are used to collect patient descriptors, they need to be reported so that other investigators, treatment personnel and policymakers can estimate the possible generalizability of the study to other treatment samples.

The importance of reporting assessment methods is illustrated by studies that yield different descriptions of patients depending on the procedure used. For example, Hesselbrock, Hesselbrock, Tennen, Meyer, & Workman (1983) assessed the presence of depression in alcohol-dependent inpatients using three measurement procedures. The Minnesota Multiphasic Personality Inventory (MMPI) Depression scale yielded a depression diagnosis in 62% of the sample. The Beck Depression Inventory (BDI) resulted in 54% being diagnosed depressed. The DSM-III criterion approach to diagnosis (using a structured interview) identified 27% as depressed.

Definition and Measurement of Outcome Criteria

The size of an outcome figure depends on the criterion selected for evaluation, how the criterion is defined, and how it is measured. A look at the definition and measurement of drinking amount and frequency will exemplify this point.

Having a "good" or "successful" drinking outcome has been variously defined as total abstinence, with no consideration given to any other kind of improvement such as significantly attenuated drinking (e.g., Stojiljković, 1969), not engaging in "destructive drinking" for at least 6 months during 2 years of evaluation (Ferguson, 1970), being either completely abstinent or having "no more than two

brief drinking episodes" during 2–20 months after inpatient discharge (Gallant, Rich, Bey, & Terranova, 1970), and "wife reported five or fewer weeks containing any episode of 'unacceptable' drinking" and "husband reported five or fewer weeks containing any 200-g or more per day drinking" in the 12–24-month period since treatment intake (Orford, Oppenheimer, & Edwards, 1976).

"Normal" ("controlled," "moderate," "nonproblem") drinking has been variously defined as drinking less than 5 oz. (148.5 ml) of ethanol on a typical day or averaging less than 1 oz. (29.5 ml) of ethanol per day over 30 days before follow-up (Ruggels, Armor, Polich, Mothershead, & Stephen, 1975), and "usually consumption of 6 oz or less of 86-proof liquor or its equivalent in alcohol content" (Sobell & Sobell, 1976).

Becoming "worse" in drinking amount or frequency has been defined as "increased drinking" (Boggs, 1967), "shorter periods of abstinence" (Aharan, Ogilvie, & Partington, 1967), "loss in months of abstinence" postadmission versus preadmission (Gibbins & Armstrong, 1957), loss in percentage of period abstinent postadmission versus preadmission (Smart, Storm, Baker, & Solursh, 1966), and alcohol "intake increased" (Tyndel, Fraser, & Hartleib, 1969).

With so much variability in the definition of drinking outcome, different studies are likely to yield discrepant findings where the discrepancy reflects variability in definition rather than in treatment effectiveness. An example of how important the criterion definition is to understanding recovery is found in the Rand report (Polich et al., 1980a, 1980b). Abstinence was defined in two ways: (a) no drinking for 6 months before the 18-month follow-up (long-term abstainers) and (b) no drinking for 1 to 5 months before the 18-month follow-up (short-term abstainers). Dramatic differences were found in the functioning of these two groups in the time between the 18-month follow-up evaluation and follow-up 4.5 years after treatment. Short-term abstainers were nine times more likely than the long-term abstainers to have died from alcohol-related causes. Also, short-term abstainers were significantly more likely to be problem drinkers at the 4.5 year follow-up vs. the long-term abstainers (29% vs. 12%, $\chi^2[1] = 8.67$, $p < .005$). In addition, 85% of the short-term abstainers were found to have had a serious alcohol problem when they last drank before the 4.5-year follow-up, a problem rate which was higher than that observed for those who were problem drinkers at the time of the 4.5-year follow-up.

Not only have drinking-outcome variables been differentially defined, data pertaining to them have been collected in numerous ways across studies. Many studies have relied on self-reported data alone through mailed questionnaires, telephone interviews, or face-to-face interviews. Others (e.g., Orford et al., 1976; Sobell & Sobell, 1976) have obtained information from collateral informants to check against patient self-reports. Still others (e.g., Miller, 1975; Polich et al., 1980a, 1980b; Sobell & Sobell, 1976) have used breath alcohol tests on a probe day basis to check against self-reports of drinking. Drinking outcome data have been collected using the time-line follow-back method as well as the quantity-frequency method (Sobell, Cellucci, Nirenberg, & Sobell, 1982). How one collects data may affect the size of whatever outcome figures are obtained.

Time of Evaluation

Time of evaluation has a major influence on outcome rates. The longer data are collected after treatment, the lower the rates of improvement tend to be on the average (Emrick, 1982). For example, at the end of intensive treatment, an abstinence rate (i.e., total and continuous abstinence from alcohol) of about 50% would not be exceptional, but after the termination of all treatment a rate of this size would be. Then the abstinence rate would more likely be on the order of 20% (Emrick, 1982). Of course, what length of evaluation is selected depends on the question one is asking. If the evaluator wants to investigate aspects of the process of treatment, a long-term follow-up is irrelevant. Also, if the evaluator wishes to assess the effectiveness of an initial, intensive in-hospital alcohol treatment program, a long-term follow-up is irrelevant inasmuch as the purpose of such programs is to diagnose and provide early treatment and then referral for medical and psychiatric conditions associated with or derived from alcohol dependence (Costello, 1982). Rehabilitation of the alcohol-dependent patient over the long haul is a function of other treatment services (viz., intermediate, outpatient, and aftercare facilities; Costello, 1982). If the evaluation question is, "Does the program work in some ultimate sense?" (Moos et al., 1982, p. 1132), a minimum follow-up interval of 12–18 months is suggested, as group outcome data do not begin to stabilize until then (see Sobell et al., 1980). Some investigators have suggested an even longer minimum period of evaluation (e.g., Nathan & Lansky, 1978) to ensure that a fairly stable estimate of treatment response is obtained. A lengthy period of assessment also provides the opportunity to identify patients who deteriorate after an initial positive response to treatment as well as those whose functioning is poor immediately posttreatment but improves over time (Moos et al., 1982; Nathan & Lansky, 1978). Of course, the longer the period of evaluation, the more posttreatment adjustment is mediated by factors other than treatment per se.

Posttreatment factors, such as life stressors, social support resources, and the patient's overall ability to cope with stress, have been shown to have a sizable influence on outcome adjustment (Moos et al., 1982). Thus, although a lengthy period of follow-up may cast informative light on the course of patients' drinking and related problems posttreatment, as the period of evaluation is lengthened less of what is observed can be attributed directly to treatment.

On the other side of the evaluation coin, pretreatment functioning needs to be assessed for comparison with posttreatment adjustment, using parallel pre–post data-gathering procedures. Again, the length of the period of evaluation may affect the size of improvement rates. For instance, a short pretreatment window of observation may capture an unrepresentative period of severe drinking that precipitates treatment. When compared to posttreatment functioning, such pretherapy data may create the impression of posttherapy improvement when in actuality patients have merely returned to their stable baseline level of functioning. Although there are difficulties in gathering valid self-report and collateral informant data retrospectively over an extended period of time, a one-year pretreatment interval is recommended for obtaining stable baseline data (Cooper, Sobell, Maisto, & Sobell, 1980).

Handling of Treatment Dropouts

The exclusion of patients who leave treatment very early (e.g., after less than 2 weeks of inpatient treatment and fewer than five outpatient visits) can result in higher improvement rates, since rapid dropouts may do less well than those who complete treatment. In a review of 384 studies, Emrick (1975) found that for alcohol-dependent individuals who had received less than 2 weeks of inpatient treatment or fewer than 5 outpatient visits, 40% had improved in drinking behavior over a period of 6 months or more. For individuals who had received more than minimal treatment, the overall drinking improvement rate was 63% over a comparable period of evaluation. Although a defensible argument can be made for excluding from evaluation patients who drop out from treatment very rapidly (since they might be considered as inadequately exposed to the benefits of therapy), patients who drop out of treatment later on, for reasons directly related to therapy, need to be evaluated. Otherwise, results are likely to be biased in the direction of the more successful patients (see Meltzoff & Kornreich, 1970, for a discussion of methods for dealing with dropouts).

Handling of Patients Lost to Follow-Up

How one handles patients who are lost to follow-up can effect treatment outcome findings. Since follow-up completion rates are often less than 75% (Moos & Bliss, 1978), the assumptions made about lost patients and the analytic procedures that reflect those assumptions can have a significant impact on outcome rates. If one assumes that all patients who are lost to follow-up are doing poorly and are therefore to be regarded as treatment failures, the evaluator may arrive at a conservative estimate of improvement. The evaluator may choose, on the other hand, to exclude from analysis anyone who is lost to follow-up and treat the obtained data as reflecting the treatment sample as a whole. This procedure may be likely to generate somewhat inflated estimates of improvement, inasmuch as patients who are difficult to interview after being contacted may fare less well than those who are more easily interviewed, despite the two subgroups having similar sociodemographic characteristics and similar functioning at intake characteristics (see Maisto & Cooper, 1980, for a review of the pertinent literature).

Inasmuch as studies with high follow-up completion rates are likely to generate less-biased data, evaluators should strive to keep to a minimum the number of patients who are lost to follow-up. Follow-up rates can be maximized by the evaluator (a) preparing patients before the end of treatment about the nature of the follow-up to be conducted, (b) obtaining at the start of treatment information about several individuals who can aid in tracking patients after they leave treatment, as well as obtaining permission to contact those individuals, and (c) having a familiar individual contact patients at least once a month (Caddy, 1980; Maisto & Cooper, 1980; Sobell et al., 1980).

Data Analysis Procedures

Varying statistical methods cast varying lights on treatment evaluation data. For example, in a study by Fuller and Williford (1980), a chi-square test yielded no statistically significant differences between three treatment groups in the percentage of patients who were abstinent for 12 months after the start of treatment. Analysis of the same data by life-table tests yielded significant differences between the groups. Furthermore, different life-table tests yielded varying results. One test (Generalized Wilcoxon) generated a statistically significant result when all three groups were compared as well as when two groups were combined and compared with the third, whereas the other test (Generalized Savage) yielded significant results only for the latter comparison. In another study Finney, Moos, and Chan (1981) found that length of stay was significantly negatively correlated with rehospitalization for halfway house patients when using a partial correlation technique, yet was not significantly correlated when using a treatment-effect correlation technique. Inasmuch as study

results are, in part, a function of the procedures used to analyze data, evaluators need to report precisely what methods they employed so that other investigators can assess the findings more fully and perhaps conduct cross-validation studies.

Internal Validity of Data

A number of variables have been discussed that affect the generalizability of data, or external validity. However, even the most generalizable of data will be at best useless and at worst misleading if they are not valid with respect to the sample under investigation (i.e., have internal validity). Efforts must be undertaken to maximize this latter type of validity, and the results of studies can be awarded more or less confidence depending on the steps taken or not taken to increase it. Some of the more notable means of improving internal validity are addressed in this section.

Investigator Bias

Follow-up needs to be conducted by evaluators who are neutral with respect to the results of the investigation yet are not so distant from the treatment environment that they lack necessary contextual understanding for obtaining comprehensive and comprehensible data. Perhaps the best arrangement is to have the evaluator "separated from the therapeutic effort but, nevertheless, integrated in some way within the treatment service so that follow-up staff know the patients whom they will follow-up and are, in turn, known by these patients" (Caddy, 1980, p. 168). Unfortunately, research on treatment of alcoholism characteristically uses as follow-up investigators program staff who are personally identified with the treatment under review. Such investigators have a high potential for bias when observing, measuring, analyzing, interpreting, and reporting treatment-outcome data (Rosenthal, 1969).

Demand Characteristics

Interviews need to be conducted in a manner that minimizes data distorting demand characteristics for patients and other interviewees (Orne, 1969). Patients need to be approached nonjudgmentally, particularly with respect to reports of drinking behavior. The interview climate needs to be one of "concerned data gathering" (Caddy, 1980, p. 166), and patients must be reinforced for the factual reporting of outcome adjustment, not for giving a report that is favorable to the treatment agency. Unfortunately, a considerable amount of program-effectiveness data misses the mark on this point. For example, as a matter of clearly stated policy, some treatment programs provide care only so long as patients remain totally abstinent from alcohol. If patients face the loss of an important relationship with the treatment program should they drink and report this event to a staff member who is evaluating the program, underreporting of drinking relapses is encouraged. A similar press works against patients who are mandated to treatment by the legal system or by employers, inasmuch as a report of resumed drinking from the program to the mandating agent poses the threat of legal consequences or job loss. Patients and other informants need to be assured that whatever information is given will have no negative effect whatsoever on the treatment, legal status, or job security of the patient. Patients may also be tempted to provide answers they think will please the interviewer. An example of this response bias is found in a study by Obitz, Wood, and Cantergiani (1977). Half of 64 patients were asked by an individual who was known as a recovering alcoholic whether they preferred group therapy meetings or Alcoholics Anonymous meetings while in the hospital. Another interviewer who was identified as a nonalcoholic asked the other half the same question. Patients more often told the recovering alcoholic that they preferred Alcoholics Anonymous meetings over group therapy, whereas group therapy was stated as the preferred intervention when the nonalcoholic asked the question, $\chi^2(1) = 20.12$, $p < .001$.

Multiple Sources of Data

Patient self-report data need to be compared with other sources of data. For example, significant others can be interviewed about a patient's functioning, and such reports can be compared with those of the patient. When reports of patients and others disagree, the latter do not necessarily reflect the "true" criterion (Midanik, 1982). The degree and direction of discrepancy appear to be a function of the sample under investigation, the kind of collateral interviewed (e.g., spouse, therapist), the frequency of contact between the patient and collateral, and the type of information sought (Midanik, 1982).

Also, self-report drinking data can be compared with biological markers. Chemical measures of blood alcohol are, of course, only valuable in checking reports of very recent drinking, and the test results are not always accurate (Maisto & Cooper, 1980; Sobell & Sobell, 1980). Liver function tests and, perhaps, the sweat patch test (Phillips & McAloon, 1980) can be helpful in corroborating reports of drinking over a longer period of time (e.g., two to four weeks for liver function tests; Sobell & Sobell, 1980). These tests are not, on the other hand, sensitive to a brief period of drinking.

Since self-report data are likely to be more accurate when information is sought at a time the patient is drug or alcohol free (Sobell, Sobell, & VanderSpek, 1979), a breath test should be given as

a routine matter before an interview to ensure that the patient is at least free of alcohol. Ideally a urine or blood sample would also be obtained at the time of interview to assess information about the presence of other drugs. Interviews later determined to have been obtained while the patient was under the influence of alcohol or drugs (other than those prescribed for a medical illness) can be repeated or at least reported separately from other interview data. Another means of validating patient self-report can be the search of police, hospital, employment and other agency records, although such records are not always accurate (Sobell & Sobell, 1980). Whatever the source of data, to the extent that findings converge across sources, increased confidence can be placed in their validity (see, e.g., Sobell & Sobell, 1980). However, it must always be remembered that convergence does not *guarantee* validity.

Selection of Evaluation Criteria

Beside the issues of internal and external validity of data, treatment assessments need to be judged with respect to breadth of investigation. In particular, the greater the number of process and outcome variables included in a study (within reason), the more one can learn about a form of therapy. Although it is superficially compelling to argue that the ultimate mark of successful alcohol treatment is total and permanent abstinence from alcohol, several investigators have argued that this criterion alone in no way provides an adequate picture of the effects of treatment, inasmuch as abstinence does not assure adequate or even marginal functioning in other life areas. Pattison (1966) observed that "abstinence may, but does not necessarily, indicate degrees of rehabilitation, or it may be responsible for the deterioration of personality functioning" (p. 62). Solomon (1981) writes, "even if drinking behavior is changed, there may still remain psychosocial disabilities that preclude successful rehabilitation" (p. 5). Consistent with these statements, in a now classic report, Gerard, Saenger, and Wile (1962) classified 50 male alcohol-dependent individuals, who had been abstinent for at least one year prior to follow-up, according to their overall psychosocial adjustment. Only 10% (n = 5) were rated as "independent successes," while 12% (n = 6) were successfully dependent on Alcoholics Anonymous. The remaining 78% were considered to be psychosocially disturbed, either overtly (54% of the total, n = 27) or inconspicuously (24%, n = 12). Wanberg (1983) found in a sample of 170 alcoholic inpatients that drinking behavior at one year follow-up accounted for *at most* 18% of the variance in functioning in eight other dimensions such as employment and health.

Because of the nonorthogonal relationships among treatment outcome domains, use of multiple outcome measures is essential. Nevertheless, drinking behavior should not be ignored when evaluating alcohol treatment. As Pattison (1966) asserts, abstinence "is not to be disregarded but should be placed in appropriate perspective along with other parameters of health and adaptation" (p. 66). Drinking behavior is a target symptom of alcohol rehabilitation treatment and thus needs assessment in its own right. Also, change in drinking behavior tends to relate in commonly expected ways with change on other dimensions even though the relationships are less than parallel. In his review of 265 treatment-outcome studies, Emrick (1974) found that improvement in drinking behavior was positively associated (beyond statistical chance) with improvement in the following life areas: affective-cognitive, work situation, interpersonal relationships, physical condition, legal situation, and social situation. On the basis of this finding he concluded that "drinking behavior can be used as a major criterion in alcoholism studies" (p. 529).

Consistent with the necessity for conducting multivariate analyses that include drinking behavior as a major criterion, alcohol treatment evaluation efforts undertaken in the last three decades have often used multivariate analyses with a major emphasis on drinking behavior. Not only have a broad range of life areas been examined in addition to drinking behaviors, but also a considerable number of variables have been used to measure outcome within each life area (Emrick, 1974; Maisto & McCollam, 1980).

Although a seemingly unlimited array of outcome criteria can be considered appropriate for evaluation, practical considerations render it necessary to select a limited number of specific indexes. In the interest of developing standards for evaluating alcohol treatment, the following criteria are suggested as core indexes to be used for all treatment-evaluation studies. Other criterion measures can be selected that particularly pertain to the sample under review or which are of special interest to the investigator.

Treatment completion is defined as a patient's completing treatment as judged by the primary care giver. This criterion can be assessed by entries in patient records noting the discharge status. Of course, this criterion will be inappropriate for evaluating treatment agencies where treatment is seen as a continuous process with no end point until the patient leaves treatment by self-selection, moving, or death.

Recidivism is defined as the number of subsequent entries into treatment for psychoactive substance abuse in a setting that is at least as restrictive as the setting in which the patient became involved

in the project. Patient self-report, collateral information, and treatment agency records could be used to measure this criterion.

Mortality is defined as physical death and is measured by the time from treatment admission to the day of death. Whenever possible, death certificates should be obtained to document the date of death and to receive information on the cause of death. Whenever possible, the death should be classified as directly related to alcohol abuse (e.g., death due to lethal interaction of alcohol and prescribed medication or a fatal accident with high blood alcohol level), indirectly related (e.g., death by pancreatitis), or unrelated (e.g., death by automobile accident when patient had a zero blood alcohol level, or death by influenza).

Treatment use is defined as the use of any treatment services for any medical problem, including psychoactive substance use/abuse. Patient self-report plus examination of treatment agency or insurance records could be used to measure this criterion.

Physical health is defined as the number of days the patient experiences medical problems, taking prescribed medication on a regular basis for a physical problem, being hospitalized for a physical problem, receiving a pension for physical disability, and patient's perception of the need for treatment for a medical problem. The Addiction Severity Index (ASI; McLellan et al., 1980) could be used to measure this.

Drinking behavior is defined as the number of days the patient is abstinent because of hospitalization, incarceration, or residential treatment; abstinent without environmental constraints; drinking moderately (less than 3 oz. [89 ml] ethanol per day); drinking heavily (more than 3 oz. [89 ml] ethanol per day); and not drinking because of prescribed medication that prohibits drinking. This criterion should be assessed by interviewing patients using the time-line follow-back method inasmuch as this method has been demonstrated to yield the most complete data (Sobell et al., 1982) and has demonstrated reliability and validity (Sobell et al., 1980). Collateral informant data and chemical test results could also be used to measure this criterion. In addition, the ASI could be used as a measurement instrument.

Other substance use is defined as the number of days the patient is abstinent from substances because of hospitalization, incarceration, or residential treatment; abstinent from substances without environmental constraints; abusing substances; and using substances without problems, including using according to physician's orders. This criterion refers to all psychoactive substances other than alcohol, whether illicit or prescribed by a physician. The time-line follow-back method, collateral informant data, chemical tests, and the ASI could be used to measure this index.

Legal problems are defined as the number of arrests or charges of any kind for alcohol-related and non-alcohol-related reasons as well as the number of days the patient engaged in illegal activities for profit. Patient self-report data, collateral informants, the ASI, and examination of public records can be used to assess this criterion.

Vocational functioning is defined as employment status, number of days worked, sources of income, patient's perception of employment problems, and patient's perception of the need for employment counseling during the period of observation. Collateral information (e.g., through employer interview) as well as the ASI could be used to assess this criterion.

Family/social functioning is defined as the patient's satisfaction with his or her interpersonal and recreational life. This criterion could be measured with the ASI.

Emotional functioning is defined as the patient's self-report of common psychiatric symptoms and perceived need for psychiatric treatment. This criterion could be measured using the interview schedule developed by McLellan et al. (1980). Other instruments commonly used in psychological and psychiatric practice (e.g., the Sixteen Personality Factor Questionnaire, the MMPI, the Symptom Checklist 90) could also be used to assess this index. The Personality Assessment Survey (PAS) (see Wanberg, 1983) might be particularly appropriate for evaluating this criterion, because it was developed using alcohol-dependent patients.

Although the ASI has been frequently cited as a useful instrument for measuring most of these core criteria, the Treatment Assessment Survey (TAS) (see Wanberg, 1983) can also be used to assess them. In addition to the core indexes, the TAS assesses marital adjustment and posttreatment perceptions of the nature and degree of involvement in treatment as well as its benefit.

Life stressors: Beside the core criteria, collection of data regarding a patient's life experiences encountered outside of and after treatment is strongly encouraged because some research has shown that an alcohol-dependent patient's outcome adjustment is influenced directly or indirectly by such experiences (Moos et al., 1982). The Life Events Questionnaire (Horowitz, Schaefer, Hiroto, Wilner, & Levin, 1977) could be used to measure the amount and recency of life stressors. Home, work, and community support environments encountered during and after treatment could be assessed with reliable and valid instruments developed by Moos and his colleagues (see Moos et al., 1982).

Discussion

A number of variables that affect treatment outcome rates have been identified and discussed, and procedures for strengthening the internal and external validity of evaluation data have been noted. Several criteria have been recommended for conducting comprehensive assessments of interventions. A context has now been established for addressing the questions posed at the beginning of this article.

Statements asserting a certain "success" or "cure" rate for a particular mode, type, or setting of treatment are scientifically meaningless when they are contextually removed from an array of variables known to influence outcome rates and when they are not derived from procedures designed to maximize internal validity. Beyond being meaningless, such rates are often misleading to consumers and policymakers. Assertions of effectiveness based on these rates may make good business sense, but they fail to generate scientific understanding about the benefits or harm of a particular treatment. A cautious response to the presentation of any treatment outcome rate requires that we ask several questions. What does the rate tell us about the effectiveness of the treatment under study? Which patients were studied? What problems did they have? At what time in the course of their disorders were data gathered? Which measurement criteria were used? Which measurement procedures were used? Who were the evaluators? What source provided the data? At what time in relation to treatment were data gathered? Which patients were included in the final data analyses? Obviously few programs conduct evaluations and report them in enough detail for others to obtain complete answers to all these questions. As a result, scientists, consumers, and policymakers are left to make judgments about treatment with only partial information and consequently are vulnerable to distorted impressions.

It would be unfortunate if, for instance, the 90% abstinence rate reported at the beginning of this article were separated from the data on which it was based and then used to proclaim that the program has a 90% recovery rate. The figure pertains only to those patients who completed an inpatient treatment program and who were then actively involved in aftercare groups. Because these patients comprised only 9.8% of the total number of patients treated, the outcome rate is highly unlikely to represent the program's population. For example, active aftercare patients may have possessed unique qualities (e.g., motivation) that made them particularly responsive to treatment (Solomon, 1981). Consistent with this possibility, the total abstinence rate for all treated patients was 60% (499/832)—a figure derived from direct and indirect information—with some patients being assessed by interviewing staff members only. Unfortunately, besides age and sex no information was reported regarding the characteristics of the sample. Data were collected, it appears, at varying lengths of time after hospital discharge but no longer than a year afterward. Patients were evaluated only on the criteria of drinking behavior (abstinence vs. drinking) and rehospitalization. Information appears to have been obtained by an individual who was personally identified with the program. It seems that interviewees knew that the interviewer had an investment in their being totally abstinent following treatment. Furthermore, although some of the patients who were actively involved in aftercare treatment appear to have been treated twice in the program, measurement of their drinking behavior occurred only subsequent to the second admission. For all treated patients, nearly one third had had a second admission. In effect, then, the better than 90% abstinence rate appears to be based on a highly biased subgroup of patients, some of whom may have been treated twice, who were interviewed in a manner conducive to data-distorting demand characteristics and who were judged to be abstinent by an evaluator with a high potential for experimenter bias.

In contrast to the validity problems with the 90% abstinence figure, the 7% long-term abstinence rate reported by Polich et al. (1980b) was based on a virtually random sample of patients treated at 8 NIAAA-funded alcohol programs. Eighty-five percent of the sample was interviewed or found to be deceased. Patients were described on numerous sociodemographic and intake functioning characteristics. Self-reports were compared with reports from collaterals on a random subsample, and blood alcohol content was measured for 95% of those interviewed. Interviews were conducted by people who were not personally identified with the treatment under study. Numerous criteria were employed to evaluate outcome (e.g., abstinence, drinking with symptoms of dependence, negative consequences from drinking, employment functioning, marital functioning, financial situation). Thus, the 7% rate appears to be internally valid and generalizable to treatment populations similar to those seen at the NIAAA-funded agencies. The 90% rate appears to be more fantasy; the 7% figure, more factual.

This article has stressed the need to collect reliable and valid treatment-outcome data on representative samples of patients. Only when assertions of treatment effectiveness are based on these kinds of data can they be considered more factual than fanciful. With factual data, pursuit of many scientific issues such as optimal patient–treatment matches becomes possible, and the effectiveness of treatment of alcoholism can be enhanced.

REFERENCE NOTES

1. Emrick, C. D. *The influence of patient–therapist–treatment factors on outcome and staying in treatment.* Unpublished manuscript, Columbia University, 1973.
2. Hansen, J. *Demographic information questionnaire for alcoholism treatment evaluation research.* Unpublished manuscript, 1983. (Available from Joel Hansen, Department of Psychology, Colorado State University, Ft. Collins, Colorado 80523.)

REFERENCES

Aharan, C. H., Ogilvie, R. D., & Partington, J. T. Clinical indications of motivation in alcoholic patients. *Quarterly Journal of Studies on Alcohol,* 1967, *28,* 486–492.

Armor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment.* New York: Wiley, 1978

Baekeland, F. Evaluation of treatment methods in chronic alcoholism. In B. Kissin & H. Begleiter (Eds.), *The biology of alcoholism: Treatment and rehabilitation of the chronic alcoholic* (Vol. 5). New York: Plenum Press, 1977.

Blane, H. T. Issues in the evaluation of alcoholism treatment. *Professional Psychology,* 1977, *8,* 593–608.

Boggs, S. L. Measures of treatment outcome for alcoholics: A model of analysis. In D. J. Pittman (Ed.), *Alcoholism.* New York: Harper & Row, 1967.

Brandsma, J. M., Maultsby, M. C., Jr., & Welsh, R. J. *Outpatient treatment of alcoholism: A review and comparative study.* Baltimore, Md.: University Park Press, 1980.

Caddy, G. R. A review of problems in conducting alcohol treatment outcome studies. In L. C. Sobell, M. B. Sobell, & E. Ward (Eds.), *Evaluating alcohol and drug abuse treatment effectiveness: Recent advances.* New York: Pergamon Press, 1980.

Cooper, A. M., Sobell, M. B., Maisto, S. A., & Sobell, L. C. Criterion intervals for pretreatment drinking measures in treatment evaluation. *Journal of Studies on Alcohol,* 1980, *41,* 1186–1195.

Costello, R. M. Alcoholism treatment and evaluation: In search of methods. *International Journal of the Addictions,* 1975, *10,* 251–275.

Costello, R. M. Alcoholism treatment effectiveness: Slicing the outcome variance pie. In G. Edwards & M. Grant (Eds.), *Alcoholism treatment in transition.* Baltimore, Md.: University Park Press, 1980.

Costello, R. M. Evaluation of alcoholism treatment programs. In E. M. Pattison & E. Kaufman (Eds.), *Encyclopedic handbook of alcoholism.* New York: Gardner Press, 1982.

Costello, R. M., & Baillargeon, J. G. Reliability analysis of the treatment difficulty scale. *American Journal of Drug and Alcohol Abuse,* 1981, *8,* 117–121.

Costello, R. M., Baillargeon, J. G., Biever, P., & Bennett, R. Therapeutic community treatment for alcohol abusers: A one-year multivariate outcome evaluation. *International Journal of the Addictions,* 1980, *15,* 215–232.

Crawford, J. J., & Chalupsky, A. B. The reported evaluation of alcoholism treatments, 1968–1971: A methodological review. *Addictive Behaviors,* 1977, *2,* 63–74.

Cronkite, R. C., & Moos, R. H. Evaluating alcoholism treatment programs: An integrated approach. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 1105–1119.

Emrick, C. D. A review of psychologically oriented treatment of alcoholism: I. The use and interrelationships of outcome criteria and drinking behavior following treatment. *Quarterly Journal of Studies on Alcohol,* 1974, *35,* 523–549.

Emrick, C. D. A review of psychologically oriented treatment of alcoholism: II. The relative effectiveness of different treatment approaches and the relative effectiveness of treatment versus no treatment. *Journal of Studies on Alcohol,* 1975, *36,* 88–108.

Emrick, C. D. Evaluation of alcoholism psychotherapy methods. In E. M. Pattison & E. Kaufman (Eds.), *Encyclopedic handbook of alcoholism.* New York: Gardner Press, 1982.

Ferguson, F. N. A treatment program for Navaho alcoholics: Results after four years. *Quarterly Journal of Studies on Alcohol,* 1970, *31,* 898–919.

Finney, J. W., Moos, R. H., & Chan, D. A. Length of stay and program component effects in the treatment of alcoholism: A comparison of two techniques for process analyses. *Journal of Consulting and Clinical Psychology,* 1981, *49,* 120–131.

Fuller, R. K., & Williford, W. O. Life-table analysis of abstinence in a study evaluating the efficacy of disulfiram. *Alcoholism: Clinical and Experimental Research,* 1980, *4,* 298–301.

Gallant, D. M., Rich, A., Bey, E., & Terranova, L. Group psychotherapy with married couples: A successful technique in New Orleans Alcoholism Clinic patients. *Journal of the Louisiana State Medical Society,* 1970, *122,* 41–44.

Gerard, D. L., Saenger, G., & Wile, R. The abstinent alcoholic. *Archives of General Psychiatry,* 1962, *6,* 83–95.

Gibbins, R. J., & Armstrong, J. D. Effects of clinical treatment on behavior of alcoholic patients: An exploratory methodological investigation. *Quarterly Journal of Studies on Alcohol,* 1957, *18,* 429–450.

Gibbs, L., & Flanagan, J. Prognostic indicators of alcoholism treatment outcome. *International Journal of the Addictions,* 1977, *12,* 1097–1141.

Hesselbrock, M. N., Hesselbrock, V. M., Tennen, H., Meyer, R. E., & Workman, K. L. Methodological considerations in the assessment of depression in alcoholics. *Journal of Consulting and Clinical Psychology,* 1983, *51,* 399–405.

Hill, M. J., & Blane, H. T. Evaluation of psychotherapy with alcoholics: A critical review. *Quarterly Journal of Studies on Alcohol,* 1967, *28,* 76–104.

Horowitz, M., Schaefer, C., Hiroto, D., Wilner, N., & Levin, B. Life event questionnaires for measuring presumptive stress. *Psychosomatic Medicine,* 1977, *39,* 413–431.

Hunter, C., Jr. Freestanding alcohol treatment centers—A new approach to an old problem. *Psychiatric Annals,* 1982, *12,* 396–408.

Jeffrey, D. B. Treatment evaluation issues in research on addictive behaviors. *Addictive Behaviors,* 1975, *1,* 23–36.

Kaplan, C. P., & Van Valey, T. L. *Census '80: Continuing the factfinder tradition.* Washington, D.C.: U.S. Department of Commerce, Bureau of the Census, 1980.

Ludwig, A. M. The design of clinical studies in treatment efficacy. In M. E. Chafetz (Ed.), *Proceedings of the 1st annual alcoholism conference of the National Institute on Alcohol Abuse and Alcoholism: Research on alcoholism: Clinical problems and special populations, Washington, D.C., June 25–26, 1971.* Rockville, Md.: National Institute on Alcohol Abuse and Alcoholism, 1973.

Maisto, S. A., & Cooper, A. M. A historical perspective on alcohol and drug treatment outcome research. In L. C. Sobell, M. B. Sobell, & E. Ward (Eds.), *Evaluating alcohol and drug abuse treatment effectiveness: Recent advances.* New York: Pergamon Press, 1980.

Maisto, S. A., & McCollam, J. B. The use of multiple measures of life health to assess alcohol treatment outcome: A review and critique. In L. C. Sobell, M. B. Sobell, & E. Ward (Eds.), *Evaluating alcohol and drug abuse treatment effectiveness: Recent advances.* New York: Pergamon Press, 1980.

Mandell, W. A critical overview of evaluations of alcoholism treatment. *Alcoholism: Clinical and Experimental Research,* 1979, *3,* 315–323.

May, S. J., & Kuller, L. H. Methodological approaches in the evaluation of alcoholism treatment: A critical review. *Preventive Medicine,* 1975, *4,* 464–481.

McGuire, F. L. Treatment of the drinking driver. *Health Psychology,* 1982, *1,* 137–152.

McLellan, A. T., Luborsky, L., O'Brien, C. P., Woody, G. E., & Druley, K. A. Predicting response to alcohol and drug abuse treatments: Role of psychiatric severity. *Archives of General Psychiatry,* 1983, *40,* 620–625.

McLellan, A. T., Luborsky, L., Woody, G. E., & O'Brien, C. P. An improved diagnostic evaluation instrument for substance abuse patients: The Addiction Severity Index. *Journal of Nervous and Mental Disease,* 1980, *168,* 26–33.

Meltzoff, J., & Kornreich, M. *Research in psychotherapy.* New York: Atherton Press, 1970.

Midanik, L. The validity of self-reported alcohol consumption and alcohol problems: A literature review. *British Journal of Addiction,* 1982, *77,* 357–382.

Miller, P. M. A behavioral intervention program for chronic public drunkenness offenders. *Archives of General Psychiatry,* 1975, *32,* 915–918.

Miller, W. R., & Hester, R. K. Treating the problem drinker: Modern approaches. In W. R. Miller (Ed.), *The addictive behaviors: Treatment of alcoholism, drug abuse, smoking, and obesity.* New York: Pergamon Press, 1980.

Moos, R., & Bliss, F. Difficulty of follow-up and outcome of alcoholism treatment. *Journal of Studies on Alcohol,* 1978, *39,* 473–490.

Moos, R. H., Cronkite, R. C., & Finney, J. W. A conceptual framework for alcoholism treatment evaluation. In E. M. Pattison & E. Kaufman (Eds.), *Encyclopedic handbook of alcoholism.* New York: Gardner Press, 1982.

Nathan, P. E., & Lansky, D. Common methodological problems in research on the addictions. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 713–726.

Neubuerger, O. W., Hasha, N., Matarazzo, J. D., Schmitz, R. E., & Pratt, H. H. Behavioral-chemical treatment of alcoholism: An outcome replication. *Journal of Studies on Alcohol,* 1981, *42,* 806–810.

Obitz, F. W. Control orientation and disulfiram. *Journal of Studies on Alcohol,* 1978, *39,* 1297–1298.

Obitz, F. W., Wood, J. D., & Cantergiani, N. Alcoholics' perceptions of group therapy and Alcoholics Anonymous. *British Journal of Addiction,* 1977, *72,* 321–324.

Orford, J., Oppenheimer, E., & Edwards, G. Abstinence or control: The outcome for excessive drinkers two years after consultation. *Behaviour Research and Therapy,* 1976, *14,* 409–418.

Orne, M. T. Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press, 1969.

Parente, R. J. (Ed.). *Instruction manual for the administration of the Addiction Severity Index* (3rd ed.). Philadelphia: Veterans Administration Press, 1980.

Pattison, E. M. A critique of alcoholism treatment concepts: With special reference to abstinence. *Quarterly Journal of Studies on Alcohol,* 1966, *27,* 49–71.

Pattison, E. M. (Ed.). *Selection of treatment for alcoholics.* New Brunswick, N.J.: Rutgers Center of Alcohol Studies, 1982.

Phillips, M., & McAloon, M. H. A sweat-patch test for alcohol consumption: Evaluation in continuous and episodic drinkers. *Alcoholism: Clinical and Experimental Research,* 1980, *4,* 391–395.

Polich, J. M., Armor, D. J., & Braiker, H. B. *The course of alcoholism: Four years after treatment* (R-2433-NIAAA). Santa Monica, Calif.: The Rand Corporation, 1980. (a)

Polich, J. M., Armor, D. J., & Braiker, H. B. Patterns of alcoholism over four years. *Journal of Studies on Alcohol,* 1980, *41,* 397–416. (b)

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press, 1969.

Ruggels, W. L., Armor, D. J., Polich, J. M., Mothershead, A., & Stephen, M. *A follow-up study of clients at selected alcoholism treatment centers funded by NIAAA: Final report.* Menlo Park, Calif.: Stanford Research Institute, 1975.

Schuckit, M. A. & Cahalan, D. Evaluation of alcoholism treatment programs. In W. J. Filstead, J. J. Rossi, & M. Keller (Eds.), *Alcohol and alcohol problems: New thinking and new directions.* Cambridge, Mass.: Ballinger, 1976.

Shyrock, H. S., Siegel, J. S., & Associates. *Studies in population: The methods and materials of demography.* New York: Academic Press, 1976.

Smart, R. G., Storm, T., Baker, E. F. W., & Solursh, L. A controlled study of lysergide in the treatment of alcoholism: I. The effects on drinking behavior. *Quarterly Journal of Studies on Alcohol,* 1966, *27,* 469–482.

Sobell, L. C. Critique of alcoholism treatment evaluation. In G. A. Marlatt & P. E. Nathan (Eds.), *Behavioral approaches to alcoholism* (NIAAA-RUCAS Alcoholism Treatment Series No. 2). New Brunswick, N.J.: Rutgers Center of Alcohol Studies, 1978.

Sobell, L. C., Cellucci, T., Nirenberg, T. D., & Sobell, M. B. Do quantity-frequency data underestimate drinking-related health risks? *American Journal of Public Health,* 1982, *72,* 823–828.

Sobell, L. C., & Sobell, M. B. Convergent validity: An approach to increasing confidence in treatment outcome conclusions with alcohol and drug abusers. In L. C. Sobell, M. B. Sobell, & E. Ward (Eds.), *Evaluating alcohol and drug abuse treatment effectiveness: Recent advances.* New York: Pergamon Press, 1980.

Sobell, L. C., & Sobell, M. B. Alcoholism treatment outcome evaluation methodology. In *Prevention, intervention, and treatment: Concerns and models* (Alcohol and Health Monograph No. 3, DHHS Publication No. [ADM] 82-1192). Washington, D.C.: U.S. Government Printing Office, 1982.

Sobell, M. B., Maisto, S. A., Sobell, L. C., Cooper, A. M., Cooper, T., & Sanders, B. Developing a prototype for evaluating alcohol treatment effectiveness. In L. C. Sobell, M. B. Sobell, & E. Ward (Eds.), *Evaluating alcohol and drug abuse treatment effectiveness: Recent advances.* New York: Pergamon Press, 1980.

Sobell, M. B., & Sobell, L. C. Second year treatment outcome of alcoholics treated by individualized behavior therapy: Results. *Behaviour Research and Therapy,* 1976, *14,* 195–215.

Sobell, M. B., Sobell, L. C., & VanderSpek, R. Relationships among clinical judgment, self-report, and breath-analysis measures of intoxication in alcoholics. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 204–206.

Solomon, S. D. *Tailoring alcoholism therapy to client needs* (DHHS Publication No. [ADM] 81-1129). Washington, D.C.: U.S. Government Printing Office, 1981.

Stojiljković, S. Conditioned aversion treatment of alcoholics. *Quarterly Journal of Studies on Alcohol,* 1969, *30,* 900–904.

Tuchfeld, B. S., & Marcus, S. H. Methodological issues in evaluating alcoholism treatment effectiveness. *Advances in Alcoholism,* 1982, *2*(Whole No. 23).

Tyndel, M., Fraser, J. G., & Hartleib, C. J. Metronidazole as an adjuvant in the treatment of alcoholism. *British Journal of Addiction,* 1969, *64,* 57–61.

Voegtlin, W. L., & Lemere, F. L. The treatment of alcohol addiction: A review of the literature. *Quarterly Journal of Studies on Alcohol,* 1942, *2,* 717–803.

Wanberg, K. W. *Differential assessment and treatment manual.* Denver: Colorado Department of Health, Alcohol and Drug Abuse Division, April 1983.

21

The Role of External Validity in Theoretical Research

John G. Lynch, Jr.

Calder, Phillips, and Tybout (1982) have criticized my analysis of the role of external validity in theoretical consumer research (Lynch 1982). In particular, they disputed my argument that theoretical researchers must concern themselves with the generalizability of their research findings. The issues raised by the debate are important because of the impact that the advice of Calder et al. could have, if accepted, upon theory development within our discipline. This paper will elaborate ways in which we can use evidence bearing on the external validity of theoretically predicted effects in an opportunistic fashion to hasten theoretical progress.

In an earlier paper, Calder et al. (1981) argued from the standpoint of a falsificationist philosophy of science (Popper 1959) and quite correctly noted that theories can be rejected if their predictions can be shown to be false for *any* subjects, settings, and events within their domain. Research practices often thought to enhance external validity—e.g., the use of heterogeneous "representative" samples of respondents and of uncontrolled field settings—were deemed unnecessary and even undesirable in theoretical research, because they inflate "error" variance and make it more difficult to detect systematic violations of the predictions of one's theory. Calder et al. concluded that because these research methods compromise the effort to provide the most rigorous possible test of one's theory (by reducing statistical power), external validity is of minimal relevance to theoretical consumer research, at least as it pertains to individual experiments.

My criticism of that thesis was that Calder et al. failed to distinguish the research methods that are commonly thought to increase external validity from the concept of external validity—i.e., the degree to which the effects of experimental manipulations are independent of the levels of supposedly irrelevant background factors (Campbell and Stanley 1966; Cook and Campbell 1979). This is not a mere semantic distinction. One of my major points was that these research practices (the use of heterogeneous, representative samples of respondents and of realistic, uncontrolled field settings) do *not* have the efficacy in increasing external validity that is commonly ascribed to them. If background factor × treatment interactions exist of which the researcher is unaware (as seems likely), these research practices can mask a substantial lack of external validity.

External validity is highly relevant to theoretical consumer research. Evidence demonstrating that theoretically predicted effects lack external validity—in that they fail to generalize across various levels of background factors presumed to be theoretically irrelevant—would indicate that the theory lacked construct validity. Certain research methods that allow a partial assessment of external validity should thus be routinely adopted. Unlike the methods for increasing external validity that Calder et al. criticized, these methods compromise neither statistical conclusion validity nor construct validity. These methods are Cook and Campbell's (1979) "model of deliberate sampling for heterogeneity" and the "selective approach" of choosing some small number of background factors to be varied orthogonally to the treatments (Lynch 1982). Both of these methods ensure a high level of within-block homogeneity—and hence, statistical power—while allowing some (incomplete) assessment of external validity. The inevitably imperfect nature of this assessment stems from the fact that the researcher cannot hope to anticipate and block upon all

From John G. Lynch, Jr., "The Role of External Validity in Theoretical Research," *Journal of Consumer Research*, 1983, 10(1), 109-111.

"background variables" that would actually interact with the treatments.

Calder et al. (1982) apparently did not appreciate the distinctions I made among (1) the concept of external validity; (2) research methods that are often erroneously believed to guarantee—or at least, to greatly enhance—external validity at some sacrifice to statistical conclusion validity; and (3) research methods that allow an imperfect assessment of external validity at virtually no cost in terms of statistical power. I welcome the opportunity to clarify these distinctions because the assumption of strong positive links between research methods of type 2 and the concept of external validity seems to be ingrained among researchers. Indeed, these links are implicitly accepted in a passage from Cook and Campbell's (1979, p. 83) classic treatment of validity issues quoted by Calder et al. (1982, p. 240).

Calder and associates use the same passage from Cook and Campbell (1979) to buttress their argument that external validity is irrelevant to theoretical research. Cook and Campbell suggest that in theoretical research, external validity (which they associate with research methods of type 2) is less important than are internal, construct, and statistical conclusion validities. But issues of priorities among validity types become relevant only when one is forced to sacrifice some degree of one validity to gain another. The methods I suggested (the "model of deliberate sampling for heterogeneity" and "the selective approach") yield increased information about external validity and hence about the (construct) validity of one's postulated nomological network, at virtually no cost in terms of other types of validity. Thus when Calder et al. (1982) maintain that theoretical consumer researchers should employ the "classic" approach of holding all "background" factors constant, they are in the difficult position of arguing for a dominated research strategy.

The "model of deliberate sampling for heterogeneity" is an attractive research strategy when the researcher believes a whole host of background factors to be theoretically irrelevant. S/he wants to provide some inductive (and therefore fallible) support for a claim that theoretically predicted effects are somewhat robust (Brinberg and McGrath 1983), and to reduce fears that these effects are paradigm-bound "epiphenomena." If theoretical predictions are shown to hold despite differences in the levels at which a great number of background factors are held constant, one's confidence in the generality of the theory's predictions is increased—although subsequent research findings could demonstrate a lack of generality of theoretically predicted effects.

The "selective approach" is more useful when the researcher is explicitly searching for "boundary conditions" on theoretically predicted effects (Brinberg and McGrath 1983). S/he is attempting to uncover conditions under which these effects will fail to replicate, in hopes that these failures will suggest ways in which the theory could be improved, modified, or more clearly circumscribed in scope. For this approach to be effective, there must be some means of prioritizing "background factors." To illustrate, one might have a well-developed theory that deems certain "background factors" to be irrelevant to the relationships among "theoretical variables"—yet findings in tangentially related literatures, casual empirical observations, or the researcher's hunches might suggest that the empirical relations among those theoretical variables would in fact depend upon the levels at which one or more background variables were held constant. While the researcher has no formal theoretical (i.e., explanatory) grounds for predicting such background factor × treatment interactions, s/he would not be surprised to discover them. Here the "selective approach" would be preferable to "deliberate sampling for heterogeneity" because in the former method, block × treatment interactions can be interpreted more easily should they emerge.[1]

EFFECTS ON THEORETICAL PROGRESS

My views and Calder et al.'s diverge primarily on the interrelated issues of the relationship of external validity to construct validity,[2] the role of tests of external validity in theory development, and the attitude that researchers should have toward so-called "background variables." Calder et al. consider background variables to be unworthy of the effort necessary to investigate their effects. Their reasoning is that the list of background variables that might interact with one's theoretical variables is endless, and that the very

[1]Calder et al. (1983) indicate that if one hypothesizes a background factor × treatment interaction (however tentatively), the background factor becomes part of one's theory. This is a reasonable view, although I would prefer to reserve the term "theoretical variables" for explanatory constructs embedded in portions of a nomological net in which one has some confidence. I agree that some background factors ultimately may be represented as part of one's theory proper.

[2]Calder et al. (1982) take issue with my claim that if theoretically predicted effects can be shown to lack external validity, this would be evidence that one's theory lacks construct validity. Their dispute stems from the fact that they use the term "construct validity" to refer to what Campbell (1960) has called "trait validity," whereas I referred to that aspect of construct validity which Campbell has called "nomological validity." Calder et al. cite several papers (e.g., Bentler and Speckart 1981; Phillips 1982) that use structural equations approaches to test the validity of some posited nomological network across multiple groups or situations. In all of these studies, each construct was measured in several different ways. All reported a high degree of convergence among different measures of the same hypothesized construct within a given group or situation. However, different relationships among the various constructs were demonstrated across the different contexts. Calder et al. represent these studies as leading to the conclusion "that one can achieve construct validity but not external validity in the context of a single study" (1982, p. 242).

Clearly, Calder et al. refer to construct validity in the sense of the trait validity of one's "measurement model" rather than of the nomological validity of one's "structural model." It is possible that despite a demonstrated lack of generalizability of theoretical relationships across contexts, multiple measures of an individual construct taken within each context might all load on a common factor. In such a case, the meaning of that common factor would be in doubt. As Cronbach and Meehl (1955) argue, the meaning of a theoretical term derives from its relationships with other terms in the nomological net. At any rate, it is *not* possible to claim support for construct validity in the nomological sense in the face of a demonstrated failure of external validity of theoretically predicted relationships. Such a result would require revision of one's theory (unless some artifact could explain the failure of theoretical predictions). This was the issue with which I was concerned.

label "background" admits that one has no firm theoretical basis for predicting which of these variables would in fact interact with the treatments.

I see these background variables as offering both a challenge and an opportunity for theory enrichment. First, each block within the "selective approach" and especially the "model of deliberate sampling for heterogeneity" can be considered to be an independent "method" by which the predictions of one's theory can be tested. It is a basic principle of construct validation (Cronbach and Meehl 1955) that one's confidence in an hypothesized nomological network increases with the number and independence of its predictions that are confirmed. Increased confidence is justified when theoretically predicted treatment effects are found *and* when block × treatment interactions are shown to be insignificant despite differences between blocks in the levels at which theoretically irrelevant background factors are held constant.

Second, when block × treatment interactions are unexpectedly shown to be significant, one is afforded the opportunity for inductive insight. When variables that theoretically should have no bearing on the operation of one's treatment variables are shown to make a difference, one is forced to revise one's theory to explain the unexpected data pattern. While not all such interactions will immediately suggest the proper form these revisions should take, such data patterns should be considered valuable grist to the theorist's mill.

More importantly, considering external validity to be unimportant can prevent researchers from thinking critically about how exogenous variables might alter the effects of theoretical variables. Such an insular attitude can lead to the acceptance of paradigmatic conventions about the levels at which certain crucial background factors are to be held constant. There are many examples of how such conventions have inhibited theoretical progress. For instance, a generation of memory researchers decided to hold the meaningfulness of stimulus materials constant at very low levels, as by the use of nonsense syllables. In retrospect, this created barriers to understanding the role of existing knowledge structures in the recall of presented information. A great many similar examples could be cited.

CONCLUSION

Despite the fact that one can never guarantee external validity, theoretical researchers should concern themselves with variables external to their theories and use procedures such as the "selective approach" and the "model of deliberate sampling for heterogeneity." This is no more a "counsel of despair" than is an exhortation to engage in theoretical research despite the logical impossibility of ever proving a theory. Indeed, the inability to guarantee external validity and the inability to prove theories both stem from exactly the same cause—the problem of induction. Despite the impossibility of ultimate success, the improvement of both theory and external validity remains a worthy goal.

[*Received January 1983. Revised March 1983.*]

REFERENCES

Bentler, Peter M. and George Speckart (1981), "Attitudes 'Cause' Behavior: A Structural Equation Analysis," *Journal of Personality and Social Psychology*, 40 (February), 226–238.

Brinberg, David L. and Joseph E. McGrath (1983), "A Validity Network Schema," paper presented at a convention of the American Educational Researchers Association, April 1983.

Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout (1981), "Designing Research for Application," *Journal of Consumer Research*, 8 (September), 197–207.

———, Lynn W. Phillips, and Alice M. Tybout (1982), "The Concept of External Validity," *Journal of Consumer Research*, 9 (December), 240–244.

———, Lynn W. Phillips, and Alice M. Tybout (1983), "Beyond External Validity," *Journal of Consumer Research*, 10 (June), 112–114.

Campbell, Donald T. (1960), "Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity," *American Psychologist*, 15, 546–553.

——— and Julian C. Stanley (1966), *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally.

Cook, Thomas D. and Donald T. Campbell (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Chicago: Rand McNally.

Cronbach, Lee J. and Paul E. Meehl (1955), "Construct Validity in Psychological Tests," *Psychological Bulletin*, 52 (July), 281–302.

Lynch, John G. Jr. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (December), 225–239.

Phillips, Lynn W. (1982), "Explaining Control Losses in Corporate Marketing Channels: An Organization Analysis," *Journal of Marketing Research*, 19 (November), 525–549.

Popper, Karl R. (1959), *The Logic of Scientific Discovery*, New York: Harper Torchbooks.

22

Beyond External Validity

Bobby J. Calder, Lynn W. Phillips, and Alice M. Tybout

As it is usually phrased, the question of external validity has to do with whether the results of a behavioral study would hold for other persons, settings, times, or places. Consistent with the original notions of Campbell and Stanley (1966) and Cook and Campbell (1977), we have argued that this concept of external validity is relatively less important than other forms of validity when the objective of research is to test theory (Calder, Phillips, and Tybout 1981, 1982). Our position is that external validity is a matter of the applicability of behavioral research. It arises primarily through severe and rigorous tests of theory rather than by attempts to incorporate "real world' variables into individual studies designed to test theory. Such variables only become important in the context of evaluating interventions based on theory.

Lynch (1982, 1983) disputes our view of the importance of external validity. He proposes an alternative approach that would put external validity on a par with construct and other validity issues in theory tests. We believe that Lynch's proposal does not adequately represent the issues we have raised. Nor does it provide an effective approach to improving the applicability of behavioral research. The fundamental flaw in Lynch's prescription for external validity is that it rests on a foundation of induction, and is similar in spirit to what we have termed effects research. In essence, it is a more sophisticated version of the notion that theoretical research must somehow include atheoretical variables that seem intuitively important.

DEVELOPING APPLICABLE THEORIES

We contend that applicable theories are a natural outgrowth of scientific progress. In accord with a falsificationist view of theory testing, theories are subjected to tests where they may fail. Falsification leads to conjecture that replaces an unsuccessful theory with a new one that accounts for the available data. Then the new theory is tested until its limitations are uncovered, and so on. In part, failures identify missing variables, which can then be incorporated into new theory. This is the very sort of process described by Lynch in discussing progress in information processing research.

It is an oversimplification to claim that we believe that "external validity is of minimal relevance to theoretical consumer research" (Lynch 1983, p. 109). It has always been our contention that theoretical research is concerned with applicability to specific persons, settings, times, and places—i.e., with what could be called external validity—but we believe that such applicability is best achieved through scientific progress and evaluation of interventions rather than through the design of any theory test.

Lynch disagrees with our view and proposes that researchers attempt to increase the external validity of their theory in the context of individual studies by blocking on a limited number of "background factors." The rationale is that some of these background factors may interact with the theoretical constructs; if this occurs, the theory lacks external validity. If no such interactions are found, Lynch argues that confidence in the theory is increased and that it has been shown to have some degree of external validity.[1] This suggestion does not have the methodological problems of more traditional efforts to enhance external validity (i.e., suggestions that heterogeneity in subjects and setting be employed), and it may serendipitously lead to revision in the theory. However, it has a critical limitation that Lynch has failed to appreciate—namely, that the conclusion of his argument must rest on induction.

Consider two competing theories, each tested in a setting in which several background factors are blocked. The predictions of theory A are obtained across all blocks, while the predictions of theory B are not (i.e., there are interactions of theory variables and background factors for theory B). On the basis of these findings, it would be inappropriate to claim that theory A has greater external validity than theory B. Background factors not yet measured may interact with theory A; in fact, the total number of background factors that interact with A could be greater than the number

[1]We are not disputing Lynch's recommendations simply on procedural grounds. There is nothing wrong per se with blocking on background variables. What we do argue against is the contention that examination of such variables is a basis for inferences about validity.

From Bobby J. Calder, Lynn W. Phillips, and Alice M. Tybout, "Beyond External Validity," *Journal of Consumer Research*, 1983, 10(1), 112-114.

that interact with B. Yet Lynch is forced to argue for precisely this kind of conclusion:

> If theoretical predictions are shown to hold despite differences in the levels at which a great number of background factors are held constant, one's confidence in the generality of the theory's predictions is increased—although subsequent research findings could demonstrate a lack of generality of theoretically predicted effects (1983, p. 110).

This is an inductive argument and one that has no basis in logic.

One might reason that the problem of induction could be escaped if one could distinguish important background variables from unimportant ones. If "important" background variables fail to interact with theoretical constructs, presumably the argument that confidence in the external validity of the theory is increased could be based on something other than pure induction. But how does one know a priori which background factors are important? Lynch offers little guidance in background factor selection. He merely suggests the use of deliberate sampling for heterogeneity, empirical observation, or researcher's intuition to identify important factors.[2] Although these recommendations relabel the problem, they do little to answer the question: on what dimensions should one sample for heterogeneity?

It is our basic argument that the theory under examination is the only efficient and logical basis for selecting independent variables a priori. If the theory guides selection, then the variables examined are not background factors but are properly viewed as part of the nomological network being examined. This is not a mere semantic issue. It represents an important distinction. Theory must be the driving force in designing theory-testing research; background factors should not be selected for inclusion in a haphazard fashion. This is not to say that we recommend ignoring background factors or are "apathetic" toward them. On the contrary, we contend merely that the impact of background variables cannot, by definition, be anticipated in theory-testing research.[3] Thus any attempt to include background factors by blocking is necessarily ad hoc, leaving research open to all manner of extra theoretical, intuitive biases that can only detract from the pursuit of theory.

In sum, we disagree with Lynch's contention that we have failed to distinguish the concept of external validity from its measurement procedure. Indeed, pursuit of the concept in the form of achieving applicable theory is the very essence of our view. However, given the confusion that appears to surround the notion of external validity—and the fact that in theoretical research such validity is actually achieved by refinements in the understanding of theoretical constructs—we would like to propose dropping the term altogether. What is needed more than new approaches to external validity is a sophisticated view of the nature of progress in theoretical research. Also, there is a greater need to consider evaluation studies designed especially for the application of theory-based interventions in specific situations.

APPLYING EXISTING THEORY

At any point in time, a theory will only have survived previous attempts at falsification. Theories are necessarily incomplete and unproven. This leads to the concern, shared by Lynch, that theory-based predictions will not obtain in natural settings where background factors are uncontrolled. It is always possible that nontheoretical variables may swamp predicted effects or that unanticipated interactions may occur. The issue is one of deciding whether *interventions* based on the application of an incomplete theory are worthwhile. Note the distinction between testing an intervention and testing the theory on which it is based. This issue is addressed by our contention that theory application must be a two-step procedure (see Calder et al. 1981 for a detailed exposition of this process). First, a theory must survive rigorous testing using controlled procedures; then it can be employed to design an intervention for a specific real world situation (e.g., information processing theory

[2]The specific approach recommended by Lynch varies as a function of whether the goal is to test the robustness or the boundaries of the theory. Yet neither goal is accompanied by clear specifications as to how background variables are to be chosen in the absence of theoretical guidance.

[3]Lynch's (1983) position is also based partly on the view that one can infer validity of a theory's constructs solely by an examination of what he now terms nomological validity—the degree to which predictions from the theory containing the constructs under scrutiny are confirmed. In Lynch's framework, nomological validity is ascertained by examining whether the patterns of association among empirical measures of a concept correspond to those predicted by theory. (This must be the case because Lynch's construal of construct validity does not seem to distinguish between a validity concept and its measurement.)

There are numerous problems associated with any attempt to ascertain the validity of a theory's construct measures and hypotheses by focusing solely on associations among empirical measures (see Calder, Phillips, and Tybout 1982, p. 242). Thus it is inappropriate to define construct validity solely in terms of whether constructs from a nomological network under investigation are confirmed. Rather, one must *first* establish convergent and discriminant validity of one's measurement to increase one's confidence that the measures are faithful indicators of the concept they represent. Confidence is increased to the extent that convergent and discriminant validity are achieved with maximally dissimilar measures. Only when convergent and discriminant validity are achieved is it meaningful to examine nomological validity criteria as a *further* basis for examining the validity of measures, concepts, and hypotheses.

The strongest support for a theory will emerge when convergent, discriminant, and nomological validity are achieved. Nevertheless, there may be instances in which concepts are faithfully measured (i.e., convergent and discriminant validity are achieved with maximally dissimilar methods, and yet the predictions expected by a theory are not confirmed. Such situations contradict Lynch's proposition that "if data supporting one's theory lack external validity, the theory lacks construct validity."

Construct and discriminant validity are necessary conditions for examination of nomological validity. Examination of nomological validity criteria alone is not meaningful because problems of measurement may result in good prediction for the wrong reasons. Thus Lynch's recommendations for how to infer the validity of a construct are incomplete and potentially misleading.

may be the foundation for developing an advertising strategy for use with a particular population). Often the intervention so designed is implemented on the presumption that it will operate in accord with the theory. We argue that such implementation is premature. The intervention must undergo its own testing to see whether it performs as anticipated in the setting of interest.[4]

Thus we advocate examining the impact of background factors, though not in the manner specified by Lynch. Instead of blocking on a small number of factors, we recommend representing the full range of background factors that the intervention will encounter in an uncontrolled fashion. It should be noted that the emphasis in an intervention test is on the *particular levels of background factors that will be present in a setting of practical interest;* the goal is not to represent *all* levels of all background factors.

If the intervention works as planned, confidence in its application is increased. If it fails, a new intervention must be designed and tested. We suggest that theories that repeatedly fail to lead to successful interventions are suspect, and that the circumstances of the intervention tests may be examined in an effort to develop hypotheses about background factors which require incorporation into the theory. Note, however, that this outcome is serendipitous. Moreover, such research will not ordinarily be a strong test of theory: once such factors are identified and subjected to theory-testing procedures, they are no longer background factors but are part of the theory itself. In fact, they can only be tested by being brought into the theoretical explanation to generate a prediction. Thus background factors and serendipity can play a role in theory testing, but not as a part of ongoing theory-testing efforts in the way suggested by Lynch.

CONCLUSION

Although we agree with Lynch about the goal of developing increasingly complete theories that can serve as a basis for explaining real world phenomena, we have distinct proposals for pursuing applicability. In spirit, Lynch appears to share the falsificationist view, but his recommendations for examining background factors as a basis for determining applicability run afoul of this perspective. Ironically, Lynch himself notes that "the inability to guarantee external validity and the inability to prove theories both stem from exactly the same cause—the problem of induction" (1983, p. 111). Because we find Lynch's perspective logically inconsistent with the falsificationist view, we cannot endorse it. We continue to hold that applicability or "external validity" should not be the objective of individual theory tests; rather, it must evolve from scientific progress and address the problems of applying incomplete theories. Two separate types of research are required: theory testing and intervention testing. Background factors should only be considered in the intervention testing stage, where the specific factors tested and the range allowed are dictated by the situation of interest rather than by ad hoc considerations. We believe that the two-stage procedure we outline is the most efficacious one for achieving the goal of applicability to specific persons, settings, times, and places. Moreover, our view has the advantage of being logically consistent with a falsificationist perspective.

[*Received March 1983.*]

[4]Our suggestion is clearly related to work in the evaluation research literature (see Phillips and Calder 1979, 1980). Note, however, that we are specifically addressing theory-based interventions, as opposed to those based mostly on practical considerations.

REFERENCES

Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout (1981), "Designing Research for Application," *Journal of Consumer Research*, 8 (September), 197–207.

———, Lynn W. Phillips, and Alice M. Tybout (1982), "The Concept of External Validity," *Journal of Consumer Research*, 9 (December), 240–244.

Campbell, Donald T. and Julius C. Stanley (1966), *Experimental and Quasiexperimental Designs*, Chicago: Rand McNally.

Cook, Thomas D. and Donald T. Campbell (1976), "The Design and Conduct of Quasiexperiments and True Experiments in Field Settings," in *Handbook of Industrial and Organizational Psychology*, ed. M. Dunnete, Skokie, IL: Rand McNally.

Lynch, John G. Jr. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (December), 225–239.

——— (1983), "The Role of External Validity in Theoretical Research" *Journal of Consumer Research*, 10 (June), 109–111.

Phillips, Lynn W. and Bobby J. Calder (1979), "Evaluating Consumer Protection Laws: I. Problem Methods," *Journal of Consumer Affairs*, 13(2), 157–185.

——— and Bobby J Calder (1980), "Evaluating Consumer Protection Laws: II. Promising Methods," *Journal of Consumer Affairs*, 14(1), 9–36.

23

External Validity and the Research Process

A Comment on the Calder/Lynch Dialogue

Joseph E. McGrath and David Brinberg

A series of five articles about external validity and generalizability has appeared in the *Journal of Consumer Research* in recent years, constituting a dialogue between Calder, Phillips, and Tybout (1981, 1982, 1983) on the one hand, and Lynch (1982, 1983) on the other. That dialogue has attracted much attention in the consumer research field. The two sets of authors have raised many important issues. They apparently disagree sharply on some of them. They are to be congratulated, we feel, both for raising the consciousness of the field on these important matters and for keeping their dialogue from deteriorating into rancorous, ad hominem "cheap shots."[1]

While Calder and Lynch apparently disagree on a number of important issues regarding external validity and the research process, they also apparently agree on a great deal. We are struck by the range of basic issues on which they agree, and on which we agree with them. At the same time, we are struck by how much the apparent disagreements between them become moot—or even become agreements—when viewed within our own, somewhat broader schema for analysis of external validity and the research process (see Brinberg and McGrath 1982).

Our response to the task at hand is a paper in three main parts. First, we discuss some of the quite extensive set of issues and themes upon which Calder et al. and Lynch agree, and in doing so we presage our own views on some of them. Second, we lay out our own schema for analysis of validity and the research process, with special attention to parts of it that bear on aspects of the Calder/Lynch discourse. Third, we try to show how many aspects of the Calder/Lynch dialogues fit, and how their apparent disagreements dissolve, when placed within our broader schema.

MAJOR POINTS OF AGREEMENT

It often happens in the course of a conversation—especially when the conversation veers toward an argument—that we fail to notice, much less to emphasize, the vast expanse of underlying agreement that must exist between the two conversants for them to carry on the conversation at all. So it is, we believe, with the Calder/Lynch discourse on external validity. We want to make explicit some features of that vast area of agreement, both to put their apparent disagreements into appropriate perspective and to provide a context for the later parts of this paper. One can easily identify several dozen points about the nature of validity and the research process on which Calder and Lynch are in apparently solid agreement. We note a number of such points here, organized around half a dozen major themes. We comment on some of them in a way that will anticipate our own interpretations of these matters.

Theme #1: *External validity plays an important and complex role in the research process.*

Calder and Lynch agree that generalizability and external validity are important matters, and that they have been given far less attention than their complexity and importance would merit. We agree. We also agree that the quite badly used term, external validity, needs to be either defined carefully into its diverse forms or dropped in favor of a set of terms that begin life with less connotative excess baggage. We concur with both Calder and Lynch that it is a serious mistake to equate external validity with realism. Moreover, all three parties agree that we must distinguish between concepts on the one hand, and relations between sets of concepts on the other. For example, we all agree that we need to make a distinction between (1) the construct validity of a concept, as reflected in the convergence (and discrimination) of some particular set of operationalizations of it, and (2) the construct validity of a relation between two concepts, as reflected in the "fit" of that relation within some nomological network. We also agree that it is the latter aspect of construct validity that is intricately linked to considerations of external validity.

Theme #2: *The scientific paradigm has inherent limitations.*

Calder and Lynch agree, and we concur, that induction is a tool with serious limitations; specifically, all parties

[1]These authors and the Journal's editors are to be congratulated, also, for asking a presumably nonpartisan third party to comment on the debate and thereby to end the dialogue (although not, of course, to end interest in or concern for the issues involved).

From Joseph E. McGrath and David Brinberg, "External Validity and the Research Process: A Comment on the Calder/Lynch Dialogue," *Journal of Consumer Research*, 1983, 10(1), 115-124.

agree that induction cannot lead to certainty. (We do not agree, however, with the characterization of induction as having "no basis in logic.") All three parties seem to agree, too, that *all empirical knowledge in science is both probabilistic* (since it is based on induction) *and contingent* (on the conceptual, methodological, and substantive conditions under which it was obtained). We regard this as a fundamental fact of life which permanently and pervasively limits the "perfectability" of scientific knowledge.

They also agree that Popper's falsification principle is at the heart of the process by which we advance both our scientific knowledge and our confidence in that knowledge. We agree, but we interpret the falsification principle as a two-edged sword, and will argue later that we can improve our state of knowledge (i.e., reduce uncertainty) both by *failing to disconfirm* and by *confidently disconfirming* our hypotheses (Popper 1959).

Calder and Lynch also seem to agree that any exploration of the robustness of a set of findings is likely to encounter boundaries or limits beyond which those findings do not replicate. We concur, but we view the dual thrust of robustness and boundaries as much more fundamental—indeed, we regard it as a reflection of the dual nature of the falsification principle itself. We hold that in the pursuit of external validity, the research community is obliged to seek not only the scope but also the limits of its findings.

Theme #3: *All specific methods have inherent limitations.*

All three parties agree that not only does the overall scientific paradigm contain serious inherent limits, but so do all specific instruments, methods, procedures, designs, and strategies within that paradigm. There is no one best set of methodological choices. Even when a particular research study is aimed at one specific and limited goal (see Theme #5), there is no "one best way" to carry it out. We would carry this point further, and hold it to be a basic characteristic of the research process: all methods are flawed, but different methods are flawed differently. We *must* use multiple methods in all aspects of the process (e.g., strategies, designs, and measures), not only to provide the triangulation basis for convergence (as in the classical multitrait-multimethod approach of Campbell and Fiske 1959), but also to let differently flawed methods shore up each others' vulnerabilities (see Webb et al. 1966; McGrath, Martin, and Kulka 1982).

Furthermore, we all agree that it is unwise to try to compromise by mixing two or more research strategies. For example, we all agree that trying to run rigorous experiments in field settings, or trying to add the trappings of "mundane realism" to experiments in laboratory settings, are by no means ways to improve the quality of our data. The potential advantages of all methods, including research strategies, are *only* potential, but the inherent weaknesses of all methods are inevitable. By mixing research strategies, as in the examples given above, we are likely to get the worst of both rather than the best of each.

Theme #4: *The researcher's understanding of both the theoretical and the substantive system is crucial to research progress.*

All three parties agree that theory and empirical research are closely interwoven. We all agree, too, that it is important for the theorist to incorporate all relevant factors within the theory. Some of the relevant factors (Lynch's "background factors") are not initially a part of the theory as formulated, but in fact interact "in nature" with key theoretical variables or relations. Efforts to identify which of myriad potentially relevant factors are indeed relevant—and important enough to incorporate into theory—are themselves in need of theoretical guidance. This is of necessity a "bootstrapping" operation in which, we all agree, the researcher's intuition—and understanding of the substantive area under study—is perhaps the most valuable single ingredient.

Theme #5: *Research is done for a variety of purposes and in a variety of ways.*

All three parties agree that different specific research studies are carried out for different purposes in pursuit of a variety of (limited) goals. Both Calder and Lynch seem to regard several of these as quite distinct bodies of research activities, motivated by different purposes and carried out by different sets of procedures. We regard them not so much as separate "kinds" of research but as separate "paths" or sequences of steps for carrying out certain portions of the same overall research process.

We all agree that these different bodies of research (or research paths) stem from different initial emphases, purposes, preferences, or values. The applied researcher and the basic researcher—corresponding more or less to the two main "paths" discussed by Calder and by Lynch (effects application and theory research)—simply have different ideas, priorities, and values about what is of most importance in research. We will elaborate these ideas, arguing that there is a third "type" (a methodologist or technician) and that there are a couple of possible paths—not just one—by which each might pursue their differing goals. In fact, the idea of multiple research paths is central to our validity network schema, although we regard them as especially important in relation to internal rather than to external validity.

Theme #6: *Sampling plays a crucial and complex role in external validity.*

All three parties agree that sampling issues are important in regard to external validity, as well as in regard to other aspects of the research process. Furthermore, we all agree that the proper focus of sampling plans should be (although in practice it seldom is) on events rather than on respondents. Here, events are regarded as behaviors of persons in time/place/situation contexts. We would carry the matter further and argue that external validity must be assessed not

only with regard to facets of events, but also with regard to facets of methods and concepts as well.

Moreover, all three parties agree that populations whose members cannot be enumerated (either because they are infinite in number or for other reasons) pose some difficult problems for sampling plans. (We would not necessarily all agree on the full implications of those difficulties for the research process.) All three parties also agree that, in any case, external validity or generalizability has to do with *a relation between past events* (already measured as part of a "finding") *and future events* (to which that "finding" is to be generalized). The population of future events is, by definition, not enumerable.

Furthermore, all three parties would agree that there are at least four major sampling strategies that might be adopted vis à vis any *one* aspect or facet of the events under study (e.g., age of the population of respondents):

1. Sampling homogeneously over the entire study (i.e., holding the facet constant, say at age 18–20)
2. Sampling several subsets, each homogeneous within subset on the facet but differing on it between subsets, so that all the subsets together span the whole range of the facet (e.g., subsets in the teens, the 20's, 30–50, and over 50)
3. Sampling heterogeneously, but in a way that yields an overall distribution of the facet among the cases within the study that reflects (is representative of) the distribution of the facet among cases "in nature" (e.g., the age distribution of the target population)
4. Sampling heterogeneously on the facet but without regard to representativeness.

These four strategies offer different opportunities for—and pose different threats to—the exploration of the external validity of any given set of findings with respect to the facet in question. We carry these concerns several steps further, by construing the search for external validity of a given set of findings as *the deliberate and systematic search, on a number of facets, for both the scope and the limits* over which that given set of findings does and does not hold.

Summary

Even though the various points of agreement under these six major themes do not constitute all of the points of agreement between Calder and Lynch, they do add up to an impressive credo of methodological fundamentals. We take that set of fundamentals as one of two starting points for examining the Calder/Lynch dialogues. Our other point of departure is our own previous work, which we call the "validity network schema" (Brinberg and McGrath 1982, 1983).

THE VALIDITY NETWORK SCHEMA

In a recent article (Brinberg and McGrath 1982) and paper (Brinberg and McGrath 1983), we have offered a framework for analysis of validity and the research process. Our schema builds upon many of the points mentioned in the preceding section (among others) to offer a systematic description of the research process and of the multiple forms of validity that need to be pursued within it. This framework offers a perspective from which, we believe, both the agreements and the apparent disagreements between Calder and Lynch can be seen as part of a broader context. We do not presume that the Brinberg and McGrath material is so widely known that most readers will already be familiar with it (as we do presume for the key materials by Campbell (1959, 1966), by Cronbach (1975, 1982), and by Calder and Lynch themselves). Thus we will present the key ideas from that schema here, as briefly as clarity will permit.

Overview

The validity network schema starts with three assumptions:

1. That research involves three interrelated but analytically distinct domains, the *conceptual*, the *methodological*, and the *substantive*
2. That research involves *elements*, and *relations* between elements, from each of those three domains
3. That the complete research process involves *three major stages*, some with several steps and alternative paths for fulfilling those steps, and that there is a different fundamental idea of validity within each of the three stages.

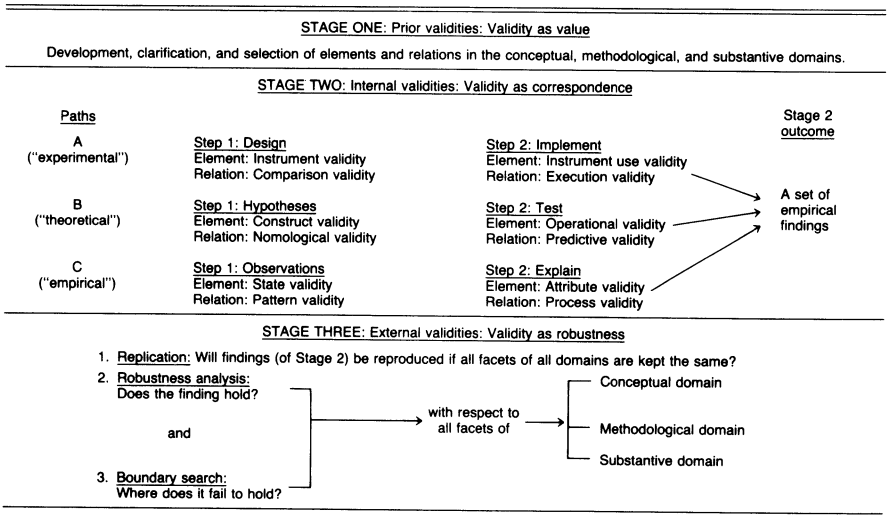
We have argued that all types of research involve the combination of some set of concepts, some set of methods for making observations or comparing sets of observations, and some set of substantive events and phenomena that are to be the focus of the study (Brinberg and McGrath 1983). Our validity network schema (1982, 1983) describes the research process as the *identification, selection, combination,* and *use* of elements and relations from the conceptual, methodological, and substantive domains:

1. The conceptual domain contains elements that are *concepts*, and relations between elements that are essentially *conceptual models* about patterns of concepts.
2. The methodological domain contains elements that are *methods*—or instruments or techniques—for making observations or manipulating variables, and relations that are structures or *comparison models* for comparing (i.e., for exploring covariation and difference in) sets of observations.
3. The substantive domain contains elements that are *events* (behaviors in temporal/spatial/situational contexts) and relations that are *phenomena* (patterns of relations among events).

Different problem areas and subareas deal with different portions of the substantive domain. Different research fields and subfields make use of sets of elements and relations from different portions of the conceptual and methodological domains. Yet any given research study makes use of

EXHIBIT 1

FORMS OF VALIDITY WITHIN THE RESEARCH PROCESS



some set of elements and relations from each of the three domains.

We have divided the complete research process into three main stages, each with several steps. Stage 1 involves development, clarification, and selection of elements and relations within each of the three domains. It is preparatory, necessary groundwork that must be done but that is often overlooked or not regarded as "research proper."

Stage 2 involves the combination and use of elements and relations from each of the three domains. It is this stage that we usually have in mind when we refer to "a research study." It involves two main steps and three different paths by which those two steps can be carried out (see Exhibit 1). All three paths lead to the same end product—a set of empirical findings.

Stage 3 involves following up the findings of Stage 2, by replication and by a systematic search for both the range and the limits of those findings. Stage 3 activity is intended to *verify, extend,* and *delimit* the set of findings that resulted from Stage 2.

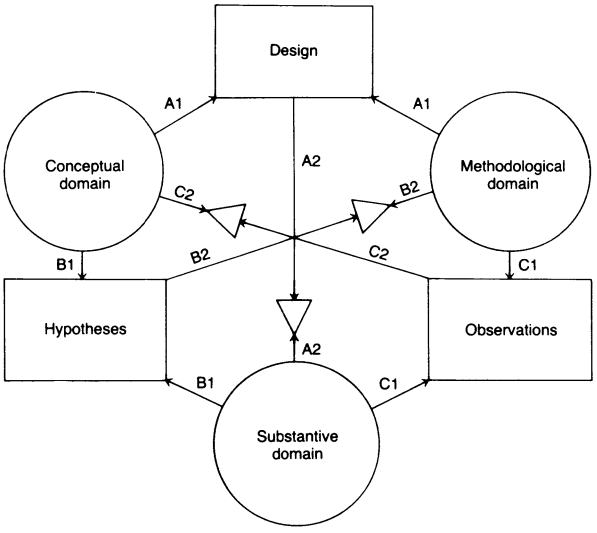
The three stages are related to one another in a kind of many-to-one-to-many relation. Stage 2 has reference to research on some specific set of problems, which we will here refer to as "the focal problem." The focal problem is defined by the substantive events, concepts, and methods that are chosen for Stage 2 activities. Stage 1 is preparatory for the study of *many* problems, of which the focal problem is but one. Stage 3 explores the range and limits of the focal problem findings of Stage 2, both when the same concepts, methods, and events are studied and when certain facets of those concepts, methods, or events are systematically varied.

Validity has different basic meanings within these three stages. In Stage 1, validity means *value*. Elements and relations are developed, clarified, and selected within each of the three domains if and only if the Stage 1 researchers consider them to be "of value"—i.e., important, interesting, or useful. Researchers engaged in Stage 2 activity—very often not the same persons who did the Stage 1 work—often accept without much thought what Stage 1 researchers have developed within one or more of the domains. That is, they borrow, often rather casually, from substantive system experts, conceptual system experts, or methodological system experts for use in relation to their own focal problem of Stage 2. Thus the values expressed in Stage 2 research often are not apparent even to the Stage 2 researcher, with respect to the elements and relations of one or more of the domains.

In Stage 2, validity has the meaning of *correspondence* or fit; that is, validity is construed as the extent to which

EXHIBIT 2

ALTERNATIVE PATHS FOR CONDUCT OF STAGE TWO OF THE RESEARCH PROCESS



Path A: (A1) Building a design, and (A2) implementing it by using it on a set of substantive events.
Path B: (B1) Building a set of hypotheses, and (B2) testing them by evaluating them with an appropriate set of methods.
Path C: (C1) Building a set of observations, and (C2) explaining them by construing them in terms of a set of meaningful concepts.

elements and relations from different domains fit when paired together. The first step in Stage 2 involves fit within the structure built by combining the elements and relations of two of the domains. The second step involves fit between that structure and elements and relations from the third domain. Since there are three ways to combine two sets of things, there are three different ways to take the first step—and hence three different "paths" through Stage 2 of the research process (see Exhibit 1). Considering both element and relation levels, and two steps within each of three paths, Stage 2 subsumes 12 distinct validity concepts—all involving the underlying notion of correspondence or fit.

In Stage 3, validity takes on the meaning of generalizability, *robustness*, or so-called external validity. Stage 3 involves activities that have to do with increasing our confidence concerning a Stage 2 finding. Specifically, it has to do with reducing our uncertainty about the range of variations—of events *and* concepts *and* methods—over which the Stage 2 findings do and do not hold. Robustness implies both the idea of *replication* (will the findings hold if I do the same thing again—as exactly as I can?) and the idea of *generalizability* (over what range of variation, on various facets of the elements and relations from *each* of the three domains, do the stage two findings hold?). The idea of generalizability, in turn, implies both the *scope* over which the findings do hold and the *limits* beyond which those findings do not hold, or do not hold in the same pattern. In this context, scope and limits refer to the range of variation on *every facet of each of the three domains*. (This is another formulation of the idea, discussed under Theme #2 in the preceding section, that all research information is contingent on the values of *all* variables—i.e., all facets of the events, concepts, and methods—under which that information was obtained.)

The complex set of relations among the three domains, the two levels (elements and relations) within each, the three stages of the research process, the three two-step alternative paths within Stage 2, and the two-step paths within Stage 3 are all reflected in Exhibit 2, which shows the relations of the domains to the two-step paths of Stage 2, and in Exhibit 1, which lists the forms of validity that arise at each level of each step of each path and in each stage.

Stage 1: *Validity as Value in the Preparatory Stage*

Research in Stage 1 involves identification, development, clarification, and selection of elements and relations from the conceptual, methodological, and substantive domains. The key idea of validity in this stage is value. The researchers' values guide (1) what aspects of the "real world" are regarded as worthwhile for study (i.e., what events and phenomena are attended to in the substantive domain); (2) what kinds of explanations are considered meaningful as interpretations of a set of observations (and hence what concepts and conceptual models are drawn from the conceptual domain); and (3) what methods for collection and analysis of data are regarded as acceptable for use in a scientific endeavor. Note, however, that different persons are liable to be "the researcher" with respect to Stage 1 developments in each of the three domains. The Stage 1 specialist in the methodological domain—the scientific tool maker, so to speak—often does the developmental work for methods and comparison models that are used by many others doing Stage 2 work on many problems. For example, tools such as Likert scales, semantic differentials, correlation and regression methods, and factor analysis all provide sets of methodological elements and relations that are drawn upon by wide segments of the behavioral science community. Similarly, there are Stage 1 specialists in the conceptual domain (one might call them theorists in some usages of that term, although we use the word "theory" somewhat differently). The Stage 1 specialists in the substantive domain are those who serve as system experts, and there are as many sets of system experts as there are "real world" systems in which Stage 2 researchers might be interested. The primary interest of a substantive system expert is usually in the elements and relations of just one specific substantive system.

The persons who carry out Stage 2 research activities are quite often not those who carry out Stage 1 activities; moreover, Stage 1 system experts tend to be different persons working within each of the three domains. Thus Stage 2 researchers are very likely to be selecting "preprogrammed" sets of elements and relations within one, two, or possibly even all three domains with which to carry on their activities. Which sets Stage 2 researchers will worry most about, and which sets they will adopt more casually, will reflect the values of those Stage 2 researchers. Yet the elements and relations that they adopt within one or more of the domains will also reflect the values of the Stage 1 specialists who did the developmental work in those domains—work that may be more or less transparent to the Stage 2 researcher. For example, whenever one uses multiple regression, factor analysis, or any other data analysis technique, the results are constrained by the sets of assumptions built into those techniques, whether or not the user realizes it or is even aware of what those techniques entail. (This is another realization of the idea, noted in Theme #2 of the preceding section, that all research information is contingent on the methods used to obtain it.) Similarly, when a Stage 2 researcher selects substantive events and phenomena for study by simply adopting a particular system expert's delineation of what is what (e.g., adopting management's specification of outcome units, communication paths, and so on), results of the subsequent research activities will be contingent on the assumptions underlying those specifications, whether or not the Stage 2 researcher intended it to be so. The same is true with regard to borrowing sets of concepts and relations—for example, assuming a linear relation between two concepts rather than a curvilinear one, or assuming that a certain pair of concepts are polar opposites rather than orthogonal dimensions. In all of these cases, the values of Stage 1 researchers in one or more domains are embedded within the Stage 2 research activities, perhaps unbeknownst to the Stage 2 researcher.

Stage 2: *Validity as Correspondence in the "Research Proper" Stage*

Stage 2 is the part of the research process that we most often regard as "doing a study." It involves the combination of elements and relations from all three domains. Logically, this is done in two main steps (Brinberg and McGrath 1982). In the first step, elements and relations from two of the domains are combined to form an intermediate or instrumental "structure." In the second step, elements and relations from the third domain are brought into combination with the structure developed in Step 1. Thus there are three possible two-step "paths" to carry out Stage 2, and three intermediate structures resulting from Step 1 of those paths. For path A (which we call the "experimental path"), Step 1 involves combining elements and relations from conceptual and methodological domains. This yields a Step 1 structure that we call a "design" (see Exhibit 1). Step 2 of that path involves implementing that design by combining elements and relations from the substantive domain with the design structure. For path B (which we call the "theoretical path"), Step 1 involves combining elements and relations from the conceptual and the substantive domains. This yields a Step 1 structure that we call a "set of hypotheses" (see Exhibit 1; in Brinberg and McGrath 1982, we called this a "theory"). Step 2 of that path involves testing that theory or set of hypotheses by combining elements and relations from the methodological domain with that set of hypotheses. For path C (which we call the "empirical path"), Step 1 involves combining elements and relations from the methodological and substantive domains. This yields a Step 1 structure that we call a "set of observations" (called a "set of data" in Brinberg and McGrath 1982). Step 2 of that path involves explaining that set of observations by combining elements and relations from the conceptual domain with that set of observations.

For Stage 2, the idea of validity has to do with correspondence or fit. In Step 1, the fit is between the elements and relations from two of the domains. In Step 2, the fit is between the elements and relations from the third domain

and the intermediate structure built in Step 1. The form of the validity question is different at the element and the relation level (as noted earlier under Theme #1 in regard to construct validity of measures and of relations). The form of the validity question is also different for Step 1 and Step 2 of each path. There are, therefore, 12 different validity questions—all including the idea of correspondence or fit—that arise from this 2 (level) × 3 (path) × 2 (step) formulation of Stage 2. The 12 validity concepts are shown in Exhibit 1. They embrace many familiar terms and concepts from the validity literature, as well as some additional forms of validity for which there is not now a generally accepted term.

The end result of Stage 2, regardless of which path is used to pursue it, is what we term "a set of empirical findings." That set of empirical findings will be different, in that it will have encountered and coped with different forms of the validity issues, depending on which path was followed. Thus the set of empirical findings will be an exemplification of an implemented design (path A), a tested theory (path B), or an explained set of observations (path C). Which of those it is will affect a number of aspects of the research activity and of our confidence in the results. For one thing, the path taken will probably reflect which domain got the most attention (one of the two embodied in Step 1) and, especially, which domain got the least attention (the domain not included in Step 1). If we imagine that researchers are likely to be more rigid about which elements and relations are to be included for the domain in which they are most interested, then the selection of elements and relations from the second domain, to construct the structure of Step 1, will be made with some adaptation to what is already fixed from the first domain. For example, if we were interested in testing certain concepts and chose the experimental path, we would probably try to find some methods that reflected the concepts in which we were interested, at both element and relation levels. But when we reached Step 2 and were trying to combine elements and relations from the third domain (which, we contend, is usually the one of least immediate interest), there would be a strong temptation to "adjust" what we selected from that domain to fit the quite limited "degrees of freedom" still remaining after building the Step 1 structure (which itself involved adaptation of domain two to accommodate interest in domain one). It is for this reason, we believe, that the researcher following the experimental path (path A) often seems to give short shrift to the substantive domain thus giving rise to the complaint that the study deals with artificial or trivial material. Similarly, the researcher following the empirical path (path C) often seems to neglect the conceptual domain and is thereby accused of "dust bowl empiricism," while the researcher following the theoretical path (path B) often seems to give short shrift to the methods domain and is accused of being cavalier about methods.

Values affect the research process in several ways, some of which have already been noted. First, in Stage 1, the values of the field (and of the surrounding culture) affect (1) the kinds of conceptual systems and methodological tools that are developed for potential use in behavioral science problems, and (2) the kinds of substantive systems that are identified as amenable to and worthy of behavioral science study. Those values become embedded in the conceptual, methodological, and substantive "systems" that are drawn upon by researchers pursuing Stage 2 of the research process. Second, since there are always more elements and relations available within each of the three domains than are used in any Stage 2 study, the values of the Stage 2 researcher influence which subsets are chosen for study. The Stage 2 researcher's values or preferences also affect the relative emphasis that is given to each of the three domains, and thus which aspects of various validity issues are actually addressed.

We can illustrate the latter points by referring back to our earlier discussion of the effects of the order of emphasis on the three domains. Suppose we posit one "type" of researcher, whom we will call an "applied researcher." Suppose further that the applied researcher starts with a dominant interest in certain elements and relations within the substantive domain. Note that there are two possible paths for conducting Stage 2: one is the path we call the "empirical path," for which Step 1 involves combining elements and relations from the substantive and methodological domains to yield a "set of observations" (as yet uninterpreted); the other path is what we call the "theoretical path," for which Step 1 involves combining elements and relations from the substantive and conceptual domains to yield a "set of hypotheses" (as yet untested). These are quite dramatically different paths, but both are amenable to the pursuit of applied goals. The "empirical path" seems to be the one that Calder and colleagues have in mind when they talk of "effects application."

Similarly, imagine a second type of researcher, termed a "basic researcher," who starts with a dominant interest in the conceptual domain. Again, there are two paths by which to pursue that interest. One is what we call the "experimental path," for which Step 1 involves combining elements and relations from the conceptual and methodological domains to yield a "design" (as yet unimplemented). This seems to be the path that Calder et al. have in mind when they talk about the "theory research" pattern. The other path available to the basic researcher is the "theoretical path" already noted under the discussion of the applied researcher's options. (It is interesting that neither of the two types discussed in the Calder and Lynch dialogues follows the path that we term the "theoretical path," although the "theory intervention" type discussed by Calder seems to correspond to Step 2 of our "theoretical path.")

Stage 3: *Validity as Robustness in the Followup Stage*

The business of Stage 3 of the research process is to verify, extend, and delimit the set of findings that result from Stage 2 activities. All of these are efforts to increase

our confidence in (reduce our uncertainty about) the Stage 2 findings. This involves three questions:

First: *If the study were repeated exactly, would the same findings occur?* This is the matter of replication. Replication implies that the Stage 3 researcher attempts to conduct a study that uses sets of concepts, methods, and substantive events that are (assumed to be) the same as those used in the Stage 2 activities that led to the findings that are to be replicated. Here the validity issue is reliability at the element level and statistical conclusion validity at the relations level. And, as in the assessment of test reliability, actually carrying out a replication requires that we relax the "sameness" constraint for at least one of the facets of concepts, methods, or events. For example, we might get a new sample of occurrences of events in question by obtaining observations on another set of days, or on another set of participants, or in another region. Yet if our intention is to replicate findings from an earlier study, we work on the premise that the "other" set—of occasions, persons, or locations—is for all practical purposes "the same" as the set used in the earlier study. This is always an assumption known to be false to some degree (as is also the case in all reliability assessments), and it is sometimes tempting to use this knowledge to "explain away" our not-infrequent failures to replicate.

Second: *If the Stage 2 study were done again, but with systematic variations on one or more facets of one or more of the domains, would the Stage 2 findings be robust over those variations?* And third: *Under what conditions—that is, for what variations in what facets in each of the domains—will the Stage 2 findings not hold?* The former question asks about the robustness or scope of the Stage 2 findings; the latter asks about their boundaries or limits. We always get partial answers to both questions at the same time, although researchers in our fields usually set out to demonstrate robustness of findings, not to search for their limits.

Note that while replication implies sameness on all facets of all domains, the dual questions of robustness and boundary search need to be asked with respect to many different facets of each of the three domains—i.e., concepts and methods, as well as substantive events. Almost all discussion of external validity has been limited either to replication or to robustness search with respect to the substantive domain only, and on only a few facets of that domain. Most attempts to assess external validity have focused on sampling with respect to respondents. As previously noted under Theme #6 and as elegantly pointed out by Lynch, the appropriate sampling unit for such "generalization" studies is the population of events, not the population of respondents or behaving units (usually individuals). To sample respondents ignores all facets of events that are carried not in the person, but in the situation, stimulus, or context.

The validity network schema elaborates the issue of external validity into a complex set of questions which entail not only replication, but also robustness and boundary search with respect to many facets of each of three domains. This is in accord with Calder's call for either abandoning the concept of external validity or differentiating it into a more useful set of conceptual distinctions (see Theme #1 in the preceding section). It also puts into a much more richly articulated context the many issues of sampling strategy discussed by Calder et al., Lynch, Cook, and Campbell (1979), and Ferber (1977), and in many other methodological works (see also Theme #6).

Finally, this elaboration of the idea of external validity—and in particular the insistence upon the dual questions of robustness and boundary search, of scope and limits—highlights the dual nature of Popper's falsification principle (see Theme #2). We gain knowledge both when we fail to disconfirm (some prior findings) and when we confidently disconfirm (those prior findings). Replication and robustness analyses rely on the former; boundary search relies on the latter.

Not only do researchers in our fields seldom deliberately set out to search for the boundaries of their findings, but when they do encounter such disconfirmations ("failures of invariance" in Wimsatt's (1981) terms), they generally regard them as negative findings to be explained away or otherwise treated as "non-findings" (see Themes #2 and #4). We would argue that there is just as much useful information in the identification of the limit of a finding (i.e., in confidently disconfirming a repeat of the finding) as there is in identifying a variation over which a given prior finding does in fact hold (i.e., a failure to disconfirm). The logic of pursuing both the scope and the limits of a given set of findings is parallel to the logic of pursuing both convergent and discriminant validity in assessing methods of measurement (Campbell and Fiske 1959). We cannot know what a concept is if we do not at the same time know what it is not. Thus we cannot know the scope of a set of findings unless we can establish the limits of those findings. And if a particular finding has no limits—if it holds for all conceivable subjects, behaviors, situations, and so on—then neither does it have any useful meaning. In other words, knowledge is always knowledge of differences (Runkel and McGrath 1972), and if a finding is unbounded, it cannot add to that knowledge.

Besides neglecting facets in the substantive domain other than respondents and avoiding all attempts to search for limits, past work on external validity has tended to neglect questions of robustness and boundary search with respect to the methodological and conceptual domains. There has been some recognition of the need for both robustness and boundary search with respect to methods of measurement, as reflected in the convergent/discriminant validity ideas already discussed. There also have been a few efforts to explore the extent to which a given finding is robust over analysis models, but that has often been incidental to attempts to shore up some limitations of an initial analysis.

The questions of robustness and boundary search are quite different when applied to the conceptual domain. Here the researcher's interest centers on two questions. On the one hand, we need to assess whether the concepts used to

interpret the substantive-methodological findings are sufficient to fully account for them (e.g., to take into account all nuances, to not imply differences where no differences are found). On the other hand, we need to assess whether these concepts are uniquely able to acccount for those findings (or could the findings be accounted for equally well—or better—by some other, perhaps more parsimonious, sets of concepts?). In other words, we need to know whether the set of concepts used to account for the findings in question is both a *necessary* and a *sufficient* set. Yet by and large, researchers in our fields seldom search the conceptual domain either for robustness or for boundaries.

SOME IMPLICATIONS OF THE VALIDITY NETWORK SCHEMA

The preceding section has foreshadowed a number of ways in which the Brinberg-McGrath validity network schema sheds light on key points at issue in the Calder/Lynch dialogues. If we accept the schema as is, along with the common themes discussed in the first section of this paper, then a number of points seemingly at issue between Calder and Lynch become either moot or resolved.

First, there is no such thing as increasing the external validity of a given study, A, within that study, A. The external validity of study A always has to be assessed in terms of results of some other study (B, C, D, and so forth). Study A may have implications for the external validity of some other study (B, or C, or K, or J), past or future, but not for itself. Furthermore, the question of whether some finding, X, obtained with sample i, would replicate if tried on some other sample, j, is merely the tip of the iceberg in regard to the overall external validity of X. This is the case regardless of the nature of i and j, and their relation to each other and to the so-called "real world."

Second, all research involves choices, and those choices reflect the values of the researcher. Often, they also reflect the underlying values, not just of the researcher conducting the study, but also of the researchers who have developed the currently dominant research paradigms and supporting norms of the field.

Third, there are indeed different "styles" or "paths" by which research is conducted. These reflect the researchers' purposes. These research styles should be regarded as alternative paths for conduct of certain stages and steps within the overall research process, rather than as alternative research processes that function independently of one another.

The underlying purposes or interests that these alternative styles reflect—i.e., focal interest in substantive, conceptual, or methodological matters—are likely to shape the emphases, attention, and care given to various aspects of the research process (and hence to various aspects of validity). Not only is a researcher's "favorite" domain likely to get first and most attention (which poses no problems in and of itself), but the "least preferred" domain is likely to get last and least attention, thereby leaving some validity issues pertinent to that domain least well served. Choices regarding elements and relations of that third domain are likely to be made to accommodate results of the prior step. Concretely, this means that a researcher following the Stage 2 "experimental path" is likely to choose samples of substantive events on the basis of their convenience for the design that has already been built in Step 1 of that path. Similarly, a researcher following the "theoretical path" is likely to choose "methods of convenience" to test a theory, while one using the "empirical path" is likely to select "concepts of convenience" to interpret a set of observations. By allowing our selection from the third domain (i.e., Step 2) to be based on convenience of fit to the structure already built in Step 1, we increase the probability that a Step 2 "fit" will be obtained. Yet we accomplish that fit at the risk of trivializing the Step 2 validity questions and of shifting the definition of the focal problem we purport to be studying.

Notice, too, that these different purposes have their major effects in Stage 2 (which we regard as dealing with our broadened notion of internal validity) and have relatively little impact on Stage 3 (which deals with external validity). We argue that, regardless of the path by which one arrives at a set of empirical findings (the common end result of all Stage 2 paths), one still needs to carry out the replication, robustness analyses, *and* boundary search activities of Stage 3 in order to have confidence in (i.e., reduce uncertainty about) those findings as interpreted.

The preceding points raise the issue of "who" it is that "must" do these things. We intend, throughout, not to be prescriptive about what the activities of any individual researcher ought to be. Rather, we wish only to prescribe what a collective "community of researchers" who are attempting to learn about a particular focal problem area must do if they are, collectively, to gain scientific information about that problem area and build confidence in that information. The individual researcher, carrying out a study that fits his or her interests or purposes, is in no way obliged to conduct research along any other path or in any other portion of the research "space" than the one proposed. The individual researcher is obliged only to do each study as well as possible within available resources, and to present it publicly for what it is: one study, in one part of the overall research process, bearing on the stated focal problem in certain limited ways. The "field," on the other hand (the collective community of researchers interested in a particular focal problem area), must ensure that all portions of the research process get sufficient attention and exploration, so that the community of researchers can increase their confidence (reduce their uncertainty) about the focal problem findings and their meanings.

These points are pertinent to the Calder/Lynch discourse. We would agree that no individual researcher is obliged to study "real" populations, to sample representatively, to be concerned with testing theoretical implications intended to be universal, or to search for the scope and boundaries of a particular set of research findings. Nor is an individual

researcher who is interested in doing Stage 2 research on a particular focal problem obliged to go back and develop completely new methods of measurement, analysis techniques, interpretations of events in some particular substantive domain, and conceptual models. Scientific information is not only inherently probabilistic and contingent (as noted in Theme #2), it is also cumulative over different research studies and different researchers. Conceptual and methodological tools, as well as substantive findings, *must be cumulative;* otherwise each of us would have to start from scratch again whenever we tried to do a research study of any kind.

On the other hand, the field is obliged to deal with all of these parts of the research process, if the field is to "know what it knows." This is equally true for both applied and basic research areas. In our view, although basic and applied research may be regarded as analytically distinct in terms of short-term purposes, the two are ultimately interdependent in terms of long-run gains in knowledge, and in terms of our confidence about that knowledge.

In this light, questions such as whether or not the basic researcher must be concerned with external validity become moot points. No particular researcher must, but the field must. The scope and limits of the basic researcher's findings—and of the applied researcher's findings as well—must be established with respect to all relevant facets of the conceptual and the methodological, as well as of the substantive domain.

In accepting these caveats, we must of course remember that the individuals who make up "the field" are the same scientists who do the separate, individual studies. While no one of them is individually responsible for doing all that is needed in the research area, each shares in the collective responsibility for doing so. Much like the well known "tragedy of the commons" (Hardin 1968), if every one of us leaves all of the preparatory work (Stage 1) and the verification and delimitation work (Stage 3) to "the field," it will not get done at all: "the field"—and all of us who comprise it—will fail. Pogo's classic statement is apropos here: "We have met the enemy, and they is us!"

CLOSING COMMENTS

Berkowitz and Donnerstein (1982) recently published an article entitled "External Validity is More Than Skin Deep." Their title meant, in part, that they rejected the idea that external validity was synonymous with realism. Calder and Lynch would agree in that rejection, and so would we (as noted previously under Theme #1). Indeed, we would hold that external validity is not only deeper than "mere realism," it is also broader than "mere population sampling," and much more complicated than merely "generalizing to"—or even "generalizing over"—variations in some single feature of the design or sample of an earlier study. In our view, fully exploring the external validity of a set of findings requires systematic efforts to verify, extend, and delimit those findings, by replication and by simultaneous robustness analysis and boundary search with respect to all relevant facets of the conceptual, methodological, and substantive domains. For both basic researchers and applied researchers, to do less is to settle for a higher level of uncertainty about those findings than is necessary. And in science, that is tantamount to throwing the game!

[*Received April 1983.*]

REFERENCES

Berkowitz, Leonard and Edward Donnerstein (1982), "External Validity Is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments," *American Psychologist,* 37(3), 245–257.

Brinberg, David and Joseph E. McGrath (1982), "Network of Validity Concepts within the Research Process," in *New Directions for Methodology of Social and Behavioral Science: Forms of Validity in Research,* eds. David Brinberg and L. Kidder, San Francisco: Jossey-Bass.

——— and Joseph E. McGrath (1983), "A Validity Network Schema," paper presented at a convention of the American Educational Researchers Association, April.

Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout (1981), "Designing Research for Application," *Journal of Consumer Research,* 8(2), 197–207.

———, Lynn W. Phillips, and Alice M. Tybout (1982), "The Concept of External Validity," *Journal of Consumer Research,* 9(3), 240–244.

———, Lynn W. Phillips, "Beyond External Validity," *Journal of Consumer Research,* 10(1), 112–114.

Campbell, Donald T. and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin,* 56, 81–105.

——— and Julian C. Stanley (1966), *Experimental and Quasi-Experimental Designs for Research,* Chicago: Rand McNally.

Cook, Thomas D. and Donald T. Campbell (1979), *Design and Analysis of Quasi-Experiments for Field Settings,* Chicago: Rand McNally.

Cronbach, Lee J. (1975), "Beyond the Two Disciplines of Scientific Psychology," *American Psychologist,* 29, 116–127.

——— (1982), *Designing Evaluation of Educational Programs,* San Francisco: Jossey-Bass.

Ferber, Robert (1977), "Research by Convenience," *Journal of Consumer Research,* 4(1), 57–58.

Hardin, G. R. (1968), "The Tragedy of the Commons," *Science,* 162, 1243–1248.

Lynch, John G., Jr. "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research,* 9(3), 225–239.

——— (1983), "The Role of External Validity in Theoretical Research," *Journal of Consumer Research,* 10(1), 109–111.

McGrath, Joseph E., Joanne Martin, and Richard A. Kulka (1982), *Judgment Calls in Research,* Beverly Hills, CA: Sage Publications.

Popper, Karl (1959), *The Logic of Scientific Discovery,* New York: Basic Books.

Runkel, Phillip J. and Joseph E. McGrath (1972), *Research on Human Behavior: A Systematic Guide to Method,* New York: Holt, Rinehart & Winston.

Webb, Eugene, Donald Campbell, Richard Schwartz, and Lee Sechrest (1966), *Unobtrusive Measures: Nonreactive Research in the Social Sciences,* Chicago: Rand McNally.

Wimsatt, W. C. (1981), "Robustness, Reliability and Overdetermination," in *Scientific Inquiry and the Social Sciences,* eds. M. Brewer and B. Collins, San Francisco: Jossey-Bass.

24

In Defense of External Invalidity

Douglas G. Mook

The greatest weakness of laboratory experiments lies in their artificiality. Social processes observed to occur within a laboratory setting might not necessarily occur within more natural social settings.

—Babbie, 1975, p. 254

In order to behave like scientists we must construct situations in which our subjects . . . can behave as little like human beings as possible and we do this in order to allow ourselves to make statements about the nature of their humanity.

—Bannister, 1966, p. 24

Experimental psychologists frequently have to listen to remarks like these. And one who has taught courses in research methods and experimental psychology, as I have for the past several years, has probably had no problem in alerting students to the "artificiality" of research settings. Students, like laypersons (and not a few social scientists for that matter), come to us quite prepared to point out the remoteness of our experimental chambers, our preoccupation with rats and college sophomores, and the comic-opera "reactivity" of our shock generators, electrode paste, and judgments of lengths of line segments on white paper.

They see all this. My problem has been not to alert them to these considerations, but to break their habit of dismissing well-done, meaningful, informative research on grounds of "artificiality."

The task has become a bit harder over the last few years because a full-fledged "purr" word has gained currency: *external validity.* Articles and monographs have been written about its proper nurture, and checklists of specific threats to its well-being are now appearing in textbooks. Studies unescorted by it are afflicted by—what else?—*external invalidity.* That phrase has a lovely mouth-filling resonance to it, and there is, to be sure, a certain poetic justice in our being attacked with our own jargon.

Warm Fuzzies and Cold Creepies

The trouble is that, like most "purr" and "snarl" words, the phrases *external validity* and *external invalidity* can serve as serious barriers to thought. Obviously, any kind of validity is a warm, fuzzy Good Thing; and just as obviously, any kind of invalidity must be a cold, creepy Bad Thing. Who could doubt it?

It seems to me that these phrases trapped even their originators, in just that way. Campbell and Stanley (1967) introduce the concept thus: "*External validity* asks the question of *generalizability:* To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (p. 5). Fair enough. External validity is not an automatic desideratum; it *asks a question.* It invites us to think about the prior questions: To what populations, settings, and so on, do we *want* the effect to be generalized? Do we want to generalize it at all?

But their next sentence is: "Both types of criteria are obviously important. . ." And ". . . the selection of designs strong in both types of validity is obviously our ideal" (Campbell & Stanley, 1967, p. 5).

I intend to argue that this is simply wrong. If it sounds plausible, it is because the word *validity* has given it a warm coat of downy fuzz. Who wants to be invalid—internally, externally, or in any other way? One might as well ask for acne. In a way, I wish the authors had stayed with the term *generalizability,* precisely because it does not sound nearly so good. It would then be easier to remember that we are not dealing with a criterion, like clear skin, but with a question, like "How can we get this sofa down the stairs?" One asks that question if, and only if, moving the sofa is what one wants to do.

But *generalizability* is not quite right either. The question of external validity is not the same as the question of generalizability. Even an experiment

From Douglas G. Mook, "In Defense of External Invalidity," *American Psychologist,* 1983, 38(4), 379–387.

that is clearly "applicable to the real world," perhaps because it was conducted there (e.g., Bickman's, 1974, studies of obedience on the street corner), will have *some* limits to its generalizability. Cultural, historical, and age-group limits will surely be present; but these are unknown and no single study can discover them all. Their determination is empirical.

The external-validity question is a special case. It comes to this: Are the sample, the setting, and the manipulation so artificial that the class of "target" real-life situations to which the results can be generalized is likely to be trivially small? If so, the experiment lacks external validity. But that argument still begs the question I wish to raise here: Is such generalization our intent? Is it what we want to do? Not always.

The Agricultural Model

These baleful remarks about external validity (EV) are not quite fair to its originators. In defining the concept, they had a particular kind of research in mind, and it was the kind in which the problem of EV is meaningful and important.

These are the applied experiments. Campbell and Stanley (1967) had in mind the kind of investigation that is designed to evaluate a new teaching procedure or the effects of an "enrichment" program on the culturally deprived. For that matter, the research context in which sampling theory was developed in its modern form—agricultural research—has a similar purpose. The experimental setting resembles, or is a special case of, a real-life setting in which one wants to know what to do. Does this fertilizer (or this pedagogical device) promote growth in this kind of crop (or this kind of child)? If one finds a significant improvement in the experimental subjects as compared with the controls, one predicts that implementation of a similar manipulation, in a similar setting with similar subjects, will be of benefit on a larger scale.

That kind of argument does assume that one's experimental manipulation represents the broader-scale implementation and that one's subjects and settings represent their target populations. Indeed, part of the thrust of the EV concept is that we have been concerned only with subject representativeness and not enough with representativeness of the settings and manipulations we have sampled in doing experiments.

Deese (1972), for example, has taken us to task for this neglect:

> Some particular set of conditions in an experiment is generally taken to be representative of all possible conditions of a similar type. . . . In the investigation of altruism, situations are devised to permit people to make altruistic choices. Usually a single situation provides the setting for the experimental testing. . . . [the experimenter] will allow that one particular situation to stand for the unspecified circumstances in which an individual could be altruistic. . . . the social psychologist as experimenter is content to let a particular situation stand for an indefinite range of possible testing situations in a vague and unspecified way. (pp. 59–60)

It comes down to this: The experimenter is generalizing on the basis of a small and biased sample, not of subjects (though probably those too), but of settings and manipulations.[1]

The entire argument rests, however, on an applied, or what I call an "agricultural," conception of the aims of research. The assumption is that the experiment is *intended* to be generalized to similar subjects, manipulations, and settings. If this is so, then the broader the generalizations one can make, the more real-world occurrences one can predict from one's findings and the more one has learned about the real world from them. However, it may not be so. There are experiments—very many of them—that do not have such generalization as their aim.

This is not to deny that we have talked nonsense on occasion. We have. Sweeping generalizations about "altruism," or "anxiety," or "honesty" have been made on evidence that does not begin to support them, and for the reasons Deese gives. But let it also be said that in many such cases, we have seemed to talk nonsense only because our critics, or we ourselves, have assumed that the "agricultural" goal of generalization is part of our intent.

But in many (perhaps most) of the experiments Deese has in mind, the logic goes in a different direction. We are not *making* generalizations, but *testing* them. To show what a difference this makes, let me turn to an example.

A Case Study of a Flat Flunk

Surely one of the experiments that has had permanent impact on our thinking is the study of "mother love" in rhesus monkeys, elegantly conducted by Harlow. His wire mothers and terry-cloth mothers are permanent additions to our vocabulary of classic manipulations. And his finding that con-

I thank James E. Deese and Wayne Shebilske for their comments on an earlier version of this article.

Requests for reprints should be sent to Douglas G. Mook, Department of Psychology, University of Virginia, Charlottesville, Virginia 22901.

[1] In fairness, Deese goes on to make a distinction much like the one I intend here. "If the theory and observations are explicitly related to one another through some rigorous logical process, then the sampling of conditions may become completely unnecessary" (p. 60). I agree. "But a theory having such power is almost never found in psychology" (p. 61). I disagree, not because I think our theories are all that powerful, but because I do not think all that much power is required for what we are usually trying to do.

tact comfort was a powerful determinant of "attachment," whereas nutrition was small potatoes, was a massive spike in the coffin of the moribund, but still wriggling, drive-reduction theories of the 1950s.

As a case study, let us see how the Harlow wire- and cloth-mother experiment stands up to the criteria of EV.

The original discussion of EV by Campbell and Stanley (1967) reveals that the experimental investigation they had in mind was a rather complex mixed design with pretests, a treatment imposed or withheld (the independent variable), and a posttest. Since Harlow's experiment does not fit this mold, the first two of their "threats to external validity" do not arise at all: pretest effects on responsiveness and multiple-treatment interference.

The other two threats on their list do arise in Harlow's case. First, "there remains the possibility that the effects . . . hold only for that unique population from which the . . . [subjects were] selected" (Campbell & Stanley, 1967, p. 19). More generally, this is the problem of sampling bias, and it raises the spectre of an unrepresentative sample. Of course, as every student knows, the way to combat the problem (and never mind that nobody does it) is to select a random sample from the population of interest.

Were Harlow's baby monkeys representative of the population of monkeys in general? Obviously not; they were born in captivity and then orphaned besides. Well, were they a representative sample of the population of lab-born, orphaned monkeys? There was no attempt at all to make them so. It must be concluded that Harlow's sampling procedures fell far short of the ideal.

Second, we have the undeniable fact of the "patent artificiality of the experimental setting" (Campbell & Stanley, 1967, p. 20). Campbell and Stanley go on to discuss the problems posed by the subjects' knowledge that they are in an experiment and by what we now call "demand characteristics." But the problem can be generalized again: How do we know that what the subjects do in this artificial setting is what they would do in a more natural one? Solutions have involved hiding from the subjects the fact that they are subjects; moving from a laboratory to a field setting; and, going further, trying for a "representative sample" of the field settings themselves (e.g., Brunswik, 1955).

What then of Harlow's work? One does not know whether his subjects knew they were in an experiment; certainly there is every chance that they experienced "expectations of the unusual, with wonder and active puzzling" (Campbell & Stanley, 1967, p. 21). In short, they must have been cautious, bewildered, reactive baby monkeys indeed. And what of the representativeness of the setting? Real monkeys do not live within walls. They do not encounter mother figures made of wire mesh, with rubber nipples; nor is the advent of a terry-cloth cylinder, warmed by a light bulb, a part of their natural lifestyle. What can this contrived situation possibly tell us about how monkeys with natural upbringing would behave in a natural setting?

On the face of it, the verdict must be a flat flunk. On every criterion of EV that applies at all, we find Harlow's experiment either manifestly deficient or simply unevaluable. And yet our tendency is to respond to this critique with a resounding "So what?" And I think we are quite right to so respond.

Why? Because using the lab results to make generalizations about real-world behavior was no part of Harlow's intention. It was not what he was trying to do. That being the case, the concept of EV simply does not arise—except in an indirect and remote sense to be clarified shortly.

Harlow did not conclude, "Wild monkeys in the jungle probably would choose terry-cloth over wire mothers, too, if offered the choice." First, it would be a moot conclusion, since that simply is not going to happen. Second, who cares whether they would or not? The generalization would be trivial even if true. What Harlow did conclude was that the hunger-reduction interpretation of mother love would not work. If anything about his experiment has external validity, it is this theoretical point, not the findings themselves. And to see whether the theoretical conclusion is valid, we extend the experiments or test predictions based on theory.[2] We do not dismiss the findings and go back to do the experiment "properly," in the jungle with a random sample of baby monkeys.

The distinction between generality of findings and generality of theoretical conclusions underscores what seems to me the most important source of confusion in all this, which is the assumption that the purpose of collecting data in the laboratory is to *predict real-life behavior in the real world.* Of course, there are times when that is what we are trying to do, and there are times when it is not. When it is, then the problem of EV confronts us, full force. When it is not, then the problem of EV is either meaningless or trivial, and a misplaced preoccupation with it can seriously distort our evaluation of the research.

But if we are not using our experiments to predict real-life behavior, what are we using them for? Why else do an experiment?

[2] The term *theory* is used loosely to mean, not a strict deductive system, but a conclusion on which different findings converge. Harlow's demonstration draws much of its force from the context of other findings (by Ainsworth, Bowlby, Spitz, and others) with which it articulates.

There are a number of other things we may be doing. First, we may be asking whether something *can* happen, rather than whether it typically *does* happen. Second, our prediction may be in the other direction; it may specify something that ought to happen *in the lab,* and so we go to the lab to see whether it does. Third, we may demonstrate the power of a phenomenon by showing that it happens even under unnatural conditions that ought to preclude it. Finally, we may use the lab to produce conditions that have no counterpart in real life at all, so that the concept of "generalizing to the real world" has no meaning. But even where findings cannot possibly generalize and are not supposed to, they can contribute to an understanding of the processes going on. Once again, it is that understanding which has external validity (if it does)—not the findings themselves, much less the setting and the sample. And this implies in turn that we cannot assess that kind of validity by examining the experiment itself.

Alternatives to Generalization

"What Can" Versus "What Does"

"Person perception studies using photographs or brief exposure of the stimulus person have commonly found that spectacles, lipstick and untidy hair have a great effect on judgments of intelligence and other traits. It is suggested . . . that these results are probably exaggerations of any effect that might occur when more information about a person is available" (Argyle, 1969, p. 19). Later in the same text, Argyle gives a specific example: "Argyle and McHenry found that targeted persons were judged as 13 points of IQ more intelligent when wearing spectacles and when seen for 15 seconds; however, if they were seen during 5 minutes of conversation spectacles made no difference" (p. 135).

Argyle (1969) offers these data as an example of how "the results [of an independent variable studied in isolation] may be exaggerated" (p. 19). Exaggerated with respect to what? With respect to what "really" goes on in the world of affairs. It is clear that on these grounds, Argyle takes the 5-minute study, in which glasses made no difference, more seriously than the 15-second study, in which they did.

Now from an "applied" perspective, there is no question that Argyle is right. Suppose that only the 15-second results were known; and suppose that on the basis of them, employment counselors began advising their students to wear glasses or sales executives began requiring their salespeople to do so. The result would be a great deal of wasted time, and all because of an "exaggerated effect," or what I have called an "inflated variable" (Mook, 1982). Powerful in the laboratory (13 IQ points is a lot!), eyeglasses are a trivial guide to a person's intelligence and are treated as such when more information is available.

On the other hand, is it not worth knowing that such a bias *can* occur, even under restricted conditions? Does it imply an implicit "theory" or set of "heuristics" that we carry about with us? If so, where do they come from?

There are some intriguing issues here. Why should the person's wearing eyeglasses affect our judgments of his or her intelligence under any conditions whatever? As a pure guess, I would hazard the following: Maybe we believe that (a) intelligent people read more than less intelligent ones, and (b) that reading leads to visual problems, wherefore (c) the more intelligent are more likely to need glasses. If that is how the argument runs, then it is an instance of how our person perceptions are influenced by causal "schemata" (Nisbett & Ross, 1980)—even where at least one step in the theoretical sequence ([b] above) is, as far as we know, simply false.

Looked at in that way, the difference between the 15-second and the 5-minute condition is itself worth investigating further (as it would not be if the latter simply "invalidated" the former). If we are so ready to abandon a rather silly causal theory in the light of more data, why are some other causal theories, many of them even sillier, so fiercely resistant to change?

The point is that in thinking about the matter this way, we are taking the results strictly as we find them. The fact that eyeglasses *can* influence our judgments of intelligence, though it may be quite devoid of real-world application, surely says something about us as judges. If we look just at that, then the issue of external validity does not arise. We are no longer concerned with generalizing from the lab to the real world. The lab (qua lab) has led us to ask questions that might not otherwise occur to us. Surely that alone makes the research more than a sterile intellectual exercise.

Predicting From and Predicting To

The next case study has a special place in my heart. It is one of the things that led directly to this article, which I wrote fresh from a delightful roaring argument with my students about the issues at hand.

The study is a test of the tension-reduction view of alcohol consumption, conducted by Higgins and Marlatt (1973). Briefly, the subjects were made either highly anxious or not so anxious by the threat of electric shock, and were permitted access to alcohol as desired. If alcohol reduces tension and if people drink it because it does so (Cappell & Herman, 1972), then the anxious subjects should have drunk more. They did not.

Writing about this experiment, one of my better students gave it short shrift: "Surely not many al-

coholics are presented with such a threat under normal conditions."

Indeed. The threat of electric shock can hardly be "representative" of the dangers faced by anyone except electricians, hi-fi builders, and Psychology 101 students. What then? It depends! It depends on what kind of conclusion one draws and what one's purpose is in doing the study.

Higgins and Marlatt could have drawn this conclusion: "Threat of shock did not cause our subjects to drink in these circumstances. Therefore, it probably would not cause similar subjects to drink in similar circumstances either." A properly cautious conclusion, and manifestly trivial.

Or they could have drawn this conclusion: "Threat of shock did not cause our subjects to drink in these circumstances. Therefore, tension or anxiety probably does not cause people to drink in normal, real-world situations." That conclusion would be manifestly risky, not to say foolish; and it is that kind of conclusion which raises the issue of EV. Such a conclusion does assume that we can generalize from the simple and protected lab setting to the complex and dangerous real-life one and that the fear of shock can represent the general case of tension and anxiety. And let me admit again that we have been guilty of just this kind of foolishness on more than one occasion.

But that is not the conclusion Higgins and Marlatt drew. Their argument had an entirely different shape, one that changes everything. Paraphrased, it went thus: "Threat of shock did not cause our subjects to drink in these circumstances. Therefore, the tension-reduction hypothesis, which predicts that it should have done so, either is false or is in need of qualification." This is our old friend, the hypothetico-deductive method, in action. The important point to see is that the generalizability of the results, from lab to real life, is not claimed. It plays no part in the argument at all.

Of course, these findings may not require *much* modification of the tension-reduction hypothesis. It is possible—indeed it is highly likely—that there are tensions and tensions; and perhaps the nagging fears and self-doubts of the everyday have a quite different status from the acute fear of electric shock. Maybe alcohol does reduce these chronic fears and is taken, sometimes abusively, because it does so.[3] If these possibilities can be shown to be true, then we could sharpen the tension-reduction hypothesis, restricting it (as it is not restricted now) to certain kinds of tension and, perhaps, to certain settings. In short, we could advance our understanding. And the "artificial" laboratory findings would have contributed to that advance. Surely we cannot reasonably ask for more.

It seems to me that this kind of argument characterizes much of our research—much more of it than our critics recognize. In very many cases, we are not using what happens in the laboratory to "predict" the real world. Prediction goes the other way: Our theory specifies what subjects should do *in the laboratory.* Then we go to the laboratory to ask, Do they do it? And we modify our theory, or hang onto it for the time being, as results dictate. Thus we improve our theories, and—to say it again—it is these that generalize to the real world if anything does.

Let me turn to an example of another kind. To this point, it is artificiality of *setting* that has been the focus. Analogous considerations can arise, however, when one thinks through the implications of artificiality of, or bias in, the *sample.* Consider a case study.

A great deal of folklore, supported by some powerful psychological theories, would have it that children acquire speech of the forms approved by their culture—that is, grammatical speech—through the impact of parents' reactions to what they say. If a child emits a properly formed sentence (so the argument goes), the parent responds with approval or attention. If the utterance is ungrammatical, the parent corrects it or, at the least, withholds approval.

Direct observation of parent–child interactions, however, reveals that this need not happen. Brown and Hanlon (1970) report that parents react to the content of a child's speech, not to its form. If the sentence emitted is factually correct, it is likely to be approved by the parent; if false, disapproved. But whether the utterance embodies correct grammatical form has surprisingly little to do with the parent's reaction to it.

What kind of sample were Brown and Hanlon dealing with here? Families that (a) lived in Boston, (b) were well educated, and (c) were willing to have squadrons of psychologists camped in their living rooms, taping their conversations. It is virtually certain that the sample was biased even with respect to the already limited "population" of upper-class-Bostonian-parents-of-young-children.

Surely a sample like that is a poor basis from which to generalize to any interesting population. But what if we turn it around? We start with the theoretical proposition: Parents respond to the grammar of their children's utterances (as by making approval contingent or by correcting mistakes). Now we make the prediction: Therefore, the *parents*

[3] I should note, however, that there is considerable doubt about that as a statement of the general case. Like Harlow's experiment, the Higgins and Marlatt (1973) study articulates with a growing body of data from very different sources and settings, but all, in this case, calling the tension-reduction theory into question (cf. Mello & Mendelson, 1978).

we observe ought to do that. And the prediction is disconfirmed.

Going further, if we find that the children Brown and Hanlon studied went on to acquire Bostonian-approved syntax, as seems likely, then we can draw a further prediction and see it disconfirmed. If the theory is true, and if *these* parents do not react to grammaticality or its absence, then *these* children should not pick up grammatical speech. If they do so anyway, then parental approval is not necessary for the acquisition of grammar. And that is shown not by generalizing from sample to population, but by what happened *in the sample.*

It is of course legitimate to wonder whether the same contingencies would appear in Kansas City working-class families or in slum dwellers in the Argentine. Maybe parental approval/disapproval is a much more potent influence on children's speech in some cultures or subcultures than in others. Nevertheless, the fact would remain that the parental approval theory holds only in some instances and must be qualified appropriately. Again, that would be well worth knowing, and *this* sample of families would have played a part in establishing it.

The confusion here may reflect simple historical accident. Considerations of sampling from populations were brought to our attention largely by survey researchers, for whom the procedure of "generalizing to a population" is of vital concern. If we want to estimate the proportion of the electorate intending to vote for Candidate *X*, and if *Y*% of our sample intends to do so, then we want to be able to say something like this: "We can be 95% confident that *Y*% of the voters, plus or minus *Z*, intend to vote for *X*." Then the issue of representativeness is squarely before us, and the horror stories of biased sampling and wildly wrong predictions, from the *Literary Digest* poll on down, have every right to keep us awake at night.

But what has to be thought through, case by case, is whether that is the kind of conclusion we intend to draw. In the Brown and Hanlon (1970) case, nothing could be more unjustified than a statement of the kind, "We can be *W*% certain that *X*% of the utterances of Boston children, plus or minus *Y*, are true and are approved." The biased sample rules such a conclusion out of court at the outset. But it was never intended. The intended conclusion was not about a population but about a theory. That parental approval tracks content rather than form, in *these children,* means that the parental approval theory of grammar acquisition either is simply false or interacts in unsuspected ways with some attribute(s) of the home.

In yet other cases, the subjects are of interest precisely because of their unrepresentativeness. Washoe, Sarah, and our other special students are of interest because they are not representative of a language-using species. And with all the quarrels their accomplishments have given rise to, I have not seem them challenged as "unrepresentative chimps," except by students on examinations (I am not making that up). The achievements of mnemonists (which show us what *can* happen, rather than what typically *does*) are of interest because mnemonists are not representative of the rest of us. And when one comes across a mnemonist one studies that mnemonist, without much concern for his or her representativeness even as a mnemonist.

But what do students read? "Samples should always be as representative as possible of the population under study." "[A] major concern of the behavioral scientist is to ensure that the sample itself is a good representative [sic] of the population." (The sources of these quotations do not matter; they come from an accidental sample of books on my shelf.)

The trouble with these remarks is not that they are false—sometimes they are true—but that they are unqualified. Representativeness of sample is of vital importance for certain purposes, such as survey research. For other purposes it is a trivial issue.[4] Therefore, one must evaluate the sampling procedure in light of the purpose—separately, case by case.

Taking the Package Apart

Everyone knows that we make experimental settings artificial for a reason. We do it to control for extraneous variables and to permit separation of factors that do not come separately in Nature-as-you-find-it. But that leaves us wondering how, having stepped out of Nature, we get back in again. How do our findings apply to the real-life setting in all its complexity?

I think there are times when the answer has to be, "They don't." But we then may add, "Something else does. It is called understanding."

---

[4] There is another sense in which "generalizing to a population" attends most psychological research: One usually tests the significance of one's findings, and in doing so one speaks of sample values as estimates of population parameters. In this connection, though, the students are usually reassured that they can always define the population in terms of the sample and take it from there—which effectively leaves them wondering what all the flap was about in the first place.

Perhaps this is the place to note that some of the case studies I have presented may raise questions in the reader's mind that are not dealt with here. Some raise the problem of interpreting null conclusions; adequacy of controls for confounding variables may be worrisome; and the Brown and Hanlon (1970) study faced the problem of observer effects (adequately dealt with, I think; see Mook, 1982). Except perhaps for the last one, however, these issues are separate from the problem of external validity, which is the only concern here.

As an example, consider dark adaptation. Psychophysical experiments, conducted in restricted, simplified, ecologically invalid settings, have taught us these things among others:

1. Dark adaptation occurs in two phases. There is a rapid and rather small increase in sensitivity, followed by a delayed but greater increase.

2. The first of these phases reflects dark adaptation by the cones; the second, by the rods.

Hecht (1934) demonstrated the second of these conclusions by taking advantage of some facts about cones (themselves established in ecologically invalid photochemical and histological laboratories). Cones are densely packed near the fovea; and they are much less sensitive than the rods to the shorter visible wavelengths. Thus, Hecht was able to tease out the cone component of the dark-adaptation curve by making his stimuli small, restricting them to the center of the visual field, and turning them red.

Now let us contemplate the manifest ecological invalidity of this setting. We have a human subject in a dark room, staring at a place where a tiny red light may appear. Who on earth spends time doing that, in the world of affairs? And on each trial, the subject simply makes a "yes, I see it/no, I don't" response. Surely we have subjects who "behave as little like human beings as possible" (Bannister, 1966)—We might be calibrating a photocell for all the difference it would make.

How then do the findings apply to the real world? They do not. The task, variables, and setting have no real-world counterparts. What does apply, and in spades, is the understanding of how the visual system works that such experiments have given us. That is what we apply to the real-world setting—to flying planes at night, to the problem of reading X-ray prints on the spot, to effective treatment of night blindness produced by vitamin deficiency, and much besides.

Such experiments, I say, give us understanding of real-world phenomena. Why? Because the *processes* we dissect in the laboratory also operate in the real world. The dark-adaptation data are of interest because they show us a process that does occur in many real-world situations. Thus we could, it is true, look at the laboratory as a member of a class of "target" settings to which the results apply. But it certainly is not a "representative" member of that set. We might think of it as a limiting, or even *defining,* member of that set. To what settings do the results apply? The shortest answer is: to any setting in which it is relevant that (for instance) as the illumination dims, sensitivity to longer visible wavelengths drops out before sensitivity to short ones does. The findings do not represent a class of real-world phenomena; they define one.

Alternatively, one might use the lab not to explore a known phenomenon, but to determine whether such and such a phenomenon exists or can be made to occur. (Here again the emphasis is on what can happen, not what usually does.) Henshel (1980) has noted that some intriguing and important phenomena, such as biofeedback, could never have been discovered by sampling or mimicking natural settings. He points out, too, that if a desirable phenomenon occurs under laboratory conditions, one may seek to make natural settings mimic the laboratory rather than the other way around. Engineers are familiar with this approach. So, for instance, are many behavior therapists.

(I part company with Henshel's excellent discussion only when he writes, "The requirement of 'realism,' or a faithful mimicking of the outside world in the laboratory experiment, applies only to . . . hypothesis testing within the logico-deductive model of research" [p. 470]. For reasons given earlier, I do not think it need apply even there.)

The Drama of the Artificial

To this point, I have considered alternatives to the "analogue" model of research and have pointed out that we need not intend to generalize our results from sample to population, or from lab to life. There are cases in which we do want to do that, of course. Where we do, we meet another temptation: We may assume that in order to *generalize* to "real life," the laboratory setting should *resemble* the real-life one as much as possible. This assumption is the force behind the cry for "representative settings."

The assumption is false. There are cases in which the generalization from research setting to real-life settings is made all the stronger by the lack of resemblance between the two. Consider an example.

A research project that comes in for criticism along these lines is the well-known work on obedience by Milgram (1974). In his work, the difference between a laboratory and a real-life setting is brought sharply into focus. Soldiers in the jungles of Viet Nam, concentration camp guards on the fields of Eastern Europe—what resemblance do their environments bear to a sterile room with a shock generator and an intercom, presided over by a white-coated scientist? As a setting, Milgram's surely is a prototype of an "unnatural" one.

One possible reaction to that fact is to dismiss the work bag and baggage, as Argyle (1969) seems to do: "When a subject steps inside a psychological laboratory he steps out of culture, and all the normal rules and conventions are temporarily discarded and replaced by the single rule of laboratory culture—'do what the experimenter says, no matter how absurd or unethical it may be' " (p. 20). He goes on to cite Milgram's work as an example.

All of this—which is perfectly true—comes in a discussion of how "laboratory research can produce the wrong results" (Argyle, 1969, p. 19). The wrong results! But that is the whole point of the results. What Milgram has shown is how easily we can "step out of culture" in just the way Argyle describes—and how, once out of culture, we proceed to violate its "normal rules and conventions" in ways that are a revelation to us when they occur. Remember, by the way, that most of the people Milgram interviewed grossly underestimated the amount of compliance that would occur *in that laboratory setting.*

Another reaction, just as wrong but unfortunately even more tempting, is to start listing similarities and differences between the lab setting and the natural one. The temptation here is to get involved in count-'em mechanics: The more differences there are, the greater the external invalidity. Thus:

> One element lacking in Milgram's situation that typically obtains in similar naturalistic situations is that the experimenter had no real power to harm the subject if the subject failed to obey orders. The subject could always simply get up and walk out of the experiment, never to see the experimenter again. So when considering Milgram's results, it should be borne in mind that a powerful source of obedience in the real world was lacking in this situation. (Kantowitz & Roediger, 1978, pp. 387–388)

"Borne in mind" to what conclusion? Since the next sentence is "Nonetheless, Milgram's results are truly remarkable" (p. 388), we must suppose that the remarks were meant in criticism.

Now the lack of threat of punishment is, to be sure, a major difference between Milgram's lab and the jungle war or concentration camp setting. But what happened? An astonishing two thirds obeyed anyway. The force of the experimenter's authority was sufficient to induce normal decent adults to inflict pain on another human being, even though they could have refused without risk. Surely the absence of power to punish, though a distinct difference between Milgram's setting and the others, only adds to the drama of what he saw.

There are other threats to the external validity of Milgram's findings, and some of them must be taken more seriously. There is the possibility that the orders he gave were "legitimized by the laboratory setting" (Orne & Evans, 1965, p. 199). Perhaps his subjects said in effect, "This is a scientific experiment run by a responsible investigator, so maybe the whole business isn't as dangerous as it looks." This possibility (which is quite distinct from the last one, though the checklist approach often confuses the two) does leave us with nagging doubts about the generalizability of Milgram's findings. Camp guards and jungle fighters do not have this cognitive escape hatch available to them. If Milgram's subjects did say "It must not be dangerous," then his conclusion—people are surprisingly willing to inflict danger under orders—is in fact weakened.

The important thing to see is that the checklist approach will not serve us. Here we have two differences between lab and life—the absence of punishment and the possibility of discounting the danger of obedience. The latter difference weakens the impact of Milgram's findings; the former strengthens it. Obviously we must move beyond a simple count of differences and think through what the effect of each one is likely to be.

Validity of What?

Ultimately, what makes research findings of interest is that they help us understand everyday life. That understanding, however, comes from theory or the analysis of mechanism; it is not a matter of "generalizing" the findings themselves. This kind of validity applies (if it does) to statements like "The hunger-reduction interpretation of infant attachment will not do," or "Theory-driven inferences may bias first impressions," or "The Purkinje shift occurs because rod vision has *these* characteristics and cone vision has *those.*" The validity of these generalizations is tested by their success at prediction and has nothing to do with the naturalness, representativeness, or even nonreactivity of the investigations on which they rest.

Of course there are also those cases in which one does want to predict real-life behavior directly from research findings. Survey research, and most experiments in applied settings such as factory or classroom, have that end in view. Predicting real-life behavior is a perfectly legitimate and honorable way to use research. When we engage in it, we do confront the problem of EV, and Babbie's (1975) comment about the artificiality of experiments has force.

What I have argued here is that Babbie's comment has force *only* then. If this is so, then external validity, far from being "obviously our ideal" (Campbell & Stanley, 1967), is a concept that applies only to a rather limited subset of the research we do.

A Checklist of Decisions

I am afraid that there is no alternative to thinking through, case by case, (a) what conclusion we want to draw and (b) whether the specifics of our sample or setting will prevent us from drawing it. Of course there are seldom any fixed rules about how to "think through" anything interesting. But here is a sample of questions one might ask in deciding whether the usual criteria of external validity should even be considered:

As to the sample: Am I (or is he or she whose work I am evaluating) trying to estimate from sam-

ple characteristics the characteristics of some population? Or am I trying to draw conclusions not about a population, but about a theory that specifies what *these* subjects ought to do? Or (as in linguistic apes) would it be important if *any* subject does, or can be made to do, this or that?

As to the setting: Is it my intention to predict what would happen in a real-life setting or "target" class of such settings? Our "thinking through" divides depending on the answer.

The answer may be no. Once again, we may be testing a prediction rather than making one; our theory may specify what ought to happen in *this* setting. Then the question is whether the setting gives the theory a fair hearing, and the external-validity question vanishes altogether.

Or the answer may be yes. Then we must ask, Is it therefore necessary that the setting be "representative" of the class of target settings? Is it enough that it be *a* member of that class, if it captures processes that must operate in all such settings? If the latter, perhaps it should be a "limiting case" of the settings in which the processes operate—the simplest possible one, as a psychophysics lab is intended to be. In that case, the stripped-down setting may actually *define* the class of target settings to which the findings apply, as in the dark-adaptation story. The question is only whether the setting actually preserves the processes of interest,[5] and again the issue of external validity disappears.

We may push our thinking through a step further. Suppose there are distinct differences between the research setting and the real-life target ones. We should remember to ask: So what? Will they weaken or restrict our conclusions? Or might they actually strengthen and extend them (as does the absence of power to punish in Milgram's experiments)?

Thinking through is of course another warm, fuzzy phrase, I quite agree. But I mean it to contrast with the cold creepies with which my students assault research findings: knee-jerk reactions to "artificiality"; finger-jerk pointing to "biased samples" and "unnatural settings"; and now, tongue-jerk imprecations about "external invalidity." People are already far too eager to dismiss what we have learned (even that biased sample who come to college and elect our courses!). If they do so, let it be for the right reasons.

[5] Of course, whether an artificial setting does preserve the process can be a very real question. Much controversy centers on such questions as whether the operant-conditioning chamber really captures the processes that operate in, say, the marketplace. If resolution of that issue comes, however, it will depend on whether the one setting permits successful predictions about the other. It will not come from pointing to the "unnaturalness" of the one and the "naturalness" of the other. There is no dispute about that.

REFERENCES

Argyle, M. *Social interaction.* Chicago: Atherton Press, 1969.

Babbie, E. R. *The practice of social research.* Belmont, Calif.: Wadsworth, 1975.

Bannister, D. Psychology as an exercise in paradox. *Bulletin of the British Psychological Society,* 1966, *19,* 21–26.

Bickman, L. Social roles and uniforms: Clothes make the person. *Psychology Today,* July 1974, pp. 49–51.

Brown, R., & Hanlon, C. Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley, 1970.

Brunswik, E. Representative design and probabilistic theory in a functional psychology. *Psychological Review,* 1955, *62,* 193–217.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1967.

Cappell, H., & Herman, C. P. Alcohol and tension reduction: A review. *Quarterly Journal of Studies on Alcohol,* 1972, *33,* 33–64.

Deese, J. *Psychology as science and art.* New York: Harcourt Brace Jovanovich, 1972.

Hecht, S. Vision II: The nature of the photoreceptor process. In C. Murchison (Ed.), *Handbook of general experimental psychology.* Worcester, Mass.: Clark University Press, 1934.

Henshel, R. L. The purposes of laboratory experimentation and the virtues of deliberate artificiality. *Journal of Experimental Social Psychology,* 1980, *16,* 466–478.

Higgins, R. L., & Marlatt, G. A. Effects of anxiety arousal on the consumption of alcohol by alcoholics and social drinkers. *Journal of Consulting and Clinical Psychology,* 1973, *41,* 426–433.

Kantowitz, B. H., & Roediger, H. L., III. *Experimental psychology.* Chicago: Rand McNally, 1978.

Mello, N. K., & Mendelson, J. H. Alcohol and human behavior. In L. L. Iverson, S. D. Iverson, & S. H. Snyder (Eds.), *Handbook of psychopharmacology: Vol. 12. Drugs of abuse.* New York: Plenum Press, 1978.

Milgram, S. *Obedience to authority.* New York: Harper & Row, 1974.

Mook, D. G. *Psychological research: Strategy and tactics.* New York: Harper & Row, 1982.

Nisbett, R. E., & Ross, L. *Human inference: Strategies and shortcomings in social judgment.* New York: Century, 1980.

Orne, M. T., & Evans, T. J. Social control in the psychological experiment: Anti-social behavior and hypnosis. *Journal of Personality and Social Psychology,* 1965, *1,* 189–200.

VI

SAMPLING AND MEASUREMENT ISSUES IN EVALUATION

Once the evaluation design has been specified, an evaluator's attention turns to considerations of the appropriate type of sampling strategy and measures to employ. During this stage, the evaluator asks questions about the appropriateness of the target population selected, the level at which the target should be specified (e.g., individual, group, community, social system), and the ability of measures to detect true changes in the target as a result of the intervention. Serious problems with reliability and validity in an evaluation could arise if incorrect sampling strategies are employed or if the measurement of key variables is not sensitive to change.

The four articles included here highlight the importance of carefully considering sampling issues before evaluation studies are conducted. The articles cover purposive sampling strategies, selectivity problems in quasi-experimental studies, sample selection bias, and measurement sensitivity. In many of the evaluation studies we reviewed, we found that the issues raised by these authors were not fully recognized or implemented by evaluators. These articles may help to alert the evaluation community to the importance of these sampling and measurement issues.

St.Pierre and Cook discuss the trade-offs of utilizing various sampling strategies for descriptive and impact assessment evaluations. Specifically, they address ways to select entities (e.g., states, school districts) in evaluations of national social programs. They conclude that, while it is desirable and theoretically possible to use a random selection of sites for impact assessments, it is usually unreasonable to do so because of practical constraints. They advocate purposive sampling of sites since this method is most realistic to implement and is able to provide useful impact evaluation data. Four types of purposive sampling strategies are reviewed: (1) sampling for heterogeneity, (2) sampling modal instances, (3) sampling on implementation, and (4) sampling on the dependent variable. They end their chapter by providing examples of purposive sampling strategies utilized in two national social program evaluations.

hen and Jöreskog describe problems with selectivity when trying to
e population parameters from nonrandom samples. If a selective sam-
nalyzed as if it were random, the results are likely to be biased and
inconsistent estimates of population parameters. Nonrandom samples
lt from sample selection, self-selection, or attrition. The authors sug-
: if the sample is nonrandom, efforts should be made to model the

selective sampling that occurred. By so doing, evaluators can generate unbiased estimates of treatment effects, even without randomization.

Berk reviews recent advances in diagnosing and correcting sample selection bias. An example of estimating this bias is provided from a study of citizen opinions of the criminal justice system. Berk notes that sample selection bias is a problem whenever researchers work with nonrandom samples, because it threatens both external and internal validity.

The final article by Lipsey addresses measurement sensitivity in program evaluation. He discusses the advantages of conducting preliminary measurement assessments that include the ability to determine if the study measures are adequate for detecting the expected treatment effect and the ability to modify the study design to reduce deficiencies in measurement sensitivity. Lipsey proposes a scheme that utilizes a components-of-variance approach permitting the examination of measurement sensitivity to criterion-level effects, reliability and generalizability, and statistical power. Examples from evaluations of a juvenile delinquency prevention program and a special education program illustrate the application of measurement sensitivity assessments.

25

Sampling Strategy in the Design of Program Evaluations

Robert G. St.Pierre and Thomas D. Cook

This paper addresses the issue of how to select entities (e.g., school districts, states, community mental health centers, Head Start centers) to be studied in evaluations of national social programs. It distinguishes between two major uses of evaluation—description and impact assessment—and argues that while the selection of a random sample of entities is appropriate for providing descriptive information, practical constraints reduce the utility of random sampling for impact evaluations. We propose purposive sampling of entities for impact evaluations as an alternative strategy, define different types of purposive samples and their uses, and give examples of implementations of purposive sampling in impact evaluations of national social programs.

DESCRIPTIVE VERSUS IMPACT EVALUATIONS

Several sources give extended definitions of evaluation, complete with typologies, classifications, and categories (e.g., Rossi, Freeman, & Wright, 1979). For our purposes we only need distinguish between two general types of evaluations: (1) those that provide descriptive information on programs and (2) those that provide information on program effectiveness. As shown in Table 1, many typical evaluation questions fall into the descriptive category, including those dealing with the use of program funds, targeting of program benefits, and program implementation. In the effectiveness category we consider studies involving the assessment of program costs and/or impacts.

From Robert G. St.Pierre and Thomas D. Cook, "Sampling Strategy in the Design of Program Evaluations," original manuscript.

Authors' Note: The two evaluations used as examples were funded by the U.S. Department of Agriculture, Food and Nutrition Service, Office of Policy, Planning and Evaluation under contracts FNS-53-3198-9-38 and FNS-53-3198-1-112 with Abt Associates Inc. Special thanks are due to Mr. Michael Puma, Dr. Jack Radzikowski, and Dr. Victor Rezmovic of FNS/OPPE for their assistance throughout these projects. This paper was presented at the joint meeting of the Evaluation Network and the Evaluation Research Society, Chicago, October 1983.

TABLE 1
Types of Evaluation

Nature of Evaluation	*Evaluation Question*	*Source of Data*	*Type of Sample*
Program description	Are program funds being used properly?	Review of extant records	Random
	Is the program reaching the intended beneficiaries?	Survey of program administrators	Random
	Is the program implemented as intended?	Survey of program administrators, program observation	Random or purposive
Program effectiveness	Is the program effective?	Measurement of program effects	Purposive
	Is the program efficient?	Measurement of program costs and effects	Purposive

The point to be made from Table 1 is that both the sources of data and the type of sample for an evaluation are related to the type of questions to be addressed and hence to the general nature of the evaluation. Descriptive questions can often be addressed by accessing existing data, such as financial records or program applications. Descriptive data on program operations are also obtained by surveys of program administrators or program participants. The reason that we typically obtain descriptive information using such techniques is that we assume the needed information is in the knowledge base of the relevant respondent. For example, in a survey designed to obtain information on program operations we assume that the respondent, most likely a local program administrator, knows how the program has been operating and can report this information in a reasonably reliable and valid manner. Similarly, when we access program records an assumption is made about their reliability and validity (though such assumptions have been questioned, e.g., by Cochran [1978]).

On the other hand, questions about program effects entail gathering data from different sources using different methods. While it is sometimes possible to obtain information on program effects from extant records (e.g., records of fuel use in evaluations of fuel conservation programs), the more usual circumstance calls for direct measurements of impacts on program participants. We would not expect to receive reliable and valid data on impacts by asking program administrators their opinions about the program—such information is not in their knowledge base.

RANDOM VERSUS PURPOSIVE SAMPLES

Now, the question arises as to what type of sample is most appropriate for gathering data in descriptive versus impact studies. In the former case it is often possible to collect information from a random sample of sites imple-

menting the program, or even to conduct a census of all sites.[1] Why is this possible?

(1) Obtaining data on a random sample of sites participating in a program allows generalization of the findings to all sites in the program. Therefore it is desirable to collect data using a random sample.
(2) The sample frame is generally available from extant records. Information on the names and addresses of participating sites, as well as at least some appropriate stratifying variables such as region of country or program size are typically available.
(3) The response burden of descriptive studies is relatively low, involving the completion of a questionnaire by one or at most a few respondents in each site, transcribing numbers from existing files, and so on.
(4) In a related area descriptive studies are relatively inexpensive, as they ususally involve a cross-sectional rather than a logitudinal design.
(5) The logistics of descriptive studies are relatively simple. Gaining the cooperation of respondents does not pose great problems since response burden is low, and participating sites do not have to enter into any long-term effort. Rather, they supply information on a one-time basis.

Hence, many of the logistical problems involved in conducting field experiments such as recruiting and maintaining sites, travel to sites, on-site data collection, and hiring on-site staff are nonexistent in descriptive studies.

The above examination of several practical considerations leads us to conclude that selection of a random sample for conducting a descriptive evaluation makes sense. It is *desirable,* because it leads to an enhanced ability to generalize. It is *possible,* because the sample frame exists. And it is *reasonable,* because response burden, expenses, and logistical difficulties are all low. This discussion should not be taken to imply that descriptive studies are simple to conduct, are inexpensive, or impose low response burden in an absolute sense. Rather, they have these characteristics *relative to* the second type of evaluation study we define—impact evaluations of program effectiveness.

Our thesis is that for conducting impact evaluations it is *desirable* to use a random sample of sites, and it is theoretically *possible.* However, several considerations act to make it *unreasonable.* First, it is desirable to conduct an impact evaluation in a random sample of program sites for the same reason that it is desirable to conduct a descriptive evaluation in a random sample of sites: It allows generalization of the results of the evaluation to the population of participating sites. Second, it is theoretically possible to conduct impact evaluations using randomly drawn sites. The real problems arise when considering the practical constraints imposed by impact evaluation, and whether it is reasonable to expect to conduct an assessment of program impact using a random sample of sites.

Considering the same practical issues raised earlier with respect to descriptive studies we find many problems. The sample frame for the evaluation may

or may not be available. Certainly there should be a list of all program sites, along with a few potential stratification variables. However, while this type of sample frame was appropriate for a descriptive evaluation, it may not be adequate for an impact evaluation where it may be important to stratify on a different set of variables, such as perceived program success. More will be made of this point later. At this time it is sufficient to say that the sample frame for a descriptive study may not be adquate for an impact study.

Next, the response burden for impact evaluation is high. The evaluator cannot simply ask a program administrator to supply opinions about program effectiveness and expect to receive valid, reliable, unbiased data. Instead, a research design must be conceived and direct measurements must usually be made on program participants at two more points in time. The data collection could involve administration of tests of some sort, face-to-face interviews, or other time consuming efforts. Therefore, while in a descriptive study the response burden may be limited to a one-time questionnaire to be completed by one or two respondents per site, an impact evaluation may involve two or more administrations of a test battery on many—perhaps 100 or more—program participants in each site. Other data on program implementation are often collected in impact assessments via face-to-face interviews or mail questionnaires for program implementors. Thus the response burden for a site participating in an impact evaluation is high.

This has clear implications for the cost of impact evaluation; they are expensive. In addition to the costs of data collection, which may involve hiring on-site staff or having evaluation staff travel from site to site in order to administer tests, conduct observations, or conduct interviews, there is the cost of designing the evaluation, negotiating with sites for their participation, maintaining good relationships with sites, and a host of related problems. These tasks tall add to the expense of impact evaluations. The same practical problems exist with respect to logistics. An impact evaluation involves considerable negotiation with sites, training sessions for site personnel, careful planning so that data collection can be carried out successfully, and a myriad similar details that are not as cumbersome in descriptive studies.

Finally, the main methodological argument for conducting an impact evaluation in a random sample of sites can be questioned. Random sampling of sites increases our confidence in generalizing the results of the evaluation to other sites, yet the burden involved in impact studies is great, and such studies are always subject to selection bias. Since it is not possible to compel sites to participate in an evaluation and since participation usually involves a fair amount of effort, sites that volunteer for impact evaluations may differ in important respects from the population of sites.

The implication of this discussion is that, while random sampling may be the best theoretical method of selecting sites for an evaluation, practical constraints mitigate against using it in impact evaluations. We have to ask ourselves whether random sampling is a viable method for obtaining information

on outcomes, or whether the expense, the volunteer bias, and logistical problems make other sampling mechanisms more attractive. Our contention is that, in order to provide information on program impact, it is appropriate to select a purposive sample of sites. With this thesis in mind, we now take a closer look at the rationale for using purposive samples in impact evaluations.

RATIONALE FOR PURPOSIVE SAMPLING

First, we assume that the evaluator is faced with the challenge of assessing the impacts of a national social program. By this we mean a federally sponsored program in which the funds and associated legislation are targeted toward the solution of a social problem. The evaluation of federally funded social programs presents some special problems because of the wide variability in activities conducted at the local level. The site-specific heterogeneity of social programs has been often documented (e.g., Berman & McLaughlin, 1978; Stebbins, St.Pierre, Proper, Anderson, & Cerva, 1978), leading to the conclusion that federal social programs tend to be funding mechanisms for allowing program participants at the local level to implement their own ideas, subject to broad constraints (Cook, 1982). A second assumption is that funds for the evaluation are limited. Though not a difficult assumption to support, the importance of recognizing limited funding in the present context highlights the fact that evaluators rarely include all local sites in a national impact evaluation. Some sampling of sites is necessary to remain within the evaluation budget and to make the evaluation feasible.

Having a rationale for sampling local sites is necessitated by another characteristic of impact evaluations—that "site-intensive evaluations with multiple indicators of success are superior to evaluations with more perfunctory site-level knowledge" (Cook, 1982). Achieving a balance between the size of the sample (number of local projects) and the amount of information collected at each site is a basic trade-off that must be made in all impact evaluations, and some of the past difficulties of large-scale evaluations are attributable to the fact that the trade-off has often favored sample size. In striving for generalizable results, evaluators have often included many local sites, and in striving to stay within budget have gathered minimal information from each. Because in an impact evaluation it is essential to collect data on program impacts, because most educational programs hypothesize a range of impacts, and because the resources available for data collection are constrained by the sample size, it is usually the most important or policy-relevant outcomes that are measured, while lesser outcomes and data on the process or implementation of the program are omitted. This leaves the evaluators in the position of measuring only a subset of the hypothesized program outcomes and, further, of not being able to explain or account for any differences in the effectiveness of local sites. The trade-off made here is between limited information on many

sites and a greater amount of information on fewer sites. In the former case it is likely that the results will be more generalizable, but in the latter it is likely that the results will be more detailed and explainable. The trade-off is thus between generalizability and depth or quality of information.

The problem, then, is that the evaluator is faced with a program that varies tremendously in the activities implemented from site to site, with limited funding, and with a mandate to conduct an impact evaluation. This would not be so troubling if impact evaluations were used to make "go/no go" decisions on programs. If this were the case, and if overall program effectiveness on one or two key outcomes were necessary for continued funding, then the omission of process information would be reasonable. If the client demands statistical generalizability, is most interested in an assessment of average or overall impact, and will be satisfied with limited information from each sites, then an assessment with a limited measurement battery is reasonable.

This is rarely the case, however. Programs live or die on the basis of a multitude of factors, and evaluative information seldom plays a critical role in overall funding decisions. If evaluations are not used to fund or de-fund programs, what are they used for? In recent documented cases of evaluation use (e.g., Boruch & Cordray, 1980), evaluations of national educational programs are used to provide information on how programs can be changed, improved, expanded, or disseminated. This type of use argues strongly for the collection of in-depth data at the site level, to allow the evaluators to examine multiple outcomes, to draw conclusions about why certain local sites were more or less effective than others, and to provide information that can be disseminated about successful sites. The solution we are proposing is, in short, to restrict the evaluation to an in-depth investigation in a limited number of sites. With the intent to select a purposive sample, difficult questions arise, e.g., What local projects should be included in the evlaution? and, How should they be selected?

SELECTING A PURPOSIVE SAMPLE

It is possible to select a variety of purposive samples. Here we describe some classes of purposive samples. The starting position in each case is that the evaluator has a target population and wants to generalize to that population. The model of representative sampling that allows formal generalization has been ruled out, and so when we speak of selecting purposive samples it is with the intent of making generalizability as strong as possible, knowing that it will not be as strong as if we had been able to draw a random sample. Four types of purposive sampling will be defined: sampling for heterogeneity, sampling modal instances, sampling on implementation, and sampling on the dependent variable.

Sampling for Heterogeneity

It is often desirable to select a purposive sample in order to ensure variability on one or more key stratifiers. For example, an evaluation of bilingual education programs might well call for the selection of sites serving different ethnic groups in order to determine if similar impacts are found across the range of this key variable. Cook and Campbell (1979) caution that evaluators take care not to focus only on the extremes when sampling for heterogeneity. In an evaluation where stratifying on city size is important it is natural to select some very large and very small cities. This should not blind us to the selection of medium-sized cities, which may in fact represent the typical instance.

Sampling Modal Instances

Rather than sampling for heterogeneity, it is also possible to concentrate on modal instances. The idea here is to define the variable(s) across which one wants to generalize and select a site or sites at the mode of each. This strategy assumes that information on sites at the extremes of a distribution is less important than information on the most typical sites—those at the mode.

Sampling on Implementation

A third strategy calls for the purposive selection of sites based on one or more variables related to implementation. Here it would be possible to select sites to represent various degrees of implementation in order to test the robustness of a program—how well it works at different levels of implementation. Alternatively, an evaluation could focus specifically on sites where the program is particularly well implemented or poorly implemented. In a related area, site selection could also be made on the basis of the transferability, or transportability of a program. Selection of programs that can be transported from site to site is particularly critical if dissemination of the program is one of the potential uses of the evaluation. A third implementation variable useful for sampling purposes is the administrative approach used by a program. For example, does the program provide a complete set of materials and instructions for use of the materials, or does it consist of a set of general objectives and rely on local-level program practitioners to decide exactly which activities to implement.

Sampling on the Dependent Variable

A final option is to select sites that are deemed ahead of time to be successful according to some criterion. Studying such "potentially successful" sites would give a program a chance to show that it can work, and would generate information on successful practices for adoption by others. This approach is particularly useful if the evaluation is exploratory in nature and is interested in looking for sites where the program makes a big difference. Two problems exist with this strategy. First, it capitalizes on chance. Hence it is particularly

useful to include replications when sampling exemplary instances. Second, the program is given a good chance at demonstrating success, and the evaluators are in a strong position to investigate the determinants of program success, but in such a study it is difficult to demonstrate a causal relationship between the treatment and observed outcomes.

In the remainder of this chapter we illustrate the sampling strategies described above with examples from large-scale social program evaluations.

SAMPLING ON THE DEPENDENT VARIABLE AND SAMPLING ON IMPLEMENTATION

The first illustration we present is based on the national evaluation of the Nutrition Education and Training (NET) program, funded by the U.S. Department of Agriculture. This evaluation combined two types of samples: a nationally representative random sample of NET projects in order to obtain descriptive information on the program, and a purposive sample of NET sites in which to conduct an impact evaluation. Further, the purposive sample was selected by sampling both on the dependent variable and on implementation characteristics.

The NET Evaluation

NET is a school-based nutrition education program. Its major target group is children in grades K–6, and it is intended to affect children's nutrition-related knowledge, attitudes, and behaviors. The program is administered via entitlement grants to states, which are responsible for hiring a NET state coordinator, deciding on a particular "model" for nutrition education within the state, and dispensing NET funds to local projects, typically in the public schools. Local projects are responsible for the actual implementation of NET activities.

The national NET evaluation was initiated in June 1979 and was completed in May 1981. When the evaluation began, little information was available on how the program was operating at the state and local levels, who was being served, how funds were being used, and impacts the program was having. Against the backdrop of upcoming reauthorization hearings the U.S. Department of Agriculture contracted for an evaluation of program operations and impacts (St.Pierre & Rezmovic, 1982).

The evaluation was motivated by several specific information needs. Since NET was newly implemented, federal program planners and decision makers needed to know what activities were being conducted nationally, how the program was being administered, how resources were being allocated, whether local projects were targeting the goals expressed by the NET legislation, and what obstacles existed to program implementation. According to the evaluation Request for Proposals (Food and Nutrition Service, 1979), findings in these areas would be used for "budget discussions with Congress and to

develop recommendations to the Administration on program continuation." (pp. 43–44).

In addition to this descriptive information, data were needed on the NET program's impacts on children's knowledge, attitudes, and eating behaviors. Federal program administrators especially wanted data on "potentially successful models of nutrition education strategies appropriate to local conditions" (FNS, 1979, p. 43). Again, NET was a new program, and anecdotal evidence indicated that some states were having problems, both in preparing state-level programs from scratch and in selecting from among previously developed programs. Thus, information on successful, transferable nutrition education programs was of primary interest at the federal level.

The Request for Proposals stated that federal staff would "use the study's products, especially those relating to successful program and project strategies for the purpose of program improvement and dissemination" (p. 44). The needs of state paralleled federal needs closely. States needed assistance in deciding what program/type of program to adopt. Some states had tentatively adopted one program but were not wedded to that choice. And finally, some states were leaving the choice of what nutrition education activities to implement up to local discretion. In these cases local personnel were in similar need of information on well thought-out programs that could be transported from place to place and that had some evidence on success in producing positive effects on knowledge. Thus, the situation was ripe for the evaluation of "exemplary" nutrition education programs.

Sample Selection Rationale

With these needs in mind, the evaluation was designed to provide descriptive information as well as data on program impacts. The descriptive part of the evaluation was straightforward and consisted of a review of the literature (Nestor & Glotzer, 1981), as well as analyses of annual plans submitted by NET state coordinators, and surveys completed by all state coordinators and a nationally representative sample of local project directors (Ferb, Glotzer, Nestor, & Napior, 1980). The objectives of the descriptive study were to provide a national picture of NET operations at the state and local levels. A randomly selected national sample of local sites was appropriate for providing this type of information, which could be collected relatively inexpensively via a structured mail survey.

In addition to descriptive information on the status of NET, the client wanted to assess program impact and provide to states information on potentially successful models of nutrition education. Therefore, the impact evaluation called for the selection of entities for assessment. The sample selection process involved making a series of decisions about the entities to be sampled. First, the decision was made not to evaluate a nationally representative set of local- or state-level projects. The rationale was that NET was a new program when the evaluation was initiated in 1979 and since implementation was incomplete, a comprehensive impact evaluation would have been unfair.

Further, resources did not permit a full-scale national assessment of program impact.

Second, the decision was made to select states, rather than local projects, as the basic entities for evaluation. As noted above, states were given the responsiblity for hiring a NET state coordinator who would develop a statewide nutrition education program. While there was no national prescription for how nutrition education was to be conducted, it was reasonable to expect some homogeneity in the activities implemented within each state, since local projects were supposed to be conducted in accordance with the state-level mandate.

Having decided not to conduct a national impact evaluation and to sample states rather than local projects as the basic entity to be studied, the issue became one of how to select the states. The general approach was to select states by sampling on the dependent variable and on implementation. Multiple sampling criteria were used. First, the program had to be *well developed.* Impact evaluations are not usually called for at the beginning of a program because program implementation is likely to be incomplete. In fact, one of the criticisms of the evaluation from state coordinators and federal program administrators was that the evaluation was premature. In an attempt to avert the problem of evaluating a nonexistent program, we opted to select state-level programs that had been developed prior to NET—ones that were well thought through and mature enough to be well implemented.

Second, the program had to have some *evidence of success* in terms of increasing children's nutrition-related knowledge. This criterion was employed because the evaluation client needed information on successful programs for dissemination to states. Using evidence of success as one criterion for selecting states for evaluation maximizes the chances of detecting positive effects and minimizes the chances of "washing out" positive effects by averaging them with negative ones. If no positive effects are found under these conditions, it is safe to say that the program will not be successful under less favorable circumstances. On the other hand, finding that a program demonstrates success when well implemented enables policymakers and program practitioners to concentrate on improving the program, or ensuring its faithful implementation, or on disseminating the tested successful versions.

Third, the program had to have *evidence of being transferable.* As noted earlier, the evaluation client hoped to disseminate information on program success to state coordinators who needed help in deciding on an approach to nutrition education. If this need was to be met it follows that the programs to be disseminated should be capable of being implemented with a minimum of problems in other locations. This is a difficult condition to impose, as the literature on knowledge diffusion is rife with instances of programs that, for one reason or another (e.g., existence of a charismatic leader, lack of proper materials or program documentation) are not transferable from one location to another. In spite of this problem, transferability is an absolutely key sam-

pling criterion since attempts at program dissemination will founder if success depends on too many characteristics that are unique to a given setting.

Fourth, the program had to be *relatively inexpensive.* This condition may appear self-evident; however, there is a range in the costs of nutrition education programs, and in an era when schools are beset with financial problems and are being pushed by the public to go "back to the basics", a nutrition education program that imposes anything other than a very minimal cost is a candiate for extinction.

Fifth, programs were selected that were maximally different from each other in terms of their *administrative approach* to the delivery of services. The intent of this criterion was to obtain some evidence as to the range of programs over which nutrition education can be expected to work. In this regard, the descriptive portion of the evaluation was used to distinguish among three administrative approaches for delivering NET services from the state level:

(a) a centralized model, under which states provide a "packaged" curriculum to be used in all particiating school districts in the state;
(b) a decentralized model, where states provide guidance, training, resources, and broad guidelines for nutrition education, but where local projects are responsible for deciding exactly which nutrition education activities to implement; and
(c) a regional model, where nutrition information and training is provided to local projects by multiple resource centers, often located within universities.

Finally, in order to be considered for inclusion in the sample, states had to be *willing to cooperate* with the evaluation. This was not the case in all instances because, as noted earlier, some state coordinators were critical of the evaluation as being premature.

To sum up, the sampling criteria called for the selection of mature, potentially successful, transferable, reasonably inexpensive state-level models of nutrition education that differed from one another in terms of administrative approach and that had state coordinators who were willing to participate in the evaluation. At the beginning of the evaluation it was not clear that any programs could be found that would meet all these criteria.

Sample Selection Procedures

The decision to draw a purposive rather than a random sample dictates a great deal about sample selection procedures. Specifically, the sample cannot be drawn by random sampling from a list of program participants in this case, a list of states. Rather, the sample must be selected on the basis of in-depth knowledge of the distribution of state programs on the sampling criteria defined earlier. This means the evaluation team had to learn about the content of state-level programs, their implementation status, transferability, evidence of success, costs, and willingness to participate in the evaluation.

There are two ways to obtain such evidence. First, it might be possible to gather quantitative data for some of the criteria. In the present case, evidence on preliminary program success was obtained by reviewing the results of evaluations already conducted. If time was available a preferred strategy might involve the conduct of short-term studies. In fact, the evaluation Request for Proposals mandated the development of "a data base which summarizes the status of NET funded and other pertinent nutrition education and training activities in the U.S.," in order to "identify from the data base a set of potentially successful models of nutrition education strategies." (FNS, 1979, p. 43).

An alternative strategy, which was also employed in the present study, is to use qualitative data from a variety of sources in order to identify programs to include in the sample. The idea here is to solicit testimonials from professionals in the field, to review relevant literature, and to ask for self-nominations, all in the service of arriving at convergent perspectives from different sources on the selection of the sample in a relatively short time frame.

The actual selection process used in the NET evaluation involved making a major presentation at a national meeting of NET state coordinators. This presentation explained the purpose of the evaluation to the state coordinators and solicited nominations of exemplary programs from state coordinators, federal officials, and others who attended the meeting. Subsequent to the meeting, additional and confirmatory nominations were obtained through phone calls to nutrition education professionals, through making inquiries of the advisory panel for the evaluation, and through a review of research literature on the effects of nutrition education programs. As a result of these efforts several states were nominated as having programs that might fit the needs of the evaluation, including California, Connecticut, Georgia, Nebraska, Pennsylvania, and West Virginia.

The next step entailed in-person visits to each of the state coordinators in these states to explain further the purpose of the study and to obtain a first hand view of the program. The on-site visits involved collecting materials explaining the nature of the program, discussing any ongoing plans for evaluation, assessing the implementation status of the program, including finding out exactly where the program had been implemented and what expansion was planned, and, finally, discussing the willingness of the state coordinator to participate in the evaluation.

Armed with as much evidence as we could obtain from as many sources as possible on program implementation, success, willingness to cooperate, and so on, a sample selection meeting was held. Meeting attendees included representatives of the evaluation contractor, the client, and the evaluation's advisory panel. From this meeting emerged a consensus to select two states for evaluation: Nebraska and Georgia.

SAMPLING FOR HETEROGENEITY AND SAMPLING MODAL INSTANCES

Our second example illustrates the use of two different purposive sampling procedures. Funded by the Food and Nutrition Service (FNS) in the U.S. Department of Agriculture, this evaluation of an ongoing demonstration program is concerned with estimating the effects of two alternatives to the donation of agricultural commodities to schools participating in the National School Lunch Program. Many of the design parameters in this congressionally mandated evaluation were set by legislation. Most relevant to the present discussion is the fact that Congress called for the evaluation to test the effects of three treatments (the current program and two alternatives) in 30 sites each, for a total of 90 sites. In this case Congress intended that sites be school districts. Congress also included specifications on how the sample was to be selected. In particular, the following somewhat contradictory stipulations were made (FNS, 1981, p. 67):

> The school districts shall be selected by stratified random sample to represent a nationwide variety.
>
> The Secretary shall allow school districts not less than 45 days or more than 60 days from the date of publication of a notice in which to apply for participation in pilot projects."

In-house evaluation staff at FNS were given the responsiblity for selecting the sample, and these statements were interpreted by FNS as meaning that a random sample was desired but that some volunteer sites should also be included in the evaluation. To meet congressional specifications and in order to assure three comparable groups of sites, FNS adopted a sampling strategy that first involved the purposive selection of a heterogeneous group of states, then called for construction of a pool of matched eligible sites representing the modal site in each state, and finally called for the random selection of a "triplet" of matched sites from each state. The three matched sites in each state were then randomly assigned to the three treatment groups (St.Pierre et al., 1981).

Heterogeneous Sampling of States

In order to achieve the widest possible geographic representation of sites, FNS decided to conduct the evaluation in many, rather than few, states. The general strategy was to purposively select 27 states, such that at least three were from each of the seven FNS geographic regions, and such that the states exhibited wide variation on variables expected to have an impact on the outcomes including degree of access to agricultural markets, prominence of agriculture as a within-state industry, and sophistication of food-related transportation and distribution systems. The selection of 27 states is thus an example of purposive sampling for heterogeneity. Several relevant variables

across which we would like to generalize were identified, and states were selected to represent as much variation as possible in those variables.

Selection of Modal Sites within States

After the purposive selection of 27 states, the sampling strategy called for each state to be represented in the final sample by a triplet of matched sites that represented the modal site in the state. Within each of the 27 selected states, all sites with student enrollments of greater than 25,000 were pulled out and became part of a special cross-state pool of "large" sites. The remaining sites were assigned to an 18-celled sample frame formed by taking all possible combinations of the following three stratifiers:

- *Program participation*—whether the site participates in
  (1) the National School Lunch Program
  (2) the National School Lunch Program *and* the School Breakfast Program
- *Enrollment*—whether the site's student population is
  (1) less than 1,199
  (2) 1,200 to 4,999
  (3) 5,000 to 24,999
- *Poverty level*—the proportion of the site's population with incomes below the Office of Management and Budget poverty level:
  (1) less than 11.9%
  (2) 12.0%–24.9%
  (3) 25.0% and over

The program participation variable has two levels, while the poverty level and enrollment variables each have three levels, yielding a total of 18 cells (2 x 3 x 3). To give a concrete example, one specific cell defined by the intersection of the above three variables contains all sites in a state that (1) participated in the National School Lunch Program but not the Breakfast Program, (2) had enrollments of less than 1,199 students, and (3) had a population in which more than 25% of the familes had incomes below the poverty level.

Because the sampling strategy called for the selection of three sites in each state that were as similar as possible, all sites in the pool of eligibles for a given state were drawn from the same cell of the sampling frame. Further, FNS matched the 27 states with the 18 cells of the sample frame such that each cell of the frame was represented by at least one state and that each state was represented by sites from a cell that were as "typical" of all the sites in the state as possible. In order to balance these competing criteria, the following iterative approach to constructing a pool of eligible school food authorities (SFAs) in each state was used:

(1) Construct initial within-state pools of sites by choosing the largest cell within each of the 27 states (i.e., the cell containing the most sites).

The rationale is that each state was to be represented by a group of modal sites.

(2) Check the resulting within-state pools of eligibles to determine if all 18 cells are represented by at least one state.

(3) If some cells are not represented, see if they would be selected by substituting the second largest cell within states that have multiple sets of modal sites.

(4) Iterate steps (2) and (3) until all cells are represented by at least one state.

This procedure produced a pool of eligible sites within each state. However, the within-state pools did not contain any large sites (enrollments of more than 25,000 students). Rather than constructing pools of eligible large sites within a given state, FNS opted for constructing pools of eligible large sites by combining SFAs from different states. This procedure was used because the removal of three large sites from the commodity program in a single state would make it impossible for the state to continue normal, cost-effective food distribution operations.

Cross-state pools of eligible large sites were therefore formed by stratifying large sites on program participation and poverty level (creating a 6-celled frame in the same fashion as the 18-celled frame for the within-state pools) and by selecting the three highest-frequency cells of the sample frame as the three pools of eligible large sites.

At this point, six sites were randomly selected from each of the within-state pools of eligible sites and the cross-state pools of large eligible sites and were asked to participate in the demonstration. The general nature of the three treatments was described, but no assignment to treatment was made. Some sites declined to participate and were dropped from the sample. These refusals were followed up with a telephone call to ascertain their reasons for nonparticipation.

As noted earlier, the congressional mandate for the evaluation called for the inclusion of some volunteers in the sample. In response to a Notice of Intent to conduct the demonstration that was published in the *Federal Register,* applications were received from 194 sites. Volunteer sites were then classified into the same state-level sampling cells that were used in defining the pools of eligible sites. Any volunteer site that fell into one of the within-state or cross-state pools was automatically eligible for the demonstration, along with the six sites randomly selected from each pool.

The final pools of eligible sites (combined volunteers and randomly selected sites) within the 27 states formed the groups from which the final selection of sites was made. Since each within-state group of six randomly selected sites had been augmented with some small number of matched volunteers, and since only three sites were needed in each state, three sites were selected randomly from each within-state and cross-state pool and were randomly assigned

to one of the three treatment groups. Sites were then contacted and asked whether they were willing to accept their assigned treatment. Those who indicated "yes" were included in the demonstration sample. Those who refused were dropped from the sample, and a replacement site was randomly selected from the appropriate pool and asked to participate. This process continued until the required number of sites was obtained.

It should be noted that this process did not guarantee representation of volunteers in the final sample, since the final set of three randomly selected sites from each within-state pool of eligibles could have, by chance, omitted all the volunteers. However, the process did greatly increase the probability of selecting volunteers for the final sample *contingent on their not being atypical.* That is, volunteers were included along with the six randomly drawn sites from each within-state pool of eligibles only when they belonged to the modal cell for that state.

To sum up, the sampling strategy first involved the purposive selection of 27 states in order to obtain a heterogeneous sample that would allow generalization to the nation as a whole. States were selected to be heterogeneous in terms of geographic location, and in terms of several argiculture-related variables. The second stage in the sampling process called for the selection of modal sites in each state. While the selection of the mode for each state was done purposively, in order to ensure that the 27 state modes filled an 18-celled sampling frame, the actual selection of sites to represent the mode was done by using a combination of randomly drawn and volunteer sites.

INTERPRETATION OF FINDINGS

To this point we have distinguished between two major types of evaluation questions—those calling for descriptive information and those requiring impact information, and have argued that practical constraints make the purposive sampling of sites in impact evaluations a more viable model than random sampling. We then defined several methods of selecting a purposive sample and gave examples of how purposive sampling has been implemented in large national evaluations. Now we review the findings from the above-referenced evaluation of the NET program in order to illustrate the way in which the purposive sampling strategy helps/hinders the interpretation of results. As noted earlier, the state-level programs in Nebraska and Georgia were selected for inclusion in the NET evaluation.

Nebraska

The Nebraska NET program was selected because it is nationally recognized, was recommended by regional and national FNS staff as well as other nutrition education professionals, and it has an approach to nutrition education that involves the three major target groups of the NET legislation: teachers, food service personnel, and children. The Nebraska program is centrally administered, with all participating school districts implementing the

same curriculum. It had preliminary evidence of effectiveness as shown by an internal evaluation conducted in three school districts (Swanson Center for Nutrition, 1979). It also had evidence of transferability since at the time of its selection for this evaluation it was either adopted or being considered for adoption in seven other states and large cities. Further evidence of transferability stemmed from the nature of the Nebraska curriculum, which includes 11 packages of instruction for grades K–6. Each package includes 12–20 class hours of instruction and specifies "activities" and "steps" for the implementation of each activity. Teachers are presented with a complete curriculum and are not required to develop any of their own materials. A one-day training session in the use of the materials is provided for teachers, food service personnel, and school administrators.

In short, the program is designed to be transferred from place to place. The program's costs are reasonable, averaging $5.15 per pupil when the fixed costs of the curriculum materials and the training are amortized over a five-year period for a total of 200 pupils. Because of a desire to disseminate the program, Nebraska State Department of Education officials were eager to help plan and participate in the evaluation.

St.Pierre, Cook, and Straw (1981) assessed the degree to which the Nebraska NET program was implemented and the results it had on children's knowledge of nutrition, on their attitudes and preferences in the nutrition domain, and on their reported and behavioral nutrition habits. A design was implemented that, at the level of schools, is close to the ideal of a randomized experiment. Data were collected from over 2300 children in 96 classrooms distributed across grades 1–6 in 20 schools spanning the state of Nebraska. The participating schools were randomly selected from 98 volunteers for NET and were assigned to treatment (13 schools) or control (seven schools) status using a modified random assignment procedure.

A battery of measures was given to children on three occasions: The full battery was administered to the full sample as a pretest in February 1980 and again as a posttest in May 1980; a subsample of NET and non-NET children were followed-up in December 1980 with a subset of the measurement battery. Thus, the pre/post time period was ten weeks and the pre/follow-up time period was ten months. Questionnaires were mailed to teachers and food service managers in May and December 1980 for the purpose of estimating the degree to which they implemented the curriculum. The major findings of the evaluation can be summarized as follows:

- Teachers did, in fact, implement the curriculum; hence the evaluation provided a reasonable test of the treatment. Teachers reported that students reacted positively to the program and that the curriculum packages were achieving their objectives.
- Strong positive impacts were found on children's nutrition-related knowledge. This finding is consistent across several different curriculum-specific and standardized measures of knowledge and across grades 1–6; the

effects are large in magnitude (between .2 and .9 standard deviations); for many measures there is a positive relationship between implementation and size of the effect; and effects on nutrition knowledge are larger and more consistent in grades 4–6 than in grades 1–3 when, in fact, the curriculum in grades 4–6 is primarily knowledge-oriented and the curriculum in grades 1–3 is experience-oriented.

- Positive impacts were found in grades 1–3 on self-report measures of food preference and in terms of an increase in children's willingness to select unfamiliar fruits and vegetables when offered a choice in the school lunch line. In grades 4–6 NET children were more willing to taste foods they did not eat before the program than were non-NET children.
- No strong, program-related effects were found on measures of food attitudes, reported food habits, or overall plate waste.
- Follow-up testing showed that knowledge gains were generally maintained and that there was no evidence of "sleeper" effects on food attitudes or reported food habits.

Two other studies allow us to broaden our perspective by providing additional information on the effects of the Nebraska NET program. First, Majure (1980) reported results from a quasi-experimental evaluation of Nebraska's materials in eight states and metropolitan areas. Findings of this study indicated significant positive treatment effects on several measures, including breakfast variety, breakfast tradition, key nutrients, food safety, food advertising, and physical fitness.

Second, Crosby and Grossbart (1980) mailed questionnaires to the parents of children who participated in the Nebraska program evaluation. Parents reported positive program effects, such as NET children being more likely than their non-NET counterparts to know about nutrition and about different foods, to ask for meal items and snacks learned about in school, and to believe that a balanced diet is important. Parents also reported considerable parent/child interaction over the program. The study is flawed by a rather low 44% response rate, which could well have biased the results in favor of NET.

To sum up, these studies find generally positive effects of the Nebraska curriculum. Though the methodological flaws of the studies would render them unconvincing if taken alone, they corroborate the findings and increase our confidence in the present evaluation.

Georgia

Georgia was the second state included in the NET evaluation. As was the case in Nebraska, Georgia had a reputedly exemplary program. Selected for study because it was recommended as particularly well thought through and implemented, Georgia's program provides an important contrast with Nebraska's in that it follows a "decentralized" approach to the implementation of nutrition education typical of that used in many of the more populous states. Rather than providing a set curriculum (as does Nebraska) the Georgia

program considers individual school systems, schools within systems, teachers, and food service personnel in those schools as the key initiators and implementors of nutrition education. The state's role is to facilitate and support local efforts by providing the conceptual framework for nutrition education, goals and objectives, extensive training, resource materials, evaluation, and follow-up. Personnel within school systems are responsible for planning, organizing, and implementing nutrition education projects that meet state goals and objectives in ways that are most feasible and effective in the particular system. In this way, the Georgia model allows nutrition education projects to be tailored to the particular administrative needs of the school system and to the needs of the student population.

Once school systems become involved in the Georgia NET program they are obliged to participate in a nutrition education training workshop and a follow-up. Originally planned as a five-day workshop and a two-day follow-up, the training is also offered in an optional one-day session. The latter option resulted from resistance at the local level to the high level of start-up effort. The training is standardized and, while schools have the flexibility to implement specific nutrition education activities as they see fit, they are all trained using the same process. Though we have no direct information on program costs for Georgia, start-up expenses should be minimal, since no particular set of materials is required and costs for the one-day training are small. Clearly, costs will jump if the longer training session is selected. Costs as implemented in a particular school will also vary depending on the materials selected by the school, whether materials are developed by participating teachers, and other similar factors. Though the Georgia program does not mandate the use of any particular set of materials, it does provide a series of goals and student competencies by grade level. Further, the Georgia NET program has developed many resource materials for use at the local level. These resource materials, the training materials, and the flexible nature of program implementation provide evidence of the Georgia program's transportability. Finally, the Georgia State Department of Education was eager to help plan and participate in the evaluation.

A note of caution is in order here regarding the transferability of the Georgia program. Though the administrative structure of the program makes it adaptable to a variety of settings, and though the resource and training materials can be transferred from place to place, the transportability of a program such as Georgia's is very different from the transportatability of a program such as Nebraska's. In the latter case it is possible for a school to buy a set of curriculum packages, have them delivered, and have teachers use them with relatively little training. The program is self-contained and does not impose a burden on teachers in terms of developing materials or lessons, preparing new products, or selecting from existing ones. The price that must be paid for this convenience is the up-front cost of the curriculum materials and the sense that

a school should use the curriculum as it finds it. Of course, lessons or even whole portions of the curriculum can be omitted or changed, but one of the key advantages of the Nebraska curriculum lies in its continuity, its comprehensiveness, and its ease of implementation. Once the school district starts tailoring the curriculum, some of these advantages are lost.

On the other hand, the Georgia program requires much more effort from teachers. The lengthy training session and the need to develop and/or select materials mean that the program requires a large amount of start-up energy. Further, the fact that the Georgia program provides an approach to the delivery of nutrition education activities rather than a set of curriculum materials means that the program will look very different from site to site. To speak of the "transferability" of such a program may be misleading, since what is transferred is not any specific content, but an approach to nutrition education, including goals, training, and lists of resource materials.

St.Pierre and Glotzer (1981) assessed the Georgia NET program in terms of the results it had upon children's knowledge of nutrition, upon their attitudes in the nutrition domain, and upon their reported nutrition habits. Further, the evaluation assessed the degree to which the measurement battery was relevant to the nutrition education activities taught in participating classrooms. The evaluation employed a research design involving the nonrandom selection of treatment schools that were already participating in NET and control schools that were not part of the program. Some 1400 children in grades 1–8 distributed across seven school districts were pretested and posttested with a limited measurement battery that was designed to detect general impact on nutrition knowledge, attitudes, and reported habits, rather than changes specific to the Georgia program. The following are conclusions of the evaluation:

- The program had strong positive effects on nutrition knowledge (at least in grades 1–4 and perhaps in grades 5 and 6). Effects were large, ranging from .2 to 1.3 standard deviations.
- No strong program-related impacts were found on food attitudes or self-reported food habits.
- The program is more effective with younger (grades 1 and 2) than with older (grades 3–6) children.

Some corroborating evidence for the findings of this evaluation is available from a study by Emory University (1980), which found pre/post knowledge gains for children in Georgia's NET program. While unable to stand on its own merits because of methodological problems, the Emory study does support the findings from the present evaluation.

Summary of Nutrition Education Impacts

The findings from evaluations of NET in Nebraska and Georgia are important, but since they represent only two states, and since those states were carefully selected by sampling both on the dependent variable and on imple-

TABLE 2

Summary of Findings from NET Evaluations

Evaluation	*Outcomes*			
	Knowledge	*Attitudes*	*Reported Habits and Preferences*	*Plate Waste and Other Behavioral Measures*
Nebraska (St. Pierre et al., 1981)	Positive effects on several measures in grades 1–6. Magnitude of effects ranges from .24 to .82.*	No effects (positive signs).	No effects (positive signs)** on reported habits. Positive effects on reported food preference in grades 1–3. Mixed positive effects and null (positive signs) on reported food preferences in grades 4–6.	No effects on total consumption. Positive effects in grades 1–3 in terms of willingness to select new foods in the school lunch line, and in grades 4–6 in terms of willingness to taste previously rejected foods.
Georgia (St. Pierre & Glotzer, 1981)	Positive effects on several measure in grades 1–4, no effects in grades 5–6 (positive signs), no effects in grades 7–8 (negative signs). Magnitude of effects ranges from –.31 to 1.27.	Positive effects in grades 1–2, negative in grade 3, null effects (positive signs) in grades 4–6.	Positive effects in grades 1–2, negative effects in grades 3–4, null effects (negative sign) in grades 5–6.	n.a.
California (Wolff, 1980)	Positive effects in grades 1–6. No effects (positive sign) in preschool and kindergarten. Magnitude of positive effects ranges from .25 to .50.	Positive effects in preschool through grade 2. No effects (positive and negative signs) in grades 3–6. Magnitude of effects is from .10 to .25.	n.a.	Positive effects on overall consumption. Positive effects for all food types except milk. Treatment group reduced plate waste by 25 percent (about 1.25 ounces) compared with 1 percent in the comparison group.
Pennsylvania (Shannon et al., 1981)	Positive effects in grades K–6. Magnitude of effects ranges from 1 to 3 items.	n.a.	n.a.	n.a.
West Virginia (West Virginia Department of Education, 1977)	Positive effects in grades K–3, 5–6.	Positive effects in grades 1–6, not in kindergarten	n.a.	Positive effects for 5 of 7 foods studied. Waste reduced from 4 to 19 percent for individual foods.
"Food is My Bag" (Applied Management Sciences, 1976)	Positive effects in grades K–12.	No overall effects. Some positive changes noted in K–3.	No effects.	Positive effects on meat, milk, bread at most grade levels. No overall effect.

*Magnitude of effect expressed in standard deviation units unless otherwise noted.
**Indicated direction of the difference, even if not statistically significant.

mentation, the findings are limited in generalizability. In order to enhance the external validity of the evaluation, we augmented the findings from Nebraska and Georgia by integrating results from recent evaluations of other NET programs. Therefore, the conclusions of the NET evaluation drew upon studies conducted in Nebraska, Georgia, California, Pennsylvania, and West Virginia, as well as a study conducted across five states. Table 2 summarizes findings from each of these evaluations.

It should be noted that the additional evaluations reviewed here (with the exception of the five-state study) are from states that were originally considered for inclusion in our NET evaluation. This is not an accidental occurrence. Though we looked to review impact evaluations of NET programs in all states, we found only a few; and the ones in California, Pennsylvania, and West Virginia were the only ones with available information on program effects. This holds in part because, like Nebraska and Georgia, these states had programs that were developed prior to NET and thus had the time to mount their own evaluations. In fact, one reason for excluding California from our study was that they already had their own evaluation in progress. Some other states had begun impact evaluations by the time we were reviewing evaluations but did not have published findings. Other states had done needs assessments or small pilot tests, but not formal impact evaluations.

Each of the evaluations summarized in Table 2 has its weaknesses; however, as a group, the studies yield some important evidence on the effects of nutrition education. First, it appears relatively easy to produce positive effects on *nutrition knowledge.* All six studies report positive findings on knowledge, findings that are not only statistically significant but are of large size (.24–1.27 standard deviations) for social science evaluations. This finding is corroborated by other reviews of the effects of nutrition education on knowledge (e.g., Contento, 1981). It may be that children have not had a great amount of exposure to nutrition concepts, and that learning these concepts is fun and relatively easy.

Effects on food attitudes and reported food habits are much more difficult to produce. Four studies reported some positive effects on *attitudes;* however, with the exception of the West Virginia study these varied by measure and grade. The California and Georgia studies did find positive attitude effects in grades 1 and 2, suggesting that it may be easier to alter attitudes for children in the early grades. Four of the studies included an examination of reported *food habits,* but none found any strong evidence of program effectiveness in this area. Evidence on *food preference* was supplied in only one study, where the Nebraska evaluation found a strong indication of positive effects on reported food preference and willingness to select new foods.

Plate waste in the National School Lunch Program was used as an outcome measure in four of the evaluations. Effects on plate waste tended to vary

by grade and food type, though positive findings were more evident in some studies (e.g., California and West Virginia) than in others (e.g., Nebraska). While the Nebraska evaluation did not find positive effects on overall plate waste, it did find strong evidence that NET children were more willing than comparison children to experiment with new foods.

The summary picture is, therefore, one in which positive effects on knowledge appear to be almost universal, while effects on attitudes, food preference, plate waste, and other behavioral measures are not consistent across studies and are confined to specific grade and food item combinations. These findings may make a good deal of sense considering the short-term nature of the programs. Knowledge is easily conveyed in the short term; to expect a three- or ten-week program consistently to affect behaviors that have been formed for several years is quite different.

CONCLUSIONS

Several points need to be made. First, all of the nutrition education programs evaluated and reviewed in this effort demonstrated large, statistically significant positive effects in terms of increasing nutrition-related knowledge. Thus, we have accomplished one of the evaluation objectives—to determine if NET programs can produce positive effects on knowledge. The purposive sampling strategy was important in that it enhanced our chances of finding positive results. Yet, the sampling strategy also leaves us dissatisfied. We are left with questions such as,

- Is it easy to produce gains on nutrition knowledge?
- Can almost any nutrition education program teach nutrition knowledge?
- Was the dependent variable—knowledge gains—too easy?
- Should we have selected programs on the basis of the potential for success on some other dependent variable?

Questions about the ease of achieving knowledge gains can be answered without major new expenditures, by continuing to review the results of other small-scale evaluations of nutrition education programs. With hindsight the only change that might have been made in the sampling strategy for the NET evaluation would be to ensure some heterogeneity when sampling on the dependent variable. That is, in addition to sampling states that were perceived to have exemplary programs, select at least one state perceived to have a mediocre program. This would have given us some handle on the range of effects we might expect.

The questions about whether programs should have been selected on the basis of potential for success on some other dependent variable is more difficult to answer and depends on one's beliefs about the causal relationships

among the many potential outcomes of a nutrition education program. The cognitive theory underlying many education programs posits that children first need to learn new information, which will then affect their beliefs and feelings about nutrition-related behaviors in questions. Finally, long-term effects might be seen on health status (Zeitlan and Formacion, 1981).

In this causal chain, changes in early or "proximal" outcomes such as nutrition-related knowledge will lead to changes in later or "distal" outcomes, such as dietary behavior or even health status. The question is, Would it have made more sense to sample NET programs based on their potential for success on an outcome more distal than knowledge? We believe not. First, at the start of the evaluation it was not at all clear that many nutrition education programs could produce knowledge gains. That was a hypothesis to be tested. Second, if we had selected programs on a more distal variable, and had not found any of them to be effective on that outcome, the validity of the evaluation would have been questioned. Even with hindsight we would not choose a different outcome variable for sampling purposes.

A second major point to make about the utility of the purposive sampling procedure is that the NET evaluation did not yield overall summative information. As noted above, we intentionally avoided the selection of a sample that would have yielded a range of results; the evaluation focused on a subset of presumably successful programs. Though we are left somewhat dissatisfied in terms of summative findings on the program as a whole, our contention from the beginning of this paper has been that field-based evaluations cannot be all things to all people, and that trade-offs must be made in order to ensure that the most important questions are addressed adequately. In the NET evaluation the most important questions dealt with documentation of potentially successful programs that could be dissemenated to other states. The fact that other questions about national program impact were not answered in the evaluation is lamentable but unavoidable, given resource levels and related practical constraints. Those other questions will have to be addressed in other evaluations.

The final point we make is to point out that, in spite of the preceding discussion about the NET evaluation, the use of purposive sampling does not automatically preclude an evaluation from addressing national summative questions. Consider the second example presented in this paper—the evaluation of alternatives to donating agricultural commodities in the National School Lunch Program. Here, the questions of interest are national questions, and it is important to produce data that can be weighted to the national level. Given the mandated sample size of 30 sites in each of three treatment groups, random selection from the population of 15,000 or so school districts could well have yielded a poor sample. Thus purposive sampling was used to enhance the quality of the sample by ensuring heterogeneity on some variables (e.g., geographic location) while selecting modal sites to represent each sampled state. Because only one site represents each treatment in each state, the sample will not allow inferences to be made at any disaggregated level

such as within states; however, the sample will allow better inferences at the national level than would a simple random sample (though a sample of 30 sites in each treatment is meager, no matter how it is selected).

Three matched sites were randomly selected from each of 27 randomly assigned to treatment groups. This process provides comparable treatment groups and enhances the internal validity of the evaluation. Since the three sites selected in each state were "modal" sites, and since states were selected in order to maximize the heterogeneity of the sample in terms of geographic and agriuculture-related variables, the external validity of the study is also maximized.

In short, the samples for the two evaluations we have been discussing were designed to allow very different types of statements to be made. Though both are impact evaluations, one focused on the evaluation of well-implemented exemplary instances, while the other is targeted to the provision of national summative data. The point with which we wish to leave the reader is that purposive sampling can help in both situations.

NOTE

1. Throughout this paper we will use "sites" to represent the local implementation of a program. A site could be a school, a school district, a Head Start Center, a Community Mental Health Center, and so on.

REFERENCES

Applied Management Sciences. (1976). *Evaluation of a comprehensive nutrition education curriculum.* Silver Springs, MD: Author.

Berman, P. & McLaughlin, M. W. (1978). *Federal programs supporting educational change, volume 8: Implementing and sustaining innovations.* Santa Monica, CA: Rand Corporation.

Boruch, R. F. & Cordray, D. S. (1980). *An appraisal of educational program evaluations: Federal, state and local agencies.* Evanston, IL: Northwestern University, Department of Psychology.

Cochran, N. (1978). "Granda Moses and the 'corruption' of data." *Evaluation Quarterly,* 2(1), 363–374.

Cook, T. D. (1982). "An evolutionary perspective on a dilemma in the evaluation of ongoing social programs." In M. B. Brewer and B. E. Collins (Eds.), *Knowing and validity: A tribute to Donald T. Campbell.* San Francisco: Jossey-Bass.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Contento, I. (1981). "Kindergarten through sixth grade nutrition education." In J. Nestor and J. Glotzer (Eds.), *Teaching nutrition.* Cambridge, MA: Abt Books.

Crosby, L., & Grosbart, S. (1980). "Memorandum to Glenda Uhrmacher." Lincoln: University of Nebraska, College of Business Administration.

Emory University. (1980). *Georgia education model for nutrition education and management training: Final report.* Atlanta, GA: Emory University School of Medicine, Program in Dietetics, Division of Allied Health Professions.

Ferb, T., Glotzer, J., Nestor, J., & Napior, D. (1980). *The Nutrition Education and Training Program: A status report, 1977–1980.* Cambridge, MA: Abt Associates Inc.

Food and Nutrition Service, U.S. Department of Agriculture. (1979). *Request for proposals for an evaluation of the Nutrition Education and Training Program.* Washington, DC: Author.

Food and Nutrition Service, U.S. Department of Agriculture. (1981). *Request for proposals for an evaluation of alternatives to the National School Lunch Program's commodity donation system.* Washington, DC: Author.

Majure, W. (1980). *Evaluation report: Eight-state consortium.* Red Oak, IA: Experience Education.

Nestor, J. & Glotzer, J. (Eds.). (1981). *Teaching nutrition.* Cambridge, MA: Abt Books.

Rossi, P. H., Freeman, H. E. & Wright, S. (1979). *Evaluation: A systematic approach.* Beverly Hills, CA: Sage Publications.

Shannon, B., Bell, P., Marbach, E., Hsu-O'Connell, L., Graves, K., & Nicely, R. (1981). "A K–6 nutrition curriculum evaluation: Instruction and teacher preparation." *Journal of Nutrition Education,* **13**(1), 9–13.

St.Pierre, R. G., Cook, T. D., & Straw, R. B. (1981). "An evaluation of the Nutrition Education and Training Program: Findings from Nebraska." *Evaluation and Program Planning,* **4**(3–4), 335–344.

St.Pierre, R. G. & Glotzer, J. (1981). *An evaluation of the Georgia Nutrition Education and Training Program.* Cambridge, MA: Abt Associates Inc.

St.Pierre, R. G. & Rezmovic, V. (1982). "Findings from the National Nutrition Education and Training Program Evaluation." Manuscript submitted for publication.

St.Pierre, R. G., Fairchild, C. K., Sullivan, D. J., Wertheimer, J., Layzer, J. I., & Light, R. J. (1981). *Evaluation of alternatives to commodity donation in the National School Lunch Program: Evaluation/analysis plan* (Vol. 1). Cambridge, MA: Abt Associates nc.

Stebbins, L. B., St.Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. (1978). "An evaluation of Follow Through." In T. D. Cook (Ed.), *Evaluation studies review annual,* (Vol. 3). Beverly Hills, CA: Sage Publications.

Swanson Center for Nutrition. (1979). *Nutrition education field test evaluation report.* Omaha, NE: Swanson Center for Nutrition.

West Virginia Department of Education, Bureau of Federal Programs and Services, Bureau of Planning, Research and Evaluation. (1977). *Evaluation report: The nutrition education team: Final report.* Charleston, WV: Author.

Wolff, H. (1980). *Evaluation of the curriculum guide for nutrition education: Preschool through grade six.* Report prepared for the California State Department of Education, Nutrition Education and Training Program, September 29.

Zeitlan, M. F. and Formacion, C. S. (1981). *Nutrition education in developing countries: Study II, Nutrition education.* Cambridge, MA: Oelgeschlager, Gunn and Hain.

26

Selectivity Problems in Quasi-Experimental Studies

Bengt Muthén and Karl G. Jöreskog

1. INTRODUCTION

Selectivity problems can occur whenever one tries to estimate population parameters from a nonrandom sample. When the sample of data is nonrandom, it is important to try to model, as realistically as possible, the process by which the observed units have been selected into the sample. Selective samples may occur because only individuals

---

AUTHORS' NOTE: *This article was presented at the Conference on Experimental Research in the Social Sciences, Gainesville, Florida, January 8-10, 1981. This project, Methodology of Evaluation Research, has been supported by the Bank of Sweden Tercentenary Foundation, project director Karl G. Jöreskog. The authors thank Bengt Dahlqvist for fast and skillful programming.*

From Bengt Muthen and Karl G. Joreskog, "Selectivity Problems in Quasi-Experimental Studies," *Evaluation Review*, 1983, 7(2), 139-174.

with certain characteristics, more or less precisely defined, are included in the sample. This may be the case in large social programs, for example, where only low-income families are eligible for the program (sample selection), or when individuals participate voluntarily in the program (self-selection). Selective samples may also occur in longitudinal studies due to attrition; that is, individuals fall out of the sample for various reasons, despite an initial random sample. Analyzing a selective sample as if it is random will result in biased and inconsistent estimates of the parameters.

Selectivity problems have been of considerable interest in recent econometric work, for example, see Stromsdorfer and Farkas (1980). Within the single-group regression framework, selectivity problems have been discussed in the context of labor force participation of married women by many writers, for example, Gronau (1974), Lewis (1974), and Heckman (1974, 1977). Selection modeling has also been applied to situations of self-selection in the choice of college education and regarding economic returns to schooling, in, for example, Griliches et al. (1978), Kenny et al. (1979), and Willis and Rosen (1979). Selectivity modeling in the analysis of longitudinal data has been considered by Hausman and Wise (1976, 1979). Selectivity problems have also been discussed in the context of evaluation of treatment effects in nonequivalent control group designs, for example by Goldberger (1972a, b), Cain (1975), in the overview by Reichardt (1979), and by Sörbom (1981).

In this article we shall discuss selectivity problems in terms of a model that in some respects is more general than those of previous writers. Selection modeling for a single group is considered in Section 2. Multiple-group issues are discussed in Section 3 and related to conventional analysis of covariance. A general model and its estimation is presented in Section 4. A simulation study is reported in Section 5, and in Section 6 an extension of the general selection model to latent variable models is discussed.

2. SELECTION IN A SINGLE GROUP

As an example from education, consider the case where y is an achievement test, x is a home background variable, and the model is

$$y = \beta_0 + \beta_1 x + \epsilon \qquad [1]$$

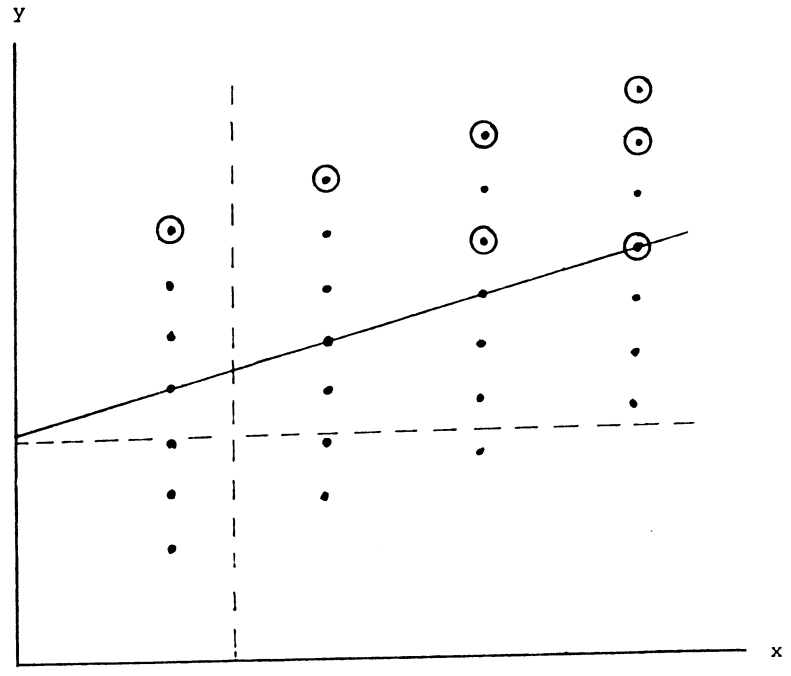


Figure 1

where ϵ is uncorrelated with x. Figure 1 shows a scatterplot of typical units in the population, say students of a certain age. (This graph is inspired by Hausman and Wise, 1976.) The straight line (equation 1) represents the population regression of y on x. If one has a random sample of observations on y and x, one can obtain unbiased estimates of β_0 and β_1 by ordinary least squares (OLS). If the sample is non-random, however, a population unit will be selected into the sample or not depending on the values of certain characteristics, which may be y, x, or other unobserved variables. If this fact is ignored, OLS will in general give biased estimates, which in turn leads to incorrect inferences for the full population. A solution to this problem is to try to model the selection process. Estimation can then be carried out for an extended model, including both the original regression relation, such as equation 1, and the selection model part. With a proper model specification, correct estimates can then be obtained for the parameters of equation 1.

In this article we shall assume that the selection process depends on a single selection variable η, and that units are selected into the sample if η exceeds a threshold value. In most cases the selection variable η is unobserved and its characteristics unknown. We shall first consider the cases when η coincides with x and y, respectively.

When η = x there is zero probability of selecting population units to the left of the vertical broken line in Figure 1. Here, students with "good" home background would be considered. However, this is of no consequence provided that units to the right of this line are selected randomly, and that the variation in x is sufficient to determine the slope β_1 of the population regression. Hence, when η = x, OLS gives unbiased estimates of β_0 and β_1.

When η = y, a unit is observed in the sample only if y exceeds a threshold. Here, only high-achieving students would actually be used in the analysis. In Figure 1 this means that population units below the horizontal broken line have zero probability of inclusion in the sample. In this case, the error term ϵ will be correlated with x in the sample, the mean of ϵ being larger for units with smaller x-values. When the threshold is zero, this corresponds to the familiar Tobit model (Tobin, 1958; Amemiya, 1973), originally proposed as a limited dependent variable model for consumption studies (no consumption if y = 0). OLS will give an estimate of the slope β_1, which is biased downward and inconsistent in large samples.

Now consider the case when η is unobserved. Units are selected if η exceeds a threshold. This is perhaps the most realistic case in the kind of application considered. Here, η may be a latent variable such as social disadvantage, where a high disadvantage results in an individual being selected. The reason for selection may be that limited funds are available, and measurement is concentrated on students that are thought to be in particular need of a certain schooling treatment. Still, the intent is to try to make inferences to the full population. The value of η represents a characteristic of the student that is not completely known to the investigator, who does not completely control the selection process. Figure 1 illustrates the probable case when η is negatively correlated with y, using encircled dots to exemplify population units that may not be included in the sample. For selectable units, the mean of ϵ is smaller for large x-values. Here there is no sharp division into selectable and nonselectable units. Ignoring selection will result in biased OLS estimates also in this case. This will be explicated below.

We will now consider the case when a proper specification of the selection process can be done, reviewing selection modeling attempted

in the literature so far. It should be noted that a weakness of selection modeling is that the general problem of misspecification may be enhanced. For instance, a misspecification of the regression relation, such as equation 1, may be mistaken for indication of selection bias (see Olsen, 1979; Stromsdorfer and Farkas, 1980: 13-41).

For the education example it may be realistic to assume that the selection variable η is linearly related to (although not completely determined by) the observed home background variable x,

$$\eta = \gamma_0 + \gamma_1 x + \delta \qquad [2]$$

We shall assume that in the total population the joint distribution of ϵ and δ is independent of x with means zero and with covariance matrix

$$\underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_{\epsilon\epsilon} & \\ \sigma_{\delta\epsilon} & \sigma_{\delta\delta} \end{bmatrix}$$

where $\sigma_{\epsilon\epsilon}$ is the variance of ϵ, $\sigma_{\delta\delta}$ is the variance of δ, and $\sigma_{\delta\epsilon}$ is the covariance between ϵ and δ. We shall also use the regression of ϵ on δ,

$$\epsilon = \omega\delta + \upsilon \qquad [3]$$

where υ is uncorrelated with δ and $\omega = \sigma_{\delta\epsilon}/\sigma_{\delta\delta}$ is the regression coefficient. Since the scale for η is arbitrary, we may without loss of generality assume that the threshold is zero and that $\sigma_{\delta\delta} = 1$, in which case $\omega = \sigma_{\delta\epsilon}$. This assumption is made throughout this section.

Consider the regression of y on x for selectable units with $\eta > 0$,

$$E(y|x, \eta > 0) = \beta_0 + \beta_1 x + E(\epsilon|\eta > 0)$$

But

$$E(\epsilon|\eta > 0) = E(\epsilon|\delta > -\gamma_0 - \gamma_1 x)$$

$$= \omega E(\delta|\delta > -\gamma_0 - \gamma_1 x)$$

so that

$$E(y|x, \eta > 0) = \beta_0 + \beta_1 x + \omega E(\delta|\delta > -\gamma_0 - \gamma_1 x) \qquad [4]$$

Let $\lambda(x) = \gamma_0 + \gamma_1 x$. Then the last term in equation 4 involves $E[\delta \mid \delta > -\lambda(x)]$, which is a monotonically decreasing function of λ denoted by $f(\lambda)$. It is clear that the ordinary least squares regression of y on x fails to give consistent estimates of β_1, unless $\sigma_{\delta\epsilon} = 0 (\omega = 0)$. This is because the ordinary regression of y on x omits the random variable $f(\lambda)$, which is correlated with x.

Let $p(z)$ denote the probability density function of δ, and let $P(x)$ denote the corresponding probability distribution function. We assume that $p(z)$ is symmetric about zero so that $p(-z) = p(z)$ and $P(-z) = 1 - P(z)$. Then

$$\begin{aligned} \Pr(\delta > -\lambda) &= 1 - \Pr(\delta \leqslant -\lambda) \\ &= 1 - P(-\lambda) \\ &= P(\lambda) \end{aligned} \qquad [5]$$

and

$$\begin{aligned} f(\lambda) &= E(\delta \mid \delta > -\lambda) \\ &= [1/P(\lambda)] \int_{-\lambda}^{\infty} z p(z)\, dz \\ &= -[1/P(\lambda)] \int_{-\infty}^{\lambda} z p(z)\, dz \end{aligned} \qquad [6]$$

Table 1 shows $\sigma_{\delta\delta}$, $p(z)$, $P(z)$, and $f(\lambda)$ for some of the well-known distributions: the normal, the logistic, the Student's t and the Laplace (this table is adapted from Goldberger, 1980). The case when δ has a standard normal distribution is of particular interest. Let $\phi(z)$ be the standard normal density function and let $\Phi(z)$ be the corresponding distribution function. Then the integral in equation 6 becomes $-\phi(\lambda)$ so that

$$f(\lambda) = \phi(\lambda)/\Phi(\lambda)$$

Therefore, in this case we have

$$E(y \mid x, \eta > 0) = \beta_0 + \beta_1 x + \omega f(\lambda) \qquad [7]$$

TABLE 1

Variance $\sigma_{\delta\delta}$ and Functions p(z), P(z), and f(λ) for Some Selected Distributions

Distribution	*Variance* $\sigma_{\delta\delta}$	*Density* $p(z)$	*Distribution Function* $P(z)$	*Truncated Mean Function* $f(\lambda)$
Normal	1	$(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}z^2}$	$(2\pi)^{-\frac{1}{2}}\int_{-\infty}^{z}e^{-\frac{1}{2}x^2}dx$	$p(\lambda)/P(\lambda)$
Logistic	$\pi^2/3$	$e^z/(1+e^z)^2$	$1/(1+e^{-z})$	$[1/P(\lambda)]\log[1/(1-P(\lambda))]-\lambda$
Student*	$n/(n-2)$	$c_n(1+z^2/n)^{-\frac{1}{2}(n+1)}$	$c_n\int_{-\infty}^{z}(1+x^2/n)^{-\frac{1}{2}(n+1)}dx$	$[(n+\lambda^2)/(n-1)]\,p(\lambda)/P(\lambda)$
Laplace	2	$\frac{1}{2}e^{-\lvert z\rvert}$	$\frac{1}{2}e^{z}$, $z\leqslant 0$ $1-\frac{1}{2}e^{-z}$, $z>0$	$1-\lambda$, $\lambda\leqslant 0$ $(1+\lambda)/(2e^{\lambda}-1)$, $\lambda>0$

SOURCE: Adapted from Goldberger (1980).

*n is the degrees of freedom parameter and $c_n=\Gamma(\frac{1}{2}(n+1))/[\Gamma(\frac{1}{2}n)\cdot(n\pi)^{\frac{1}{2}}]$.

The conditional variance of y can also easily be obtained using equation 3 and known results for the truncated normal distribution (see Johnson and Kotz, 1972: 81-83).

$$\begin{aligned}\mathrm{Var}(y|x, \eta > 0) &= \mathrm{Var}(\epsilon|\eta > 0) \\ &= \omega^2 \mathrm{Var}(\delta|\delta > -\lambda) + \mathrm{Var}(v) \\ &= \omega^2 [1 - \lambda f(\lambda) - f^2(\lambda)] + \sigma_{\epsilon\epsilon} - \omega^2 \\ &= \sigma_{\epsilon\epsilon} - \omega^2 f(\lambda)\,[\lambda + f(\lambda)] \qquad [8]\end{aligned}$$

Hence, the true regression of y on x is nonlinear and heteroscedastic. Figure 2 shows $\phi(\lambda)$, $\Phi(\lambda)$, and $f(\lambda)$ for $-4 \leqslant \lambda \leqslant 4$. Figure 3 shows the mean (equation 7) and the variance (equation 8) of y as a function of x for $\beta_0 = \gamma_0 = 0$, $\beta_1 = 1$, $\gamma_1 = -1$ and $\omega = -1$. This corresponds to the third selection situation of Figure 1. It is seen that linear regression, ignoring selectivity, will here produce a downward biased estimate of β_1. Figures 4a-d show the straight line $\beta_0 + \beta_1 x$ and the true mean function for some combinations of the signs of β_1, γ_1 and ω.

Generalizing the previous model to an arbitrary number q of explanatory variables $\underset{\sim}{x}' = (x_1, x_2, \ldots, x_q)$, of which one may be the constant 1, and using vectors of regression coefficients $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$, the model

$$y = \underset{\sim}{\beta}'\underset{\sim}{x} + \epsilon \qquad [9]$$

$$\eta = \underset{\sim}{\gamma}'\underset{\sim}{x} + \delta \qquad [10]$$

$$y: \begin{cases} \text{observed if } \eta > 0, \\ \text{not observed, otherwise} \end{cases} \qquad [11]$$

with normally distributed errors can be seen as a generalized Tobit model, where the assumption $\eta = y$ has been relaxed (see Cragg, 1971). In addition to consumption and labor force studies in econometrics, where the y-variable is limited, this model has been used to model selectivity in various applications of the kind discussed in Section 1. This generalized Tobit model is the basic model we will use henceforth. For a recent survey of the statistical treatment of Tobit models, see Amemiya (1982).

The generalized Tobit model may be interpreted in two parts corresponding to the two relations in equation 10 and equation 9. With

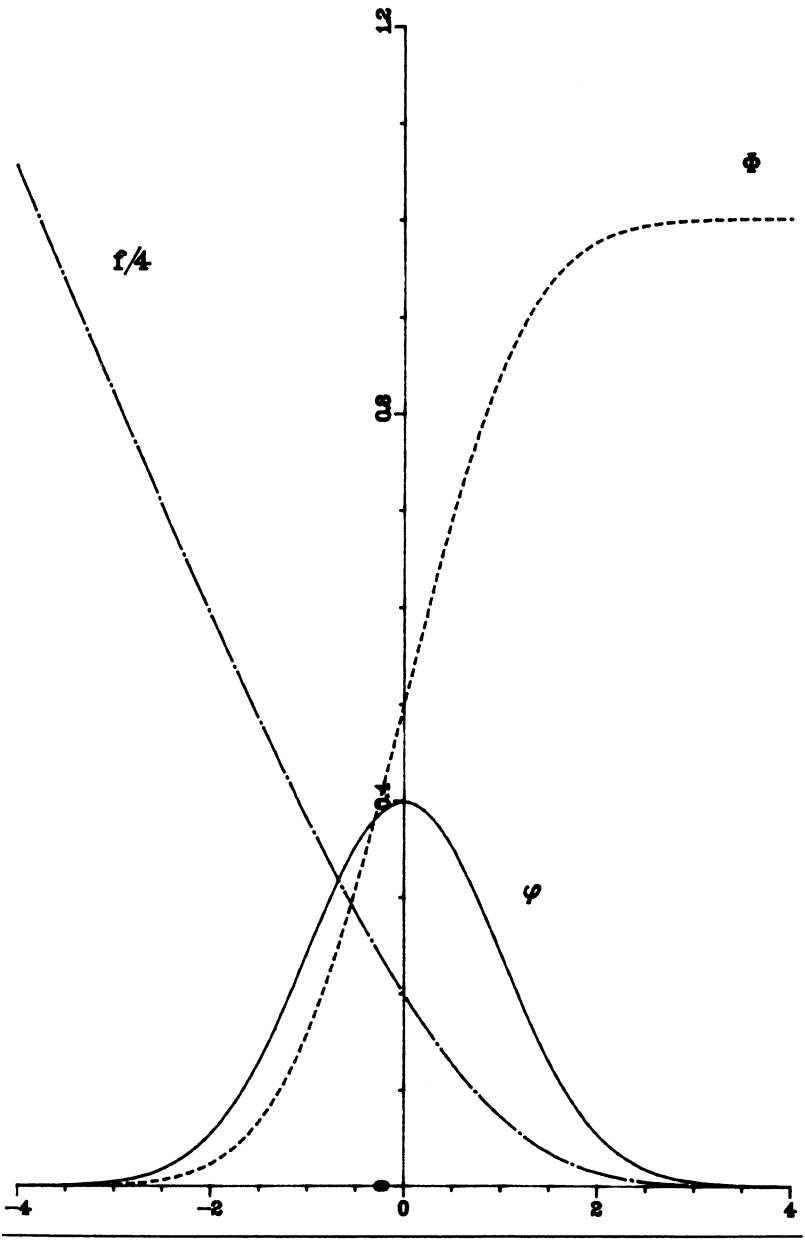


Figure 2: Functions $\phi(\lambda)$, $\Phi(\lambda)$, and $f(\lambda)$ for $-4 \leqslant \lambda \leqslant 4$

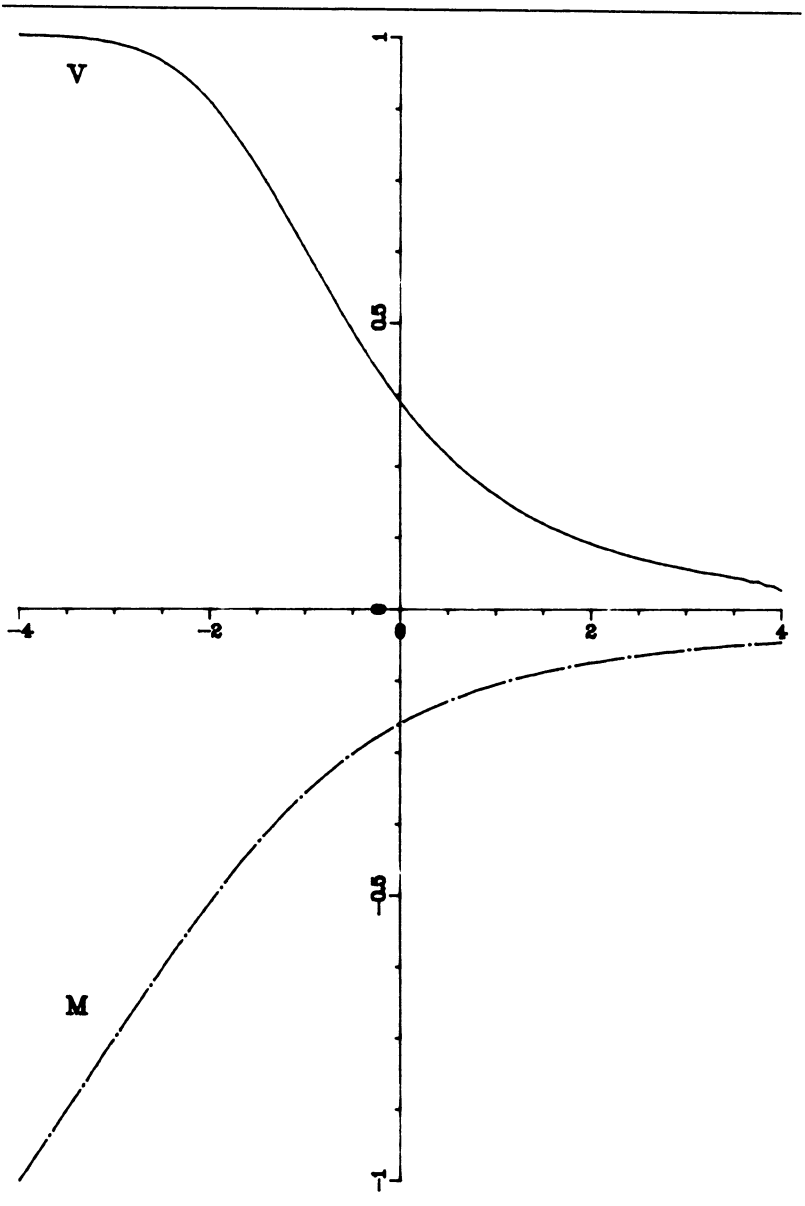


Figure 3: Mean Function $M(x) = \frac{1}{4}[\beta_0 + \beta_1 x + \omega f(\lambda)]$
Variance Function $V(x) = \sigma_{\epsilon\epsilon} - \omega^2 f(\lambda)[\lambda + f(\lambda)]$ for
$\beta_0 = \gamma_0 = 0$, $\beta_1 = 1$, $\gamma_1 = -1$, $\omega = -1$ and $\sigma_{\epsilon\epsilon} = 1$ and $-4 \leqslant x \leqslant 4$

FIGURE 4A

$\beta = 1, \gamma = -1, \omega = -1$

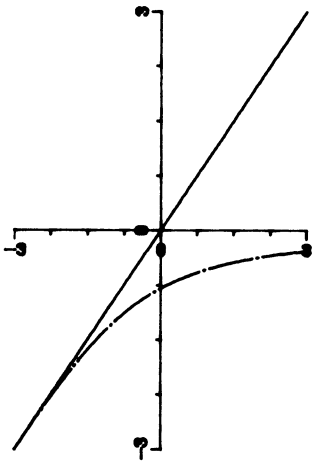


FIGURE 4B

$\beta = 1, \gamma = 1, \omega = 1$

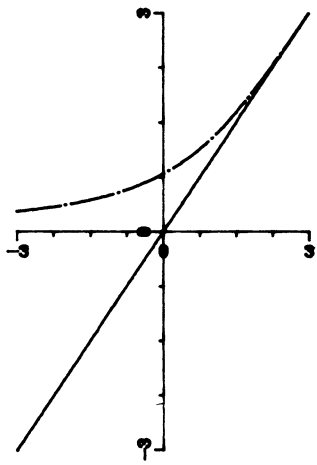


FIGURE 4C

$\beta = -1, \gamma = -1, \omega = 1$

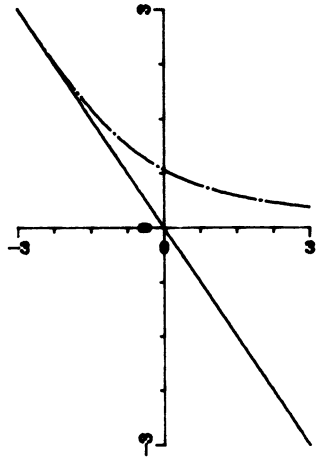
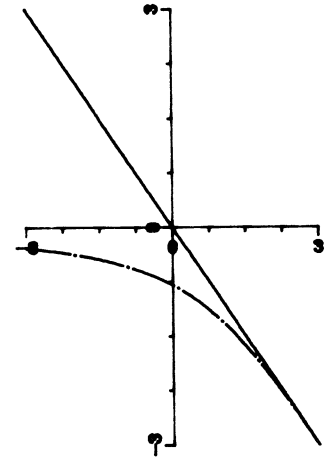


FIGURE 4D

$\beta = -1, \gamma = 1, \omega = -1$



Figures 4A-D: Linear Function $L(x) = \beta x$ and Mean Function $M(x) = \beta x + \omega f(\gamma x)$ for Four Combinations of β, γ, and ω and $-3 \leqslant x \leqslant 3$.

$\sigma_{\delta\delta} = 1$, the probability of the event y observed ($\eta > 0$) follows a Probit model,

$$\Pr(\text{y observed } |\underset{\sim}{x}) = \Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}), \qquad [12]$$

With $\Pr(\text{y not observed } |\underset{\sim}{x}) = 1 - \Phi(\underset{\sim}{\gamma}'\underset{\sim}{x})$. The second part describes the distribution of y given $\underset{\sim}{x}$ and $\eta > 0$,

$$E(y|\underset{\sim}{x}, \eta > 0) = \underset{\sim}{\beta}'\underset{\sim}{x} + \omega f(\underset{\sim}{\gamma}'\underset{\sim}{x}). \qquad [13]$$

Two types of samples must be distinguished. In the Probit model (equation 12), it is assumed that the sample includes units for which $\underset{\sim}{x}$ is recorded also for those with $\eta \leqslant 0$. This will be referred to as the *censored* case. When such units cannot occur in the sample, we have the *truncated* case.

Several techniques have been proposed for the estimation of equations 9, 10, and 11, using maximum-likelihood methods (e.g., see Griliches et al., 1978; Hausman and Wise, 1979), and various two-stage estimators applicable to the censored case only (e.g., see Heckman, 1979; Maddala and Lee, 1976). We will consider maximum-likelihood estimation, but the Heckman estimator will also be reported in the simulation study in Section 5.

In the first step of the Heckman estimator, $\underset{\sim}{\gamma}$ is estimated by maximum-likelihood Probit analysis. In the second step, OLS is applied to equation 13 in the truncated sample using the estimated $f(\lambda)$ as an additional x variable.

Recent contributions to selection modeling in the single-group case include studies pertaining to the robustness against deviations from the assumed functional form and error structure (e.g., see Goldberger, 1980; Hurd, 1979; Nelson, 1979; Olsen, 1979; Ray et al., 1980), and generalizations to more than one selection relation (e.g., see Tunali et al., 1980; Venti and Wise, 1980).

3. MULTIPLE-GROUP COMPARISONS

In this section we consider the analysis of treatment (intervention) effects for nonequivalent group designs. Such quasi-experimental designs are common in the evaluation of social programs or social experiments. Of particular concern is the case where outcome measurements are made before and after the intervention, for a control group,

and one or several treatment groups. In practice, randomization is infrequently accomplished, and the problem is how to separate the potential treatment effect(s) from group differences only produced by the way individuals were assigned to the different groups. Although originally intended for experimental settings, analysis of covariance (ANCOVA) is frequently used in this situation. Such ANCOVA applications have been strongly criticized, and attempts have been made to adjust the technique to fit the quasi-experimental setting. For a good summary of the problems and the various adjustment techniques, see Reichardt (1979) and Weisberg (1979).

As in the previous section, the nonequivalence of the control and treatment groups may be due to the investigator choosing to treat a certain subset of individuals (such as particularly needy ones) or due to self-selection by the individuals (such as volunteers in a new program). Nonequivalent groups may also arise due to attrition, despite initial randomization.

Data of this sort may be viewed as samples from different groups (populations) to be compared (see Thorndike, 1942). However, for one or several of the groups the sample(s) is (are) nonrandom or selective in the sense defined previously. Contrary to the ANCOVA approach, this article argues for the explicit modeling of the selection processes in order to avoid bias due to comparisons of nonequivalent groups.

Consider a quasi-experiment related to the education example discussed in Section 2. Say that there is one experimental group (E), one control group (C), and a single posttest y. Complete randomization has not been undertaken. It is suspected that the groups are different with respect to certain background characteristics, of which an important part is the variable x, say. The ANCOVA approach is to consider the regressions

$$\left.\begin{aligned} y^C &= \mu + \beta^C x^C + \epsilon^C \\ y^E &= \mu + \alpha + \beta^E x^E + \epsilon^E \end{aligned}\right\} \qquad [14]$$

assumed to hold for each of the two populations (groups). Given $\beta^C = \beta^E$, α is taken as the treatment effect.

With random sampling from each population (i.e., complete randomization), ANCOVA analysis, given $\beta^C = \beta^E$, consistently estimates the treatment effect α. If in fact x does not influence y, there would be no point in including it in equation 14, resulting in a simple analysis of variance. If x does influence y, its inclusion increases precision.

However, in many quasi-experimental studies there may be non-random selection of units to both controls and experimentals, and it is not necessarily the same selection variable that governs the selection process for controls and experimentals. A more appropriate model for this situation is

Controls:
$$y^C = \mu + \beta^C x^C + \epsilon^C \quad [15]$$

$$\eta^C = \gamma_0^C + \gamma_1^C x^C + \delta^C \quad [16]$$

$$y^C: \begin{cases} \text{observed if } \eta^C > 0 \\ \text{not observed, otherwise} \end{cases} \quad [17]$$

Experimentals:
$$y^E = \mu + \alpha + \beta^E x^E + \epsilon^E \quad [18]$$

$$\eta^E = \gamma_0^E + \gamma_1^E x^E + \delta^E \quad [19]$$

$$y^E: \begin{cases} \text{observed if } \eta^E > 0 \\ \text{not observed, otherwise} \end{cases} \quad [20]$$

Independent random sampling is assumed from the two populations in this new model. The specification is completed by a choice of bivariate distributions for the two sets of error terms ϵ and δ, given x. The selection relations (equations 16 and 19) are the needed auxiliaries to the causal relations (equations 15 and 18), in order to obtain unbiased treatment effect estimates.

The random error terms δ^C and δ^E include variables other than x, influencing selection. It is an important advantage of selection modeling that such variables need not be explicitly included, as in the ANCOVA model. Given equations 15 through 20 as the true model, the ANCOVA covariate x in the analysis of equation 14 does not completely control for the existing selectivity.

Again, if $\beta^C = \beta^E$, it is reasonable to take α as a measure of the treatment effect. Hence, we need a technique to analyze data from the two groups simultaneously under selection models in which some parameters are constrained to be equal in the two groups. These equality constraints should not be taken for granted, however, but must be tested by means of the data.

With the education example, γ^E and the correlation between ϵ^E and δ^E are presumably both negative. Individuals with an unusually high

social disadvantage score, that is, a high δ^E value, are those more likely to be selected, and they are also the ones likely to have lower achievement scores, or low ϵ^E value. For the selectable experimentals, the true regression of y on x is then nonlinear and of the same shape as in Figure 3. Ignoring selectivity, ANCOVA for the experimentals and controls will give biased results. This will be studied further in Section 5.2, in a similar artificial data example.

Barnow et al. (1980) and Goldberger (1979) formulated a special form of selection, where for one experimental group and one control group,

$$y^E: \begin{cases} \text{observed if } \eta^E > 0 \\ \text{not observed, otherwise} \end{cases}$$

$$y^C: \begin{cases} \text{observed if } \eta^C \leqslant 0 \\ \text{not observed, otherwise} \end{cases}$$

where $\eta^E = \eta^C$. They illustrated their model by the well-known Head Start Compensatory Education program, so that y is the posttest achievement score. Here, a single selection variable (related to family income of the child) defines group membership. In terms of our model, their model implies that γ parameters and error covariance parameters differ only in signs between controls and experimentals. This specification seems too restrictive.

4. A GENERAL MODEL AND ITS ESTIMATION

In the previous section we formulated selection modeling for a single explanatory variable, x, and for one or two groups, the emphasis being on the basic ideas of the model. In this section we generalize the model to an arbitrary number of exogenous (explanatory) variables and to an arbitrary number of groups. The data will be regarded as sampled from G groups or populations, and for each group a univariate regression relation and a single selection relation is assumed. This formulation is chosen for simplicity; it may be generalized to a multivariate system (a structural equation system) for each group, to multivariate selection relations for each group, and also to categorical response variables.

It is essential to distinguish between two parts of the model: the causal relation and the selection relation. For each group g, g = 1, 2, . . . , G, we assume a causal relation of the form:

$$y^{(g)} = \underset{\sim}{\beta}^{(g)\prime}\underset{\sim}{x}^{(g)} + \epsilon^{(g)} \qquad [21]$$

where $\underset{\sim}{\beta}^{(g)}$ is a q × 1 vector of parameters, $\underset{\sim}{x}^{(g)}$ is a vector of random explanatory variables, and $\epsilon^{(g)}$ is a random residual, independent of $\underset{\sim}{x}^{(g)}$. We do not assume random sampling from the populations of equation 21. Instead, in addition to equation 21 we assume the selection relations for g = 1, 2, . . . , G

$$\eta^{(g)} = \underset{\sim}{\gamma}^{(g)\prime}\underset{\sim}{x}^{(g)} + \delta^{(g)} \qquad [22]$$

$$y^{(g)}: \begin{array}{l} \text{observed if } \eta > 0 \\ \text{not observed, otherwise} \end{array} \qquad [23]$$

where $\eta^{(g)}$ is a latent selection variable, $\underset{\sim}{\gamma}^{(g)}$ is a q × 1 vector of parameter, $\underset{\sim}{x}^{(g)}$ is as before, and $\delta^{(g)}$ is a random residual, independent of $\underset{\sim}{x}^{(g)}$. Let $\sigma^{(g)}_{\epsilon\epsilon}$, $\sigma^{(g)}_{\delta\epsilon}$, $\sigma^{(g)}_{\delta\delta}$ be the variance of $\epsilon^{(g)}$, the covariance between $\epsilon^{(g)}$ and $\delta^{(g)}$, and the variance of $\delta^{(g)}$, respectively. We assume for each group a bivariate normal distribution for $\epsilon^{(g)}$ and $\delta^{(g)}$, g = 1, 2, . . . , G. The groups are assumed to be independent, and for each group g we consider random sampling from the population given by equations 21, 22, and 23.

Each parameter of the model will be allowed to be any of three kinds: a free parameter, a parameter fixed to a certain value, or a parameter constrained to be equal to another parameter. For instance, both $\underset{\sim}{\beta}^{(g)}$ and $\underset{\sim}{\gamma}^{(g)}$ may contain parameters fixed to zero so that the same exogenous variables do not necessarily operate in equations 21 and 22. Also, group invariance of parameters can be tested by applying equality restrictions.

Group level differences are captured by $\beta^{(g)}$ parameters corresponding to unit x variables. The ANCOVA model, with possible group invariance of slopes and residual variances, is a special case of the above formulation, where $\sigma^{(g)}_{\delta\epsilon} = 0$ for g = 1, 2, . . . , G. Then the relation (equation 22) is inconsequential (see also the likelihood expressions that follow). A selection process may operate ($\sigma^{(g)}_{\delta\epsilon} \neq 0$) in one or more of the groups, and may operate differently in different groups.

Consider the bivariate distribution of y and η given $\underset{\sim}{x}$ and $\eta > 0$. For simplicity, the group index is omitted. Let $\phi(z; \mu, \sigma^2)$ denote the normal density for a variable z with mean μ and variance σ^2, let $\phi_{y\eta}$ denote the bivariate normal density for y and η given x, and let $\Phi(a)$ denote the probability that a standard normal variable falls below a. Let $\sigma_\delta = \sqrt{\sigma_{\delta\delta}}$.

The probability that $\eta > 0$, given $\underset{\sim}{x}$, may then be written as $\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1})$. The density for the singly truncated bivariate normal distribution for y and η, given $\underset{\sim}{x}$, is

$$p_{y\eta} = \begin{cases} 0, \text{ if } \eta \leqslant 0 \\ \phi_{y\eta}/\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1}), \text{ otherwise.} \end{cases} \qquad [24]$$

From equation 24 we obtain the marginal distribution for y as

$$p_y = \int_0^\infty \phi_{y\eta}\, d\eta/\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1})$$

$$= \phi(y; \underset{\sim}{\beta}'\underset{\sim}{x}, \sigma_{\epsilon\epsilon}) \times \Phi(\mu_{\eta\cdot y}\sigma_{\eta\cdot y}^{-1})/\Phi(\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1}) \qquad [25]$$

where

$$\mu_{\eta\cdot y} = \underset{\sim}{\gamma}'\underset{\sim}{x} + \sigma_{\delta\epsilon}\,\sigma_{\epsilon\epsilon}^{-1}\,(y - \underset{\sim}{\beta}'\underset{\sim}{x}) \qquad [26]$$

$$\sigma_{\eta\cdot y}^2 = \sigma_{\delta\delta} - \sigma_{\delta\epsilon}^2\,\sigma_{\epsilon\epsilon}^{-1} \qquad [27]$$

Equation 25 gives the density of y, given $\underset{\sim}{x}$, in a truncated sample. The event $\eta > 0$ has probability one in such a sample. This means that population units with $\eta \leqslant 0$ cannot be included in the sample; not only do we not observe y, but we do not observe $\underset{\sim}{x}$ either. In the contrary case of a censored sample, a unit for which y is not observed ($\eta \leqslant 0$), given $\underset{\sim}{x}$, occurs with the probability $\Phi(-\underset{\sim}{\gamma}'\underset{\sim}{x}\sigma_\delta^{-1})$. In this case, the density of y given $\underset{\sim}{x}$ when $\eta > 0$ is that of equation 25 except that the denominator cancels out.

The likelihood for both the truncated and the censored case may be summarized in the following way. For g = 1, 2, . . . , G, let $N^{(g)}$ denote

the sample size, let $N_t^{(g)}$ denote the number of sample units for which $\underset{\sim}{x}^{(g)}$ is observed, but not $y^{(g)}$, let

$$s^{(g)} = \begin{cases} 1, \text{ if selection occurs in group g} \\ 0, \text{ otherwise (random sample assumed)} \end{cases} \quad [28]$$

and let

$$t^{(g)} = \begin{cases} 1, \text{ if } \eta \leqslant 0\text{-units cannot occur in the} \\ \text{sample for group g (truncated sample)} \\ 0, \text{ otherwise (censored sample)} \end{cases} \quad [29]$$

The log likelihood for independent samples from the G groups may then be written

$$\log L = \sum_{g=1}^{G} \left[s^{(g)}(1 - t^{(g)}) \sum_{i=1}^{N_t^{(g)}} \log f_1(\underset{\sim}{x}_i^{(g)}) + \sum_{i=N_t^{(g)}+1}^{N^{(g)}} \log f_2(y_i^{(g)}, \underset{\sim}{x}_i^{(g)}) \right] \quad [30]$$

where

$$f_1(\underset{\sim}{x}_i^{(g)}) = \Phi(-\underset{\sim}{\gamma}^{(g)\prime} \underset{\sim}{x}_i^{(g)} \sigma_\delta^{-1}) \quad [31]$$

$$f_2(y_i^{(g)}, \underset{\sim}{x}_i^{(g)}) = \phi(y_i^{(g)}; \underset{\sim}{\beta}^{(g)\prime} \underset{\sim}{x}_i^{(g)}, \sigma_{\epsilon\epsilon}^{(g)}) \times \left\{ \Phi\left(\mu_{\eta \cdot y_i}^{(g)} \sigma_{\eta \cdot y_i}^{(g)-1} \right) \right\}^{s^{(g)}} \times \left\{ \Phi\left(\underset{\sim}{\gamma}^{(g)\prime} \underset{\sim}{x}_i^{(g)} \sigma_\delta^{(g)-1} \right) \right\}^{-t^{(g)}s^{(g)}} \quad [32]$$

Maximum-likelihood (ML) estimates are obtained from equation 30 in a straightforward fashion. The numerical optimization may

however be nontrivial, since the shape of the likelihood function can be complicated, yielding convergence problems. This may be particularly pressing in cases where the model does not fit well, with small samples or with poor starting values. In the censored case, reasonable starting values may be obtained from a separate Probit regression and an OLS regression in the truncated sample. The truncated case is relatively more difficult since separate Probit type information is not available for the estimation of the γ-parameters. We note that for the truncated case when $\sigma_{\delta\epsilon} = 0$ holds exactly, γ will in fact be indeterminate.

Of some importance is the choice of parameterization in the actual computations. To ensure positive values for the variance expressions in the likelihood, we use the following parameterization. Consider the new parameter $\sigma^*_{\epsilon\epsilon}$, defined by $\sigma_{\epsilon\epsilon} = e^{\sigma^*_{\epsilon\epsilon}}$ yielding positive $\sigma_{\epsilon\epsilon}$. Also, the indeterminacy of the scale of η is used to set $\sigma^2_{\eta\cdot y} = 1$; that is, $\sigma_{\delta\delta}$ is not a free parameter, but is restricted as $\sigma_{\delta\delta} = 1 + \sigma_{\delta\epsilon} e^{-\sigma^*_{\epsilon\epsilon}}$. Hereby, all parameters in the optimization are free to vary from minus to plus infinity. In the actual reporting of the estimates, however, we find it convenient to revert to the more conventional parameterization with $\sigma_{\epsilon\epsilon}$ and $\sigma_{\delta\delta} = 1$. Also, instead of $\sigma_{\delta\epsilon}$, we will report the correlation between the errors, denoted by ρ. Standard errors will be given for this latter set of parameter estimates.

Let d_i be defined such that

$$\log L = \sum_{g=1}^{G} \sum_{i=1}^{N^{(g)}} d_i^{(g)}$$

and let

$$\underset{\sim}{A} = \sum_{g=1}^{G} \sum_{i=1}^{N^{(g)}} (\partial d_i^{(g)}/\partial \underset{\sim}{\theta} \times \partial d_i^{(g)}/\partial \underset{\sim}{\theta}')_{\underset{\sim}{\theta}=\underset{\sim}{\hat{\theta}}} \qquad [33]$$

where $\underset{\sim}{\hat{\theta}}$ is the ML estimate of the parameter vector $\underset{\sim}{\theta}$. The squared, asymptotic standard errors of $\hat{\theta}$ may then be found on the diagonal of A^{-1} (see also Griliches et al., 1978).

For the iterative optimization involved, the so-called FLEPOW algorithm is used (see Gruvaeus and Joreskog, 1970). This algorithm is based on a rapidly converging quasi-Newton method that makes use of first-order derivatives of the likelihood function, and a positive definite weight matrix $\underset{\sim}{E}$ that is built up during the iterations to approximate the

inverse of the Hessian matrix at the solution point. Starting values for the parameter estimates must be provided and also a starting value of $\underset{\sim}{E}$. For reasonably good starting values, a starting value of $\underset{\sim}{E}$ may be obtained by evaluating $\underset{\sim}{A}$ at that point and using $\underset{\sim}{E} = \underset{\sim}{A}^{-1}$. A similar algorithm was presented in Berndt et al. (1974) and has been applied, for example, in work by Hausman and Wise (1976, 1979).

5. ANALYSIS OF SIMULATED DATA

The aim of this section is to illustrate the models and selectivity issues discussed in previous sections by analysis of data sets generated from three different basic models. No more than two samples of different size will be drawn in each case, hence this study is more limited in scope than a rigorous Monte Carlo investigation. A similiar study was carried out by Wales and Woodland (1980) for the original Tobit model in the single group case.

5.1. SINGLE-GROUP DATA

Two basic models will be used here. Model 1 is specified with a single x,

$$y = 0.0 + 1.0x + \epsilon \qquad [34]$$

$$\eta = 0.0 - 1.0x + \delta \qquad [35]$$

$\sigma_{\delta\epsilon} = \sigma_{\delta\delta} = 1.0$, $\rho = -0.5$, and the mean and variance of x are chosen as $\mu_x = 0.0$, $\sigma_{xx} = 1.0$. Also, x is taken to be normally distributed.

Model 2 is specified with three x variables, where not all variables are operating in both the regression and selection relation,

$$y = 1.2 + 2.0x_1 + 1.0x_2 + 0.0x_3 + \epsilon \qquad [36]$$

$$\eta = 1.0 + 1.0x_1 + 0.0x_2 + 2.0x_3 + \delta \qquad [37]$$

where $\sigma_{\epsilon\epsilon} = 1.44$, $\sigma_{\delta\delta} = 1.0$, $\rho = 0.5$, and $\mu_{x_1} = \mu_{x_2} = \mu_{x_3} = 0.0$, $\sigma_{x_1x_1} = 1.0$, $\sigma_{x_2x_2} = 0.04$, $\sigma_{x_3x_3} = 0.01$, $\rho_{x_1x_2} = 0.8$, $\rho_{x_1x_3} = 0.3$, $\rho_{x_2x_3} = 0.5$. The x variables are taken to be trivariate normal.

Given the two basic models, trivariate and five-variate normal vectors corresponding to (y x η) for Model 1 and (y x_1 x_2 x_3 η) for Model 2 are generated. Two basic sample sizes are used in each case, N = 1000 and N = 4000. For each model and basic sample, a truncated subsample is created by including only selectable units, for which $\eta > 0$. In the corresponding censored case, the sample size is maintained as 1000 or 4000, where units with $\eta \leqslant 0$ are considered as nonselected, lacking observed y values. The truncated and censored cases may be viewed as corresponding to different real-life situations, where different amounts of data information are available.

For each basic model and sample size, several analyses are made. Using only the truncated sample, ordinary regression ignoring selection is reported. This is compared with the correct model formulation, that is, the truncated case of the respective basic model estimated by ML according to Section 4. With the censored sample case, ML Probit regression for the estimation of $\underset{\sim}{\gamma}$ will be reported together with the Heckman estimator for β and ω. This is compared with the full model formulation, the censored case of the respective basic model, estimated by ML according to Section 4. We can also study the gain in precision of the estimates, comparing the truncated and censored case, estimated under the correct model formulation.

The results are given in Table 2 for Model 1 and in Table 3 for Model 2. For Model 1 there is strong selectivity, where only about half of the full population consists of selectable units. For Model 2 the corresponding figure is about three-quarters. Hence, we find overall more markedly biased estimates from ordinary regression for the first model. The columns "Truncated Case" and "Censored Case" give ML estimates in accordance with Section 4. In the truncated case, this estimator performs well, and the estimates are in no case more than twice the standard errors from the true values. For N = 1000, some of the standard errors are, however, rather large.

With information corresponding to the censored case, the Probit estimator for $\underset{\sim}{\gamma}$ works extremely well in all cases. It is in fact comparable to the also high performance full information ML estimator (censored case), also with respect to precision in the estimates. The Heckman estimator for the β parameters performs very well, and is also comparable to the ML estimator. Note that $\sigma_{\epsilon\epsilon}$ is not consistently estimated (underestimated) and that the standard errors that are given are only approximate and too low, since these quantities are obtained via

TABLE 2
Parameter Estimates for Data Simulated According to Model 1*

Parameter	Population Value	Regression	Probit	Heckman	Truncated Case	Censored Case
			$N_t = 496, N = 1000$			
β_0	.0	−.373 (.054)		.101 (.278)	−.209 (.119)	.074 (.179)
β_1	1.0	.788 (.052)		1.048 (.062)	.931 (.095)	1.033 (.114)
$\sigma_{\epsilon\epsilon}$	1.0	.985 (.065)		.979 (.064)	.982 (.076)	1.126 (.131)
ω				−.587 (.333)		
γ_0	.0		.001 (.046)		.991 (1.599)	.013 (.046)
γ_1	−1.0		−1.033 (.067)		−3.448 (4.542)	−1.040 (.068)
ρ	−.5			−.593	−.248 (.413)	−.522 (.164)
			$N_t = 1963, N = 4000$			
β_0	.0	−.435 (.027)		.002 (.149)	−.223 (.137)	.013 (.083)
β_1	1.0	.807 (.027)		1.059 (.088)	.965 (.084)	1.065 (.054)
$\sigma_{\epsilon\epsilon}$	1.0	.916 (.029)		.911 (.028)	.978 (.056)	1.054 (.062)
ω				−.539 (.183)		
γ_0	.0		.020 (.023)		.851 (.723)	.021 (.023)
γ_1	−1.0		−1.040 (.032)		−1.277 (.346)	−1.043 (.032)
ρ	−.5		−.542		−.521 (.122)	−.538 (.078)

*Standard errors in parentheses.

TABLE 3
Parameter Estimates for Data Simulated According to Model 2*

Parameter	Population Value	Regression	Probit	Heckman	Truncated Case	Censored Case
			$N_t = 239$, $N = 1000$			
β_0	1.2	1.462		1.270	1.388	1.192
		(.046)		(.123)	(.064)	(.076)
β_1	2.0	1.848		1.979	1.923	2.033
		(.076)		(.106)	(.088)	(.092)
β_2	1.0	.397		.415	.450	.414
		(.396)		(.397)	(.431)	(.415)
β_3	0.0	.442		.710	.482	.834
		(.477)		(.497)	(.537)	(.529)
$\sigma_{\epsilon\epsilon}$	1.44	1.323		1.318	1.351	1.475
		(.071)		(.070)	(.084)	(.115)
ω				.522		
				(.314)		
γ_0	1.0		1.007		3.075	1.009
			(.060)		(.998)	(.057)
γ_1	1.0		.994		2.126	.982
			(.104)		(.922)	(.107)
γ_2	0.0		−.090		.038	−.009
			(.497)		(3.570)	(.509)
γ_3	2.0		1.897		1.214	1.754
			(.603)		(3.652)	(.613)
ρ	0.5			.454	.670	.602
					(.263)	(.113)
			$N_t = 984$, $N = 4000$			
β_0	1.2	1.432		1.171	1.265	1.160
		(.023)		(.066)	(.074)	(.039)
β_1	2.0	1.773		1.966	1.942	1.973
		(.039)		(.058)	(.073)	(.047)
β_2	1.0	1.049		1.028	.939	1.030
		(.197)		(.196)	(.276)	(.204)
β_2	.0	−.156		.280	.058	.303
		(.251)		(.268)	(.346)	(.268)
$\sigma_{\epsilon\epsilon}$	1.44	1.356		1.348	1.446	1.506
		(.035)		(.034)	(.057)	(.055)
ω				.690		
				(.167)		

(continued)

TABLE 3 (Continued)

Parameter	*Population Value*	*Regression*	*Probit*	*Heckman*	*Truncated Case*	*Censored Case*
γ_0	1.0		.900		1.547	.991
			(.029)		(.478)	(.029)
γ_1	1.0		.994		1.352	1.000
			(.051)		(.307)	(.050)
γ_2	.0		−.083		−.779	−.104
			(.252)		(1.323)	(.246)
γ_3	2.0		2.054		1.502	2.023
			(.314)		(1.639)	(.308)
ρ	.5			.594	.590	.585
					(.093)	(.057)

*Standard errors in parentheses.

ordinary regression (see Heckman, 1979; with corrections in Stromsdorfer and Farkas, 1980: ch. 2, where a consistent estimator for $\sigma_{\varepsilon\varepsilon}$ and appropriate standard errors are given).

For Model 2 the zero population coefficients result in poorly estimated β_2, β_3 parameters for the N = 1000 case. Here, the regression relation is misspecified, since x_3 is included. It seems as if the nonlinear influence of x_3 via the selection relation is picked up by a linear term in the regression relation. Fixing $\beta_3 = 0$ gives an improved overall picture, with β_2 estimate in the censored case .713(.363).

In the N = 4000 cases, the gain in precision when using information from nonselected units is well reflected in the lower standard errors for the columns "Censored Case" as compared to the columns "Truncated Case." Going from N = 1000 to N = 4000, we would expect a reduction of standard errors to about half the size. This pattern holds for the censored case, but not for the truncated case; suggesting that the large-sample approximation of the standard errors is rather poor in the truncated case for the smaller sample sizes used.

We finally report the computing time for the ML estimator in the censored case and for Model 2, starting with the regression and Probit estimates, and $\rho = 0$. On the IBM 370/158, the time used was about two minutes for N = 1000 and about nine minutes for N = 4000.

5.2. DATA FOR TWO GROUPS

By means of a model for two groups, Model 3, we will now illustrate the Section 3 issues regarding the estimation of treatment effects using nonequivalent groups. We chose a model for a control group and an experimental group in line with the example of Section 3. For the control group, Model 3 states

$$y^C = -0.4 + 0.8x^C + \epsilon^C \quad [38]$$

and that random sampling is the case. Here, $\mu_x^C = 0.0$, $\sigma_{xx}^C = 1.0$, $\sigma_{\epsilon\epsilon}^C = 0.9$. For experimentals, Model 3 is the same as Model 1,

$$y^E = 0.0 + 1.0x^E + \epsilon^E \quad [39]$$

$$\eta^E = 0.0 - 1.0x^E + \delta^E \quad [40]$$

with means, variances, and covariances as before. In fact, the same sample will be used for this group. We will only study the case of $N^{(g)} = 4000$, where $N^{(g)}$ is the basic sample size in each of the groups. We may view the full population regressions (equations 38 and 39) in the following way. We first consider a single parent population. Treatment produces two new subpopulations, one for controls, which is the same as before; while, as is seen for experimentals, the treatment affects both the intercept and the slope. Thus, there is a positive true treatment effect for large x values ($x > -2$), and the effect increases with increasing x.

We first study ANCOVA, which, ignoring selectivity, is carried out on the truncated sample. In this case, there are still 4000 controls, but only 4000 - 1963 = 2037 experimentals. Ordinary ANCOVA assumes group-invariant slopes and residual variances, so that the treatment effect is taken to be the difference in intercepts. Testing this invariance hypothesis for the data at hand, we obtain $\chi^2(2) = .047$, using a standard likelihood-ratio test. Due to selectivity, ANCOVA is therefore unable to reject this hypothesis. The results are presented in Table 4. The treatment effect, $\beta_0^E - \beta_0^C$, is estimated as -.016, but is not significant, $\chi^2(1) = .356$. Hence, selection of the most needy ones to the experimental group masks the positive true treatment effect. This is because the ANCOVA covariate, x, does not completely control for the non-

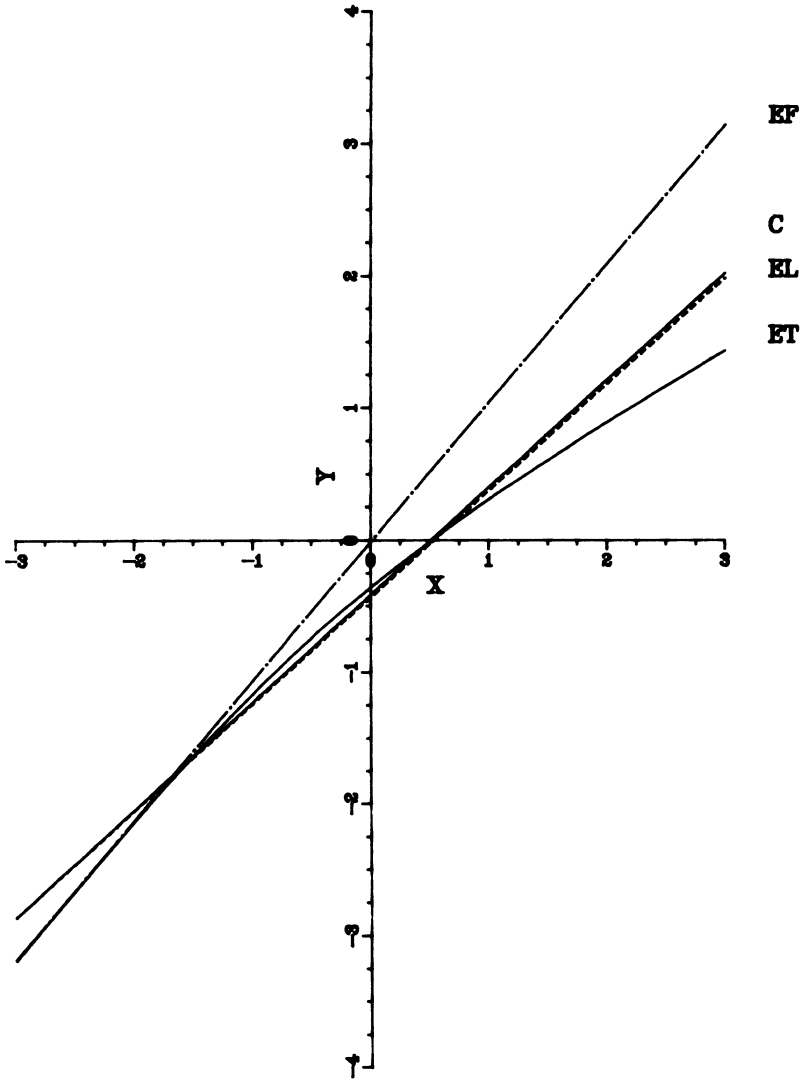
TABLE 4
Parameter Estimates for Control Group and Experimental Group Data Simulated According to Model 3*

Parameter	Population Value	ANCOVA	Truncated Case	Censored Case
		Controls		
β_0^C	−.4	−.416	−.416	−.416
		(.015)	(.015)	(.015)
β_1^C	.8	.812	.813	.813
		(.013)	(.014)	(.014)
$\sigma_{\epsilon\epsilon}^C$	.9	.915	.914	.914
		(.017)	(.020)	(.020)
		Experimentals		
β_0^E	.0	−.432	−.223	.013
		(.023)	(.137)	(.083)
β_1^E	1.0	.812	.965	1.065
		(.013)	(.084)	(.054)
$\sigma_{\epsilon\epsilon}^E$	1.0	.915	.978	1.054
		(.017)	(.056)	(.060)
γ_0^E	.0		.851	.021
			(.723)	(.023)
γ_1^E	−1.0		−1.277	−1.043
			(.346)	(.032)
ρ^E	−.5		−.521	−.538
			(.122)	(.078)

*Standard errors in parentheses.

equivalence of the groups due to selectivity. Figure 5 shows this situation graphically using the different estimated regressions.

Allowing for selectivity in the experimental group, the ML estimator of Section 4 was applied to the same data. The test of group-invariant slopes and error variances resulted in $\chi^2(2) = 7.002 (p < .05)$ for the truncated case and $\chi^2(2) = 13.462 (p < .005)$ in the censored case. In both analyses, the hypothesis is correctly rejected. The estimates are given in the two right-most columns of Table 4. In both cases the estimated regression lines correspond well to the true lines.



EF = Regression line for experimentals, estimating the full population regression.
ET = Regression curve for experimentals, estimating the regression in the selected subpopulation.
EL = Regression line for experimentals, estimating the linear approximation to the regression in the selected subpopulation.
C = Regression line for control, estimating the full population.

Figure 5: Estimated Regressions for Control Group and Experimental Group Data Simulated According to Model 3

6. LATENT EXOGENOUS VARIABLES

We will now discuss the analysis of selective samples in the context of latent (unobserved) variable models. A general approach to the study of latent variable models has been given, for example, by Jöreskog (1977; see also Sörbom and Jöreskog, 1981). For simplicity we will here limit ourselves to the case where the latent variables occur on the right-hand side in the regression relation of interest.

6.1. GENERAL RESULTS ON SELECTION

It will be useful to review some classical results on selection in multivariate distributions due to Pearson (1912) and Lawley (1943-1944), and applied to factor analysis models by Meredith (1964).

Pearson and Lawley considered influences of selection on a random vector variable $\underset{\sim}{z}$, say. For a set of selection variables, here denoted by $\underset{\sim}{\eta}$, selection is of a general type, changing the density of $\underset{\sim}{z}$, p_z, into p_z^*. Given that the regression of $\underset{\sim}{z}$ on $\underset{\sim}{\eta}$ in the total population is linear and homoscedastic, it is shown that

$$\underset{\sim}{\mu}_z = \underset{\sim}{\mu}_z^* - \underset{\sim}{\Sigma}_{z\eta}^* \underset{\sim}{\Sigma}_{\eta\eta}^{*-1} (\underset{\sim}{\mu}_\eta^* - \underset{\sim}{\mu}_\eta) \qquad [41]$$

$$\underset{\sim}{\Sigma}_{zz} = \underset{\sim}{\Sigma}_{zz}^* - \underset{\sim}{\Sigma}_{z\eta}^* (\underset{\sim}{\Sigma}_{\eta\eta}^{*-1} - \underset{\sim}{\Sigma}_{\eta\eta}^{*-1} \underset{\sim}{\Sigma}_{\eta\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{*-1}) \underset{\sim}{\Sigma}_{\eta z}^* \qquad [42]$$

$$\underset{\sim}{\Sigma}_{\eta z} = \underset{\sim}{\Sigma}_{\eta\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{*-1} \underset{\sim}{\Sigma}_{\eta z}^* \qquad [43]$$

where we use the general notation $\underset{\sim}{\mu}_u$ for the mean vector of the random variable u and $\underset{\sim}{\Sigma}_{uv}$ for the covariance matrix of the random vectors $\underset{\sim}{u}$ and $\underset{\sim}{v}$. Quantities with asterisks refer to the distribution in the subpopulation of selectables and the corresponding quantities without asterisks to the total population. This gives

$$\underset{\sim}{\mu}_z^* = \underset{\sim}{\mu}_z + \underset{\sim}{\Sigma}_{z\eta} \overset{\cong 1}{\underset{\sim}{\Sigma}}_{\eta\eta} (\underset{\sim}{\mu}_\eta^* - \underset{\sim}{\mu}_\eta) \qquad [44]$$

$$\underset{\sim}{\Sigma}_{zz}^* = \underset{\sim}{\Sigma}_{zz} + \underset{\sim}{\Sigma}_{z\eta} \underset{\sim}{\Sigma}_{\eta\eta}^{-1} (\underset{\sim}{\Sigma}_{\eta\eta}^* - \underset{\sim}{\Sigma}_{\eta\eta}) \underset{\sim}{\Sigma}_{\eta\eta}^{-1} \underset{\sim}{\Sigma}_{\eta z} \qquad [45]$$

We note that the selection situations studied in the previous sections are of this type. We considered the conditional distribution of y and η

for given $\underset{\sim}{x}$. Due to the bivariate normality of the errors, linearity and homoscedasticity is ensured, so that the mean and variance of y, given $\underset{\sim}{x}$ and $\eta > 0$, can be obtained by equations 44 and 45.

Now consider the factor analysis model (see Lawley and Maxwell, 1971)

$$\underset{\sim}{z} = \underset{\sim}{\nu} + \underset{\sim}{\Lambda}\underset{\sim}{\xi} + \underset{\sim}{\zeta} \qquad [46]$$

where $\underset{\sim}{\nu}$ is a vector of location parameters, $\underset{\sim}{\Lambda}$ is a matrix of factor loadings, $\underset{\sim}{\xi}$ is a vector of latent factors, and $\underset{\sim}{\zeta}$ is a vector of residuals (unique variables or measurement errors). Following Sörbom (1974), let $\underset{\sim}{\mu}_{xi} = \underset{\sim}{\theta}$. With the usual assumptions.

$$\mu_z = \nu + \Lambda\theta \qquad [47]$$

$$\Sigma_{zz} = \Lambda\Phi\Lambda' + \Psi \qquad [48]$$

where $\underset{\sim}{\Phi}$ is the covariance matrix of the factors, and $\underset{\sim}{\Psi}$ is the covariance matrix of the residuals, usually assumed to be diagonal.

Assume that a set of selection variables $\underset{\sim}{\eta}$ are directly related to $\underset{\sim}{\xi}$ only, but not to $\underset{\sim}{\zeta}$ or $\underset{\sim}{z}$ ($\underset{\sim}{\eta}$ is indirectly related to $\underset{\sim}{z}$). Applying the Pearson-Lawley formulas, it is found that the factor analysis model holds in the selected population and that $\underset{\sim}{\nu}$, $\underset{\sim}{\Lambda}$, and $\underset{\sim}{\Psi}$ are unaffected by the selection (see Meredith, 1964; Olsson, 1978), and that

$$\underset{\sim}{\mu}_z^* = \underset{\sim}{\nu} + \underset{\sim}{\Lambda}\underset{\sim}{\theta}^* \qquad [49]$$

$$\underset{\sim}{\Sigma}_{zz}^* = \underset{\sim}{\Lambda}\underset{\sim}{\Phi}^*\underset{\sim}{\Lambda}' + \underset{\sim}{\Psi} \qquad [50]$$

where

$$\underset{\sim}{\theta}^* = \underset{\sim}{\theta} + \underset{\sim}{\Sigma}_{\xi\eta}\underset{\sim}{\Sigma}_{\eta\eta}^{-1}(\underset{\sim}{\mu}_\eta^* - \underset{\sim}{\mu}_\eta) \qquad [51]$$

$$\underset{\sim}{\Phi}^* = \underset{\sim}{\Phi} + \underset{\sim}{\Sigma}_{\xi\eta}\Sigma_{\eta\eta}^{-1}(\underset{\sim}{\Sigma}_{\eta\eta}^* - \underset{\sim}{\Sigma}_{\nu\nu})\Sigma_{\eta\eta}^{-1}\underset{\sim}{\Sigma}_{\eta\xi} \qquad [52]$$

Invariance properties of this kind are utilized in multiple-group factor analyses and structural equation modeling where different subpopulations are compared in a simultaneous analysis (see Jöreskog, 1971; Sörbom, 1974; Jöreskog and Sörbom, 1980; and Sörbom and Jöreskog, 1981).

6.2. SELECTION MODELING WITH LATENT EXOGENOUS VARIABLES

In Sörbom (1978, 1981; see also Sörbom and Jöreskog, 1981), the multiple-group factor analysis is extended to handle ANCOVA situations with latent variables. Of particular interest here is the case where the covariates (the exogenous variables) are imperfectly measured, assuming a factor analytic measurement model. Allowing for measurement error in the covariates avoids biased results (e.g., Reichardt, 1979, and references therein). We will consider a simple case of this type of model and introduce the added complication of selective samples.

Consider the following model for groups g = 1, 2, . . . , G,

$$y^{(g)} = \beta_0^{(g)} + \underset{\sim}{\beta}^{(g)\prime}\underset{\sim}{\xi}^{(g)} + \epsilon^{(g)} \qquad [53]$$

$$\eta^{(g)} = \gamma_0^{(g)} + \underset{\sim}{\gamma}^{(g)\prime}\underset{\sim}{\xi}^{(g)} + \delta^{(g)} \qquad [54]$$

$$\underset{\sim}{x}^{(g)} = \underset{\sim}{\nu} + \underset{\sim}{\Lambda}\underset{\sim}{\xi}^{(g)} + \underset{\sim}{\zeta}^{(g)} \qquad [55]$$

$$y^{(g)}, \underset{\sim}{x}^{(g)}: \begin{cases} \text{observed, if } \eta^{(g)} > 0 \\ \text{not observed, otherwise} \end{cases} \qquad [56]$$

where $\epsilon^{(g)}$ and $\delta^{(g)}$ have covariance $\sigma_{\delta\epsilon}^{(g)}$, and both $\epsilon^{(g)}$ and $\delta^{(g)}$ are assumed to be independent of $\xi^{(g)}$ and $\zeta^{(g)}$. Here, equation 55 shares the assumptions of equation 46. Note that $\underset{\sim}{x}^{(g)}$ is included in equation 56, since in this case the model also restricts the marginal distribution of $\underset{\sim}{x}^{(g)}$, so that we consider the joint $y^{(g)}$, $\underset{\sim}{x}^{(g)}$ distribution, not only the conditional distribution of $y^{(g)}$ given $\underset{\sim}{x}^{(g)}$, as before. Here we consider the truncated case only.

With $\sigma_{\delta\epsilon}^{(g)} = 0$ for g = 1, 2, . . . , G, relation (equation 54) is inconsequential and we obtain a special case of Sörbom (1978). The measurement model (equation 55) is assumed to have group-invariant $\underset{\sim}{\nu}$ and $\underset{\sim}{\Lambda}$. The groups may consist of a control group and several treatment groups. The latent variable vector $\underset{\sim}{\xi}^{(g)}$ contains the covariates. Given group-invariant slope parameters $\underset{\sim}{\beta}^{(g)}$, treatment effects are obtained as differences in the $\beta_0^{(g)}$ (g = 1, 2, . . . , G) parameters.

Now consider the full model given by equations 53 through 56 for each group g. Note that the specification allows separate exogenous variables to operate in equations 53 and 54, and that those in equation

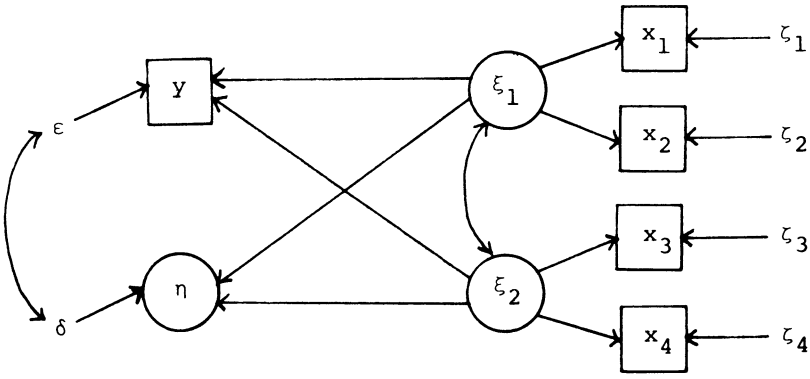


Figure 6: Path Model for Selection in the Presence of Two Latent Exogenous Variables

54 can be specified to be directly observed variables measured without error. For each group g, there is selectivity in the structural regression (equation 53) of $y^{(g)}$ on $\xi^{(g)}$ whenever the error covariance $\sigma^{(g)}_{\delta\epsilon}$ is nonzero.

We will illustrate the selectivity issues by means of the model of Figure 6, where squares denote observed variables and circles denote latent variables.

To continue the example of Section 3, y and ξ_1 represent achievement post- and pretest scores, where for simplicity the posttest is taken to be measured without error, while ξ_1 is the pretest true score. (Random measurement error in y causes no bias since it is absorbed in the residual ϵ.) Here, ξ_2 may represent, say, true home background score. True scores rather than actually observed scores are assumed to influence the selection variable η. Assume that the prerequisites for the Pearson-Lawley formulas hold in a certain population, for example, by multivariate normality for all variables involved. Consider the analysis of a random sample from the subpopulation $\eta > 0$, that is, a selective, truncated, sample from the full population. We note that by the Pearson-Lawley results, the factor analysis model (the measurement model) for the marginal distribution of x_1, x_2, x_3, x_4 in this subpopulation will hold with invariant $\underset{\sim}{\nu}$, $\underset{\sim}{\Lambda}$, and $\underset{\sim}{\Psi}$, since η is indirectly related to $x_1 - x_4$ via ξ_1, ξ_2. In the distribution of y, x_1, x_2, x_3, x_4, we note that y can be considered as an additional measure of both ξ_1 and ξ_2 in a five-variate factor analysis model. Since the selection variable is directly related to y, this model will not hold in the subpopulation of

selectables. Estimating the structural regression of y on ξ_1 and ξ_2 by the methods of Sörbom (1978), Sörbom and Jöreskog (1981) gives biased results in a way analogous to previous sections.

A simple ad hoc estimator using the methods of Section 4 seems possible, however. Since the measurement model for $x_1 - x_4$ holds in the subpopulation, we may use the truncated sample to estimate factor scores $\hat{\xi}_1$ and $\hat{\xi}_2$, properly scaled to approximate the covariance matrix Φ^* (see Lawley and Maxwell, 1971: ch. 8). In a second step, y and η are regressed on $\hat{\xi}_1$ and $\hat{\xi}_2$ according to the truncated case of the selection model in Section 4. This enables us to obtain an approximate test for selectivity and approximately estimate the full population regression coefficients.

If, however, any of the observed exogenous variables $x_1 - x_4$ influences η directly (see also Goldberger, 1972a), the measurement model will also be incorrectly specified. Hence, ignoring selectivity may give seriously biased results from analyses by the methods of Sörbom (1978), Sörbom and Jöreskog (1981).

With the model of equations 53 through 56, any of the above selection situations can be specified.

7. CONCLUSION

In this article we have shown the statistical and computational feasibility of correctly analyzing selective samples by selection modeling. Analysis methods originally proposed in econometric studies have been shown to be of potential use in general quasi-experimental studies, particularly regarding selectivity in the context of treatment effect evaluation with nonequivalent groups. In such evaluations, randomization is seen not to be essential to unbiased treatment effect estimation.

> Indeed, the critical difference for avoiding bias is not whether the assignments are random or nonrandom, but whether the investigator has *knowledge of and can model* this selection process [Cain, 1975: 304].

This suggests that the investigator should gather extensive information on the selection processes involved, and seek to be in control of

them by systematic selection in a consistent manner. For instance, in the context of our model, we have seen the benefits of using censored sample information rather than truncated sample information. Say that nonrandom selection into the experimental (treatment) group is desirable from an ethical point of view. With the language of our model we may take an initial random sample for which $\underset{\sim}{x}$ is observed. From this sample, units can be selected in a nonrandom but consistent way, resulting in censored sample information. Indeed, we may consider the selectivity problem in the opposite way. Instead of trying to *find* the correct selection model, we could select *according to* a prescribed model, attempting to determine by design the selection variable η in terms of a set of background variables $\underset{\sim}{x}$. Of course, the selectivity problem of attrition will remain.

Admittedly, the selection modeling of this article may certainly be an oversimplification for many practical quasi-experimental situations. Selectivity problems, however, seem unavoidable by design, implying that more flexible statistical specifications should be investigated.

REFERENCES

AMEMIYA, T. (1982) Tobit Models: A Survey. Palo Alto, CA: Rhodes Associates.

———(1973) "Regression analysis when the dependent variable is truncated normal." Econometrica 41, 6: 997-1016.

BARNOW, B. S., G. G. CAIN, and A. S. GOLDBERGER (1980) "Issues in the analysis of selection bias, in E. Stromsdorfer and G. Farkas (eds.) Evaluation Studies Review Annual, Vol. 5. Beverly Hills, CA: Sage.

BERNDT, E. B., B. HALL, R. HALL, and J. A. HAUSMAN (1974) "Estimation and inference in non-linear structural models." Annals of Economic and Social Measurement 3: 653-656.

CAIN, G. C. (1975) "Regression and selection models to improve nonexperimental comparisons," pp. 297-317 in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment, Some Critical Issues in Assessing Social Programs. New York: Academic Press.

CRAGG, J. G. (1971) "Some statistical models for limited dependent variables with application to the demand for durable goods." Econometrica 39: 829-844.

GOLDBERGER, A. S. (1980) "Abnormal selection bias." Social Systems Research Institute, University of Wisconsin, Madison.

———(1979) "Methods for eliminating selection bias." Department of Economics, University of Wisconsin, Madison.

———(1972a) "Selection bias in evaluating treatment effects: some formal illustrations." Discussion paper 123-72. Madison, WI: Institute for Research on Poverty.

———(1972b) "Selection bias in evaluating treatment effects: the case of interaction." Discussion paper 129-72. Madison, WI: Institute for Research on Poverty.

GRILICHES, Z., B. H. HALL, and J. A. HAUSMAN (1978) "Missing data and self-selection in large panels." Annales de l'INSEE 30-31: 137-176.

GRONAU, R. (1974) "Wage comparisons—a selectivity bias." J. of Pol. Economy 82: 1119-1144.

GRUVAEUS, G. T. and K. G. JÖRESKOG (1970) "A computer program for minimizing a function of several variables." Research Bulletin 70-14. Princeton, NJ: Educational Testing Service.

HAUSMAN, J. A. and D. A. WISE (1979) "Attrition bias in experimental and panel data: the Gary income maintenance experiment." Econometrica 47: 455-474.

———(1976) "The evaluation of results from truncated samples: the New Jersey income maintenance experiment." Annals of Economic and Social Measurement 5: 421-445.

HECKMAN, J. (1979) "Sample selection bias as a specification error." Econometrica 47: 153-161.

———(1977) "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)." University of Chicago. (mimeo)

———(1974) "Shadow prices, market wages, and labor supply." Econometrica 42: 679-694.

HURD, M. (1979) "Estimation in truncated samples when there is heteroscedasticity." J. of Econometrics 11: 247-258.

JOHNSON, N. and S. KOTZ (1972) Distributions in Statistics: Continuous Multivariate Distributions. New York: John Wiley.

JÖRESKOG, K. G. (1977) "Structural equation models in the social sciences: specification, estimation and testing," pp. 265-286 in P. R. Krishnaiah (ed.) Applications of Statistics. Amsterdam: North-Holland.

———(1971) "Simultaneous factor analysis in several populations." Psychometrika 36: 409-426.

———and D. SÖRBOM (1980) Simultaneous Analysis of Longitudinal Data from Several Cohorts. Research Report 80-5. Department of Statistics, University of Uppsala.

KENNY, L. W., L. LEE, G. S. MADDALA, and R. P. TROST (1979) "Returns to college education: an investigation of self-selection bias based on the Project Talent data." Int. Econ. Rev. 20, 3: 775-789.

LAWLEY, D. N. (1943-1944) "A note on Karl Pearson's selection formulae." Proceedings of the Royal Society Edinburgh, Section A (Mathematics and Physics Section) 62, 1: 28-30.

——and A. E. MAXWELL (1971) Factor Analysis as a Statistical Method. London: Butterworth.

LEWIS, H. G. (1974) "Comments on selectivity biases in wage comparisons." J. of Pol. Economy: 1145-1155.

MADDALA, G. S. and L-F. LEE (1976) "Recursive models with qualitative endogenous variables." Annals of Economic and Social Measurement 5: 525-545.

MEREDITH, W. (1964) "Notes on factorial invariance." Psychometrika 29: 177-185.

NELSON, F. D. (1979) "The effect of and a test for misspecification in the censored-normal model." Social Science Working Paper 291, California Institute of Technology.

OLSEN, R. J. (1979) "Tests for the presence of selectivity bias and their relation to specifications of functional form and error distribution." Working paper 812, Yale University.

OLSSON, U. (1978) "Selection bias in confirmatory factor analysis." Research Report 78-4. Department of Statistics, University of Uppsala.

PEARSON, K. (1912) "On the general theory of the influence of selection on correlation and variation." Biometrika 8: 437-443.

RAY, S. C., R. A. BERK, and W. T. BIELBY (1980) "Correcting sample selection bias for bivariate logistic distribution of disturbances." University of California.

REICHARDT, C. S. (1979) "The statistical analysis of data from non-equivalent group designs," in T. D. Cook and D. T. Campbell (eds.) Quasi-Experimentation: Design & Analysis for Field Settings. Chicago: Rand McNally.

SÖRBOM, D. (1981) "Structural equation models with structured means," in K. G. Jöreskog and H. Wold (eds.) Systems Under Indirect Observation: Causality, Structure, Prediction. Amsterdam: North-Holland.

——— (1978) "An alternative to the methodology for analysis of covariance." Psychometrika 43: 381-396.

——— (1974) "A general method for studying differences in factor means and factor structures between groups." British J. of Mathematical and Statistical Psychology 27: 229-239.

——— and K. G. JÖRESKOG (1981) "The use of structural equation models in evaluation research." Presented at the conference on Experimental Research in Social Sciences, Gainesville, Florida, January 8-10.

STROMSDORFER, E. and G. FARKAS (1980) Evaluation Studies Review Annual, Vol. 5. Beverly Hills, CA: Sage.

THORNDIKE, R. L. (1942) "Regression fallacies in the matched groups experiment." Psychometrika 7: 85-102.

TOBIN, J. (1958) "Estimation of relationships for limited dependent variables." Econometrica 26: 24-36.

TUNALI, F. I., J. R. BEHRMAN, and B. L. WOLFE (1980) "Identification, estimation and prediction under double selection." Presented at the 1980 Joint Statistical Meetings of American Statistical Association and Biometric Society, Houston, Texas.

VENTI, S. and D. A. WISE (1980) "Test scores, educational opportunities, and individual choice." Discussion Paper Series, Kennedy School of Government, Harvard University.

WALES, T. J. and A. D. WOODLAND (1980) "Sample selectivity and the estimation of labor supply functions." Int. Econ. Rev. 21, 2: 437-468.

WEISBERG, H. I. (1979) "Statistical adjustments and uncontrolled studies." Psych. Bull. 86: 1149-1164.

WILLIS, R. J. and S. ROSEN (1979) "Education and self-selection." J. of Pol. Economy 87, 5: 7-36.

27

An Introduction to Sample Selection Bias in Sociological Data

Richard A. Berk

Sampling has long been central in discussions of sociological research methods. Yet, with a few exceptions (e.g., Tuma et al., 1979; Rossi et al., 1980; Berk et al., 1981), recent developments on the nature of sampling bias have not filtered into sociological practice. This neglect represents a major oversight with potentially dramatic consequences. More than external validity is threatened. Internal validity is equally vulnerable even if statements are made conditional upon the available data.

This paper undertakes a brief review of recent advances in the diagnosis of and corrections for "sample selection bias." Key points are illustrated with analyses taken from real data sets. Thus, the paper is no substitute for a careful reading of the primary source material and recent more lengthy, technical overviews. My goal is to direct the attention of the sociological community to a significant methodological problem while stressing major themes and intuitive reasoning.

* Direct all correspondence to: Richard A. Berk, Department of Sociology, University of California, Santa Barbara, CA 93106.

The research reported in this paper was supported by a grant from the National Institute of Justice (grant No. 80-IJ-CX-0037). I am also grateful for the help in data collection provided by Anthony Shih and Jimy Sanders. Finally, Karl Schuessler, Kenneth Land, and Phyllis Newton provided helpful comments on an earlier draft of the paper.

WHAT'S THE PROBLEM?

Sample selection bias can be intuitively understood through the usual bivariate scatter plot interpreted within the framework of the general linear model. Given a fixed regressor (gener-

From Richard A. Berk, "An Introduction to Sample Selection Bias in Sociological Data," *American Sociological Review*, 1983, 48, 386-398.

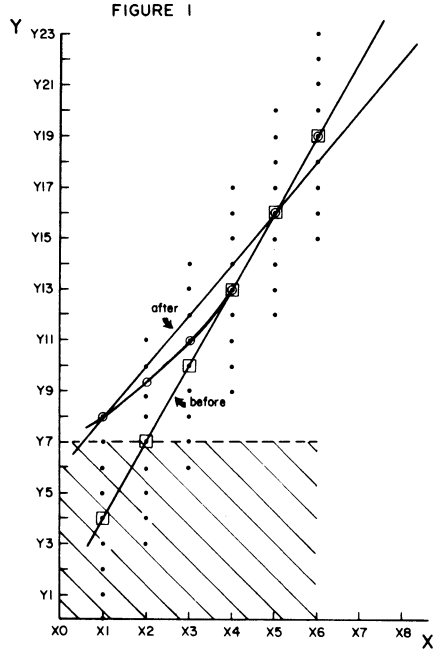
alizations to stochastic regressors are easily accomplished, e.g., Pindyck and Rubinfeld, 1981:274–78), one assumes a linear relationship between an exogenous and an endogenous variable. One also assumes that the endogenous variable is affected additively by a disturbance (error) term characterized by an expected value of zero for each value of the exogenous variable.[1] If these two assumptions are met, the disturbance term is uncorrelated with the exogenous variable, which guarantees unbiased least squares estimates of the slope and intercept. Other assumptions about the disturbance term that are typically made need not concern us.[2]

Figure 1 is a scatter plot for an endogenous variable Y and an exogenous variable X. Assume the data are a simple random sample from some population of interest, that in this population the linear form is correct, and that for each value of X the mean of the disturbances is zero. Implied is that the regression line passes through the expected value of Y for each value of X. In Figure 1 these expected values are represented by boxes, and the regression line is labeled "before."

In Figure 1, suppose that observations with values on Y equal to or less than Y7 cannot be obtained. For example, suppose that Y is a measure of the seriousness of incidents of wife battery, and that police only make an arrest in such incidents if the dispute exceeds some level of seriousness (Berk et al., 1983). Then, if one's data are taken exclusively from police arrest reports, less serious incidents will be systematically underrepresented. In Figure 1 observations in the shaded area are missing.

For low values of X in Figure 1 the new expected values are represented by circles. Thus, for all observations with X equal to X1, the expected value of Y has shifted from Y4 to Y8. Likewise, for all observations with X equal to X2, the expected value of Y has shifted from Y7 to Y9.5. As X increases, the size of the shift is reduced until by X4, the new and old expected values are virtually identical.

FIGURE 1



The new expected values for Y means that the original regression line no longer fits the data. The relationship between X and Y is no longer linear; the slope becomes steeper as X increases (up to X4). Consequently, *any* attempt to fit a straight line will produce a specification error. Basically, one is using the wrong functional form. In Figure 1 the second regression line labeled "after" shows the result that might materialize. Compared to the true relationship, the estimated relationship has been attenuated.

What are the implications? First, external validity has been undermined. The regression line estimated from the scatter plot in Figure 1 will systematically underestimate the slope of the population regression line. If X is the number of prior wife battery incidents, the estimated causal effect of such priors on the seriousness of the immediate incident will be substantially smaller than the causal effect in the population. Excluding less serious incidents

[1] Actually, one assumes that for each observation, the expectation of the disturbance term is zero; this implies that the expectations for each value of the exogenous variable are zero. The assumption of linearity allows for nonlinear relationships that can be transformed into linear ones (e.g., Pindyck and Rubinfeld, 1981:107–110). With time series data, one sometimes makes a distinction between exogenous variables and predetermined variables. All regressors are predetermined, including lagged values of the endogenous variable. However, lagged values of the endogenous variable are not exogenous. A further discussion of such issues can be found in Engle et al. (1983).

[2] In order to obtain efficient estimates of the regression coefficients and unbiased estimates of their standard errors, one must assume that the disturbances are uncorrelated with one another and that all have the same variance. Then in "small" samples, one must assume for significance tests that the disturbances are normally distributed. Asymptotically, the normality assumption is unnecessary. Discussion of the assumptions for least squares procedures can be found in virtually any econometrics text.

attenuates the causal effect in this instance. Clearly, one should not try to generalize from the sample in Figure 1 to all incidents of wife battery. Such problems are well understood by most sociologists.

Second, and not commonly recognized, internal validity is also jeopardized *even if one is prepared to make causal inferences to a population of less serious battery incidents*. In Figure 1, for low values of X, the regression line falls on or above the expected values, while for high values of X, the regression line falls on or below the expected values. For low values of X, therefore, negative disturbances will predominate, while for high values of X, positive disturbances will predominate. This implies that X will be positively correlated with the disturbance term. As a result, least squares estimates of the slope and intercept will be biased (and inconsistent as well), even if one is only interested in the causal relationship between the seriousness of the incident and the number of priors for the subset of more serious incidents. Put another way, effects of the exogenous variable and the disturbance term are confounded, and causal effects are attributed to X that are really a product of random perturbations.

The confounding of X and the disturbance term follows in this example *even if one's sole concern is with more serious wife battery incidents*. One cannot dismiss the problem by claiming interest only in the nonrandom subset of cases represented by the sample at hand. By excluding some observations in a systematic manner, one has inadvertantly introduced the need for an additional regressor that the usual least squares procedures ignore (Heckman, 1976, 1979); in effect, one has produced the traditional specification error that results when an omitted regressor is correlated with an included regressor (e.g., Kmenta, 1971:392–95).

Figures 3 through 5 present in schematic fashion other examples of outcomes obtained when segments of some population cannot be observed. Figure 2 is a new representation of Figure 1 and serves as a benchmark.

Suppose in Figure 3 that Y is income and X is education and that the sample only includes individuals with income below the poverty line. The estimated regression line is again biased downward with both external validity and internal validity weakened. One cannot generalize the estimated causal relationship to all adults nor is the relationship between education and income properly represented, *even for individuals with incomes below the poverty line*.

Both Figure 2 and 3 depict exclusion through a threshold for the endogenous variable under scrutiny. Goldberger (1981), borrowing from Lord and Novick (1968), has called this manner of selection "explicit." Alternatively, one might use the term "direct" for reasons that will be apparent shortly.[3]

Figure 4 shows a more complicated selection process. The lower right-hand section of the scatter plot has been eliminated, but not in a way that reflects a single threshold on Y. How might this happen? Suppose that Y is the amount of money spent on medical care, and X is the amount a person smokes. Also suppose that people who smoke more are more likely to have fatal illnesses, other things being equal. Clearly, one cannot observe the amount of money spent on medical care for individuals who are no longer alive.

It is important to stress that no threshold is defined in terms of medical costs or even the amount of smoking. Rather, the threshold involves a new variable, physiological viability, that, for purposes of illustration, has been assumed not to play a role in the relationship of interest (i.e., the effect of smoking on medical costs). When physiological viability falls below the threshold of death, the case is excluded. Goldberger, again drawing from Lord and Novick, has called such selection processes "incidental." Alternatively, one might use the term "indirect."

As before, both external validity and internal validity are jeopardized. Once again, the exclusion of a nonrandom subset of observations introduces a nonlinear relationship between X and Y. When, in this instance, a straight line is fitted, the estimated causal relationship is inflated, and effects attributed to X include the impact of the disturbance term.

Consider a second example. Suppose that Y is length of time retail stores remain in business, and X is amount of capital the stores had when they opened. Suppose also that in 1970 one obtains a random sample of retail stores just opening for business and that data are collected until 1980. However, not all stores fail in the ten-year interval; for some fraction of the cases the time to failure cannot be observed. Such incidental selection is called right-hand censoring in the failure time literature (e.g., Lawless, 1982; Tuma, 1982) and can lead to distorted scatter plots as in Figure 4. Less common, left-hand censoring is also possible

[3] Explicit selection can be generalized so that the threshold is not a constant (Goldberger, 1981), but the generalization has not had a substantial impact on empirical work. There seems no need, therefore, to complicate the discussion with variable thresholds.

FIGURE 2

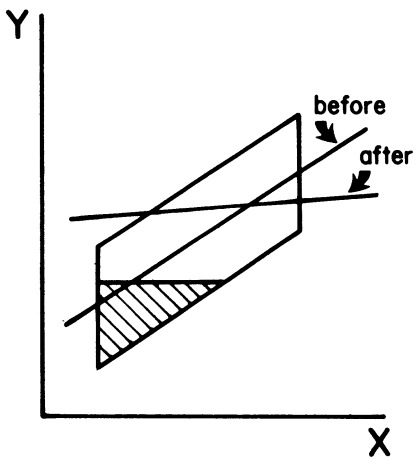


FIGURE 3

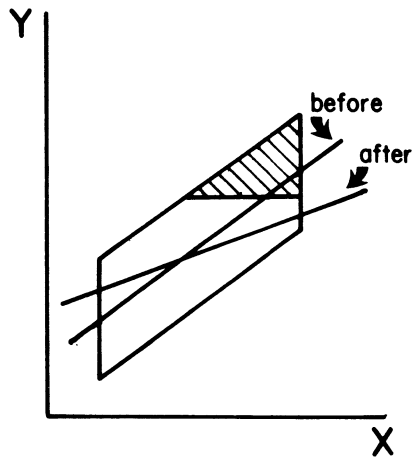


FIGURE 4

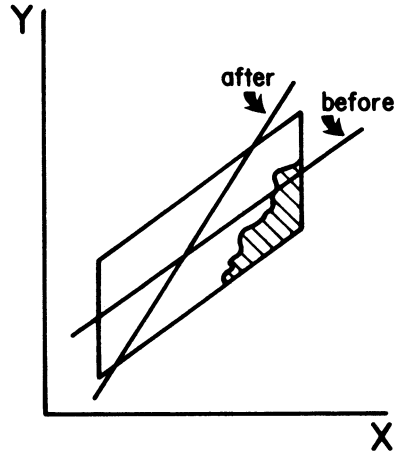
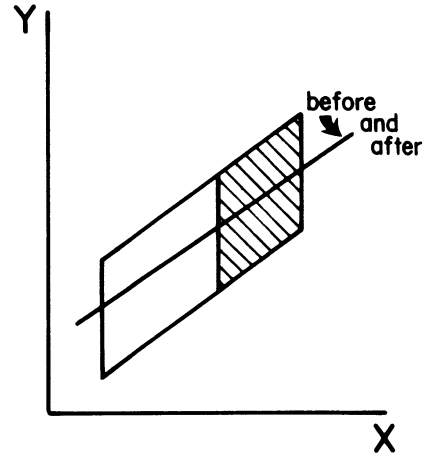


FIGURE 5



(i.e., the data collection begins after some units have failed).

Figure 5 shows a pattern in which a threshold for exclusion is defined for the exogenous variable. Suppose that people of all incomes are included, but people with greater than a high school education are not. If the relationship between education and income is really linear across the full range of educational levels, external validity and internal validity are unscathed.[4] One can generalize to a population that includes individuals with more than a high school education and estimate the effect of education in an unbiased manner.

There are, thus, three initial lessons to be

[4] The danger is that by excluding observations one has a smaller sample and perhaps less variation in X. Both reduce one's statistical power; standard errors will be increased.

learned. First, if potential observations from some population of interest are excluded from a sample on a nonrandom basis, one risks sample selection bias. Nonrandom exclusion of certain observations can be caused by data collection procedures or by processes inherent in the phenomena under study. For example, skip patterns are meant to weed out nonrandom subsets of respondents for whom some questions do not apply. Such procedures risk sample selection bias when the remaining (nonrandom) observations are analyzed. In this situation, a researcher's data collection procedures recapitulate nonrandom selection in the social world. The general point is that the prospect for sample selection bias is pervasive in sociological data. Circumstances under which the prospect becomes a reality will be addressed below.

Second, it is difficult to anticipate whether the biased regression estimates overstate or understate the true causal effects. The direction and size of the bias depends in the bivariate case on the number and location of observations that are excluded; the situation is enormously more complicated in multivariate models. When sample selection bias is present, one is essentially flying blind. One is faced with the same kinds of problems one finds in multiple regression analyses with conventional specification or measurement errors. Only in special cases[5] can the direction of the distortions be known.

Third, the problems caused by nonrandom exclusion of certain observations are manifested in the expected values of the endogenous variable. When the usual linear form is fit to the data, the expected values of the disturbances for each value of X are no longer zero. The bad news is that the disturbances are then correlated with the exogenous variable. The good news is that in the nonlinear form lies a potential solution.

A MORE FORMAL STATEMENT OF THE PROBLEM

The social science literature contains several formal introductions to sample selection bias (e.g., Heckman, 1976; Goldberger, 1981) and several textbook-level discussions (e.g., Judge et al., 1980: ch. 14; Berk and Ray, 1982). Probably the best known and most accessible formulation is by Heckman (1979). I have drawn heavily on his exposition.[6]

Consider a random sample of I observations with two equations of interest:

$$Y_{1i} = X_{1i}\beta_1 + U_{1i} \quad (1a)$$

$$Y_{2i} = X_{2i}\beta_2 + U_{2i} \quad (i = 1, \ldots, I), \quad (1b)$$

where each X is a vector of exogenous variables which may, or may not, be the same, and the betas are vectors of conformable regression coefficients. In both equations, the expected values of the disturbances are taken to be zero, which implies that both equations are properly specified. More generally, each equation by itself is assumed to meet the usual assumptions for ordinary least squares. Across equations, however, the disturbances are correlated and are assumed to behave as if drawn from a bivariate normal distribution. Thus, equations 1a and 1b represent a pair of seemingly unrelated equations (e.g., Pindyck and Rubinfeld, 1981:323–24). One has nothing more than a pair of regression equations with correlated disturbances.

Suppose that on sociological grounds one cares about the first equation; equation 1a can be thought of as the "substantive equation." However, one can only observe the endogenous variable in that equation if the endogenous variable in the second equation exceeds (or does not exceed) some threshold. The second equation can be called the "selection equation."

To make this more concrete, suppose that the first equation is a causal model of the length of prison sentences given to convicted felons. Yet, convictions can only result if the strength of the evidence implies guilt beyond a reasonable doubt; one can think of the second equation as a causal model for the strength of evidence. Individuals for whom reasonable doubt exists are excluded from sentencing. Since the same parties are involved in both the determination of guilt and the determination of sentence length, the disturbances in the two equations are plausibly correlated. That is, random perturbations (e.g., how aggressive the prosecutor is) will simultaneously affect both endogenous variables.

---

[5] Goldberger (1981) discusses the situations in which the direction and size of the bias can be determined. If one can assume the data come from a multivariate normal distribution, then in the case of explicit selection, all regression coefficients are attenuated. For incidental selection, even under the assumption of multivariate normality, the direction and size of the bias cannot be determined.

[6] The problem has a long history. Pearson and Lee (1908) wrestled with truncated distributions, the econometrics community was first introduced to the problem of explicit selection by James Tobin (1958), and biometricians have worried about left-hand and right-hand censoring at least as long (e.g., Lawless, 1982:34–44).

Equations 2a and 2b show the results of the selection process.

$$E(Y_{1i} \mid X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i \qquad (2a)$$

$$E(Y_{2i} \mid X_{2i}, Y_{2i} \geq 0) = X_{2i}\beta_2 + \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i \qquad (2b)$$

For equation 2a the conditional expectation of the endogenous variable is equal to the expected value of the original substantive equation (1a) plus a new term. For equation 2b the conditional expectation of the endogenous variable is equal to the expected value of the original selection equation (1b) plus a somewhat different new term. Focusing first on the substantive equation 2a (e.g., the equation for sentence length), the new term can be divided into two parts. The first part is the ratio of the covariance between the disturbances in equations 1a and 1b to the standard deviation of the disturbances in equation 2a. The ratio, therefore, serves as a regression coefficient; if the covariance between the two disturbances is zero, *the extra term disappears*. If the disturbances are uncorrelated, the usual least squares procedures will suffice.

The meaning of the second component can be understood through the following equations:

$$\lambda_i = \frac{f(z_i)}{1-F(z_i)} \qquad (3)$$

$$z_i = -\frac{X_{2i}\beta_2}{(\sigma_{22})^{1/2}} \qquad (4)$$

The z in equation 4 is the negative of the predicted value from a probit equation in which one models the likelihood that in the selection equation (1b) the threshold will be equaled or exceeded. In our example, the probit equation models the likelihood that a conviction will occur. However, since the predicted value is multiplied by -1, one is ultimately capturing the likelihood that a conviction will not occur; the issue is really which cases will be *excluded*.

More specifically, the predicted value from a probit equation is a normally distributed, random variable with a mean of zero and a standard deviation of 1.0. The negative of this random variable is then used in equation 3, where the numerator is the variable's density, and the denominator is 1.0 minus the variable's (cumulative) distribution. The ratio is called the hazard rate, which represents for each observation the instantaneous probability of being excluded from the sample conditional upon being in the pool at risk (Tuma, 1982:8–10). The larger the hazard rate, the greater the likelihood that the observation will be discarded.

Equally important, the hazard rate captures the expected values of the disturbances in the substantive equation after the nonrandom selection has occurred. It was precisely these expected values that are the source of the biased estimates. By including the hazard rate as an additional variable, one is necessarily controlling for these nonzero expectations. Alternatively stated, the deviations of the expected values from the regression line result from an omitted variable that has now been included. The key, then, to consistent parameter estimates is to construct a hazard rate for each observation. And it cannot be overemphasized that it is the selection process that *introduces* the need for a new variable.

Turning to equation 2b, the hazard rate is constructed from the same equation in which it is then used; the distinction between the selection and substantive equations disappears. Referring back to our earlier terminology, the two-equation model (equations 2a and 2b) represents incidental or indirect selection. The one-equation model (equation 2b) represents explicit or direct selection. The latter is also known as a Tobit Model (Tobin, 1958).[7]

To summarize, whenever one has a nonrandom sample, the potential for sample selection bias exists. Examples are easy to construct. Studies of classroom performance of college students rest on the nonrandom subset of students admitted and remaining in school. Studies of marital satisfaction are based on the nonrandom subset of individuals married when the data are collected. Studies of worker productivity are limited to the employed. And, potential problems are complicated by inadequate response rates.

Alternatively stated, the difficulty is that one risks confounding the substantive phenomenon of interest with the selection process. The impact of a mother's level of education on a child's college grade point average may be

[7] When the selection process eliminates observations solely for the endogenous variable, one commonly speaks of censoring. When observations are missing in the exogenous variables as well, one commonly speaks of truncation (Heckman, 1976:478). Here, only censored samples are considered. Truncation causes far more serious difficulties that are well beyond the scope of this paper. An introduction to the issues and a good bibliography can be found in Berk and Ray (1982). It is also important not to confuse sample selection censoring or truncation with legitimately bounded endogenous variables where no observations are lost. For example, analyses of some kinds of survey questions must respond to ceiling and floor effects and, in a sense, these effects truncate the endogenous variable. However, floor and ceiling effects imply a nonlinear functional form (e.g., a logistic) and not a failure to observe certain values on the endogenous variable.

confounded with its impact on the child's likelihood of getting into college. The impact of a husband's income on the amount of leisure time a couple shares may be confounded with its impact on the likelihood that the couple will be married at all. The impact of seniority on output per hour may be confounded with its impact on the likelihood of being employed. Finally, the impact of a respondent's race on any of these phenomena may be confounded with its impact on the likelihood of responding to a questionnaire.

There is also the problem of infinite regress. Even if one has a random sample from a defined population, that population is almost certainly a nonrandom subset from a more general population. Suppose one has a random sample of all felony arrests in a given state in a given year. The random sample of felony arrests is a nonrandom sample of all reported felonies in that state in a given year. The reported felonies are a nonrandom sample of actual felonies committed. The felony arrests in a given state are also not a random sample of felonies in all states. In principle, therefore, there exists an almost infinite regress for any data set in which at some point sample selection bias becomes a *potential* problem. As for traditional specification errors and measurement errors, the question is not typically whether one has biased (or even consistent) estimates.[8] The question is whether the bias is small enough to be safely ignored.

Given the almost universal potential for sample selection bias, the critical issue becomes when that bias is likely to materialize. The key lies in the correlation between the disturbances for the substantive and selection processes. Under explicit selection, the substantive and selection processes are captured in a single equation. The two disturbance terms are, therefore, identical and correlate perfectly. Thus, *any* nonrandom (explicit) selection produces biased and inconsistent estimates of the regression coefficients, with the bias a function of the proportion of the sample excluded. If one is prepared to assume that the data (exogenous and endogenous variables) are drawn from a multivariate normal distribution, the bias is proportional to the probability of exclusion (Goldberger, 1981). And the probability can be estimated from the proportion of cases for which no observations on the endogenous variable are available. Explicit selection seems to be relatively rare in sociological data.

The situation for incidental selection is more complicated. One rarely knows much about the likely sign and magnitude of the correlation between the disturbances. Perhaps the easiest case is found when one can point to an obvious variable omitted from both the substantive and selection equations that is also *uncorrelated with the regressors included*. The omitted variable will cause the disturbances to be correlated, but since the omitted variable is uncorrelated with the included regressors, the equations (prior to sample selection) are properly specified (under the usual definition of specification error).

For example, one might be interested in victimization from natural disasters such as tornadoes, floods, and earthquakes (Rossi et al., 1982). Suppose that questionnaires are given to a random sample of adults and that response rates are virtually 100 percent. In an analysis of the amount of damage done in "the most recent" disaster, a large number of respondents would have nothing to report. Indeed, skip patterns in the questionnaire are designed to spare them from such items.

Almost regardless of how one conceives the substantive and selection processes, the severity of the natural disaster to which respondents were exposed (from no experience to a devastating one) should affect the likelihood of reporting a firsthand experience and also the amount of damage that resulted. However, if no external measure of disaster severity is available, no external measure of severity can be included in either equation. Should that omitted variable be correlated with regressors that are included, one has the traditional omitted variable specification error. If, however, one can argue that the severity of the natural disaster is probably uncorrelated with the included regressors, one can alternatively assert that sample selection bias will be present when data from the subset of disaster victims are analyzed. Given the processes that determine the location and magnitude of tornadoes, for example, the arguments for sample selection bias (rather than traditional omitted variable bias) may well be plausible. For example, it is unlikely that the probability of damage from a tornado is related to education, income, or attitudes toward risk.

In most sociological research, the issues are muddier. One must first justify the model specifications for the substantive and selection equations (no small feat) and then carefully address whether the disturbances are likely to be correlated. There are probably grounds for concern when the substantive and selection processes unfold with the same actors, and/or in the same physical locations, and/or at about the same time. Under these conditions, random perturbations will have a significant opportunity to affect jointly the selection and

[8] Randomized experiments come the closest to eliminating such problems.

substantive outcomes. The sentencing example is surely a good illustration. Studies of the wages earned by women are among the best known examples in the economics literature. One can only observe wages for women who are employed, and employed women are a nonrandom subset of all women. Moreover, random perturbations are likely to affect simultaneously both the probability of getting a job and wages once the job begins (Heckman, 1980). More generally, however, the social science community still has very limited experience with the sample selection problem, and there are as yet no compelling guidelines.

APPLICATIONS

Examples of corrections for explicit selection are readily found elsewhere, often under the rubric of Tobit Models (Tobin, 1958; Greene, 1981; Berk et al., 1983). In the pages ahead, analyses will be presented in which incidental selection is at issue.

In the analyses to be discussed shortly, the following steps are followed:

1. A probit model of the selection process is estimated with the dummy endogenous variable coded "0" when the observation on the substantive endogenous variable is missing and "1" when it was present.
2. The predicted values from the probit equation are saved. These predicted values represent a random, normal variable.
3. From the predicted values, the hazard rate is constructed. The predicted values are first multiplied by -1.0, and the density and distribution values calculated. The results are plugged into equation 3.
4. The hazard rate is then treated as a new variable and included in any substantive equations.
5. The bulk of the substantive analyses are done with ordinary least squares, although spot checking with other procedures (e.g., generalized least squares) is also undertaken.

The data are from a study of citizen opinions of various parts of the criminal justice system. For each of four county criminal justice agencies (a Police Department, The Office of the Court Administrator, The Public Defender's Office, and a Victim/Witness Assistance Program in the District Attorney's Office), self-administered questionnaires were mailed to random samples of individuals shortly after these individuals had an encounter with the agency in question. Here, we will rely exclusively on material from people who were called for jury duty. Overall, the problem was an effort to determine if accurate and cost-effective ways could be developed to provide rapid citizen feedback on the performance of the criminal justice system (Berk and Shih, 1982).

We anticipated low response rates. Therefore, we collected from official records considerable information on *all* prospective respondents, expecting to model failures to return the questionnaire.

Table 1 shows the three selection equations for nonresponse. The results on the far right derive from a probit model which rests on what we have been assuming so far: the two disturbances are bivariate normal. If one is prepared to assume that the disturbances are bivariate logistic, then the selection equation should be logistic (Ray et al., 1980). Finally, if one is prepared to assume that the disturbances in the selection equation follow a rectangular distribution and that the disturbances in the substantive equation are a linear function of the disturbances in the selection equation, the linear probability model may be used to model selection (Olsen, 1980b).

The probit approach is by far the most popular, and we will continue to rely on it. However, there is some concern in the literature about what happens if bivariate normality is violated, including what the appropriate options may be (Olsen, 1980a,b; Greene 1981; Arabmazar and Schmidt, 1982). The three sets of results are presented to stress that there are options to the assumption of bivariate normality, that these options are easy to implement, and to consider whether in this instance the results depend on the option chosen.

Five conclusions follow.[9] First, the response rate is nearly 70 percent, which is certainly respectable by social science standards. Thus, there may be too few observations excluded to introduce serious selection bias.

Second, using the full sample of 498, none of the three equations is very successful at explaining nonresponse. All are able to account for 5 percent of the variance. This may result from the omission of important exogenous variables or from near random patterns of nonresponse. If the former, proper corrections may not be feasible. If the latter, the hazard rate to be constructed will have little variance and will be unlikely to have a statistically significant regression coefficient in the substan-

[9] The coding conventions reported at the bottom of Table 1 follow from the derivation of the "hazard rate" for each of the three models. For all three, the goal is to construct a variable that captures the likelihood of exclusion from the sample (i.e., nonresponse). For the linear and logistic, this is accomplished in the way nonresponse is initially coded. As we pointed out earlier, for the probit form this is accomplished later when the new variable is constructed.

Table 1. Selection Equation for Non-Response (Response Rate = 69% of 498 Cases)

Variable	Linear[a]		Logistic[b]		Probit[c]	
	Coeff.	t-Value	Coeff.	t-Value	Coeff.	t-Value
Intercept	0.650	5.48	0.944	1.56	−0.546	1.44
Female Respondent (dummy)	−0.030	−0.72	−0.156	−0.77	0.096	0.78
Age (years)	−0.005	−3.44	−0.025	−3.49	0.015	3.26
Age "missing" (dummy)	0.102	1.38	0.507	1.47	−0.311	−1.54
Respondent Employed (dummy)	0.020	0.44	0.113	0.50	−0.079	−0.59
Respondent Served on Jury (dummy)	−0.107	−1.44	−0.545	−1.46	0.275	1.21
Served × Criminal Trial (dummy)	0.023	0.39	0.116	0.37	−0.036	0.20
Served × Length of Trial (dummy)	0.004	0.54	0.022	0.60	−0.009	−0.43
Served × Defendant Won (dummy)	0.077	1.07	0.371	1.07	−0.21	−1.00
Length of Jury Selection (dummy)	−0.088	−1.25	−0.494	−1.32	0.315	1.35
	R^2 = .05 F = 2.59 P = <.01		D = .05 χ^2 = 23.87 P = <.01		R^2 = .05 F = 2.74 P = <.05	

Descriptive Statistics for Instruments

	N	Mean	Standard Deviation	Minimum	Maximum
Linear	498	−0.70	0.10	−0.99	0.50
Logistic	498	0.30	0.10	0.07	0.53
Probit	498	0.48	0.14	0.12	0.81

[a] 0 = replied, 1 = did not reply.
[b] 0 = replied, 1 = did not reply.
[c] 1 = replied, 0 = did not reply.

tive equation. Near the bottom of the table are shown descriptive statistics for the "hazard rate" variables (a kind of instrumental variable) constructed from the three equations.[10]

Third, keeping in mind the coding conventions listed at the bottom of the table (see footnote 9), the story across the three equations is virtually identical. Perhaps the easiest way to compare across the equations is to examine the three t-values for each parameter estimate. Alternatively, there are approximate transformations between the three sets of coefficients (Amemiya, 1981). For example, if each of the regression coefficients in the probit model is multiplied by .40, approximations of the linear coefficients follow.

Fourth, only one variable has a statistically significant effect on the likelihood of nonresponse at conventional levels. Interpreting the linear coefficient, for each ten years of age the probability of nonresponse decreases by 5 percent. There is also a hint that if the respondent was subjected to a more lengthy jury selection process or was selected to serve as a juror, the likelihood of nonresponse declines. Perhaps greater involvement at the courthouse leads to greater involvement in the questionnaire. Finally, for about 10 percent of the cases age was not available from the official records. For these individuals, the mean was inserted. To control for some distortions that might result, a dummy variable was included, coded "1" for those cases. However, since the official measure of age was routinely obtained from very short questionnaires mailed to all prospective jurors by the Jury Commissioner, we suspected that individuals who did not cooperate fully with the Jury Commissioner would be less cooperative with us. There is a bit of evidence that this is true. Still, the results in Table 1 are not especially instructive.[11]

Fifth, all three "hazard rates" were constructed and correlations were calculated among them. For these data, the lowest correlation is .98. Clearly, it would not matter (and in fact does not matter) which version of the "hazard rate" is used. There is, however, no reason to believe that this is a general result

[10] The "hazard rate" from the linear probability model is equal to the predicted probability of nonresponse minus 1.0. The "hazard rate" from the logit model is simply the predicted probability of nonresponse.

[11] It is possible to find selection effects in one's substantive equation, even if one cannot find systematic selection effects in the selection equation itself (Heckman, 1979:155). However, only the intercept in the substantive equation is altered.

Table 2. Ordinary Least Squares Analysis of Overall Dissatisfaction

"All in all, how would you rate your experience of being called for jury duty?"

Very Satisfied = 3	Somewhat Satisfied = 2	Somewhat Dissatisfied = 1	Very Dissatisfied = 0
39.9% (135)	37.6% (127)	15.7% (53)	6.8% (23)

	Uncorrected		Probit Correction	
Variable	Coeff.	t-Value	Coeff.	t-Value
Intercept	1.47	5.31	2.26	5.91
Hazard Rate	—	—	−1.26	−2.97
Female	0.28	2.91	0.26	2.67
Employed	0.03	0.32	0.15	1.44
White	0.16	1.18	0.12	0.92
Served	0.37	2.07	0.21	1.13
Served × Criminal Trial	0.11	0.73	0.10	0.70
Served × Length of Trial	−0.01	−0.68	−0.00*	0.25
Served × Defendant Won	0.07	0.40	0.21	1.20
Length of Jury Selection	−0.17	−1.07	−0.33	−1.98
1st Time Called	0.16	1.57	0.20	1.99
# Days Notice	0.11	1.85	0.12	1.98
Does Not Drive	0.05	0.31	0.07	0.48
	R^2 = .10		R^2 = .13	
	F = 3.41		F = 3.94	
	P = < .001		P = < .001	

* Negative, but smaller than 0.00.

and may be a consequence of the small amount of variance explained in each of the three selection equations; all three constructed "hazard rates" may be insufficiently variable to reveal properly their different forms.

Table 2 shows the results for one of the questionnaire items. Among those who returned the questionnaire, nearly 80 percent were at least somewhat satisfied with the experience. For most of the other questions, similar sentiment was expressed (Berk and Shih, 1982).

Turning to the multivariate equations, the left-hand side shows the usual least squares coefficients. The results on the right-hand side have been corrected through the addition of the hazard rate instrument. Perhaps the most important message is that the uncorrected and corrected results differ substantially. With the addition of the hazard rate, 3 percent more variance is explained, and the regression coefficient for the hazard rate is statistically significant at well beyond conventional levels (t=−2.97). The sign of the regression coefficient indicates that individuals who are less likely to return the questionnaire are more critical of the jury experience; complainers are less inclined to respond.

More important, the uncorrected equations include *false positives and false negatives*. Statistically significant coefficients would have been overlooked for the length of the jury selection, whether the respondent had previously been called for jury duty, and the number of days' notice given (assuming a two-tailed test). After corrections are made,[12] respondents are more positive if they are first timers, if they are given more notice, and the time taken for jury selection is shorter. All three effects have important policy implications (Berk and Shih, 1982) that would have been lost had corrections for sample selection bias not been undertaken. Note also that the relative importance of the jury selection variable has been substantially altered.

In the uncorrected equation, there is one false positive; individuals who served on a jury are incorrectly deemed more positive. In other words, one would have falsely concluded that serving on a jury by itself led to more favorable assessments.

Finally, only one causal effect holds in both the corrected and uncorrected equations. Female respondents are in both instances more complimentary. This was a general result over a wide variety of items.

There is, however, at least one important ambiguity. With the correction, there is a substantial change in the intercept, perhaps implying an increase in the mean of the endogenous variable. That is, the change in the intercept suggests that the original regression

[12] The corrected results are nothing more than ordinary least squares with the hazard rate included. Technically, generalized least squares is superior, but by using procedures outlined by Heckman (1976:483) little of interest changes. Still better would have been maximum likelihood procedures, but no software was available.

Ray, Subhash C., Richard A. Berk and William T. Bielby
1980 "Correcting for sample selection bias for a bivariate logistic distribution of disturbances." Paper presented at the 1980 meetings of the American Statistical Association.

Rossi, Peter H., Richard A. Berk and Kenneth J. Lenihan
1980 Money, Work, and Crime: Some Experimental Results. New York: Academic Press.

Rossi, Peter H., James D. Wright and Eleanor Weber-Burdin
1982 Natural Hazards and Public Choice. New York: Academic Press.

Sickles, Robin C. and Peter Schmidt
1978 "Simultaneous equation models with truncated dependent variables: a simultaneous Tobit model." Journal of Economics and Business 33:11–21.

Tobin, James
1958 "Estimation of relationships for limited dependent variables." Econometrica 26:24–36.

Tuma, Nancy B.
1982 "Nonparametric and partially parametric approaches to event-history analysis." Pp. 1–60 in Samuel Leinhardt (ed.), Sociological Methodology, 1982. San Francisco: Jossey-Bass.

Tuma, Nancy B., Michael T. Hannan and Lyle P. Groenveld
1979 "Dynamic analysis of event histories." American Journal of Sociology 84:820–54.

28

A Scheme for Assessing Measurement Sensitivity in Program Evaluation and Other Applied Research

Mark W. Lipsey

In program evaluation and other areas of applied social research it is essential that the measures be sensitive to the treatment effects of interest if the researcher is to minimize the risk of concluding that a treatment or program is ineffective when in fact it is the research that has failed. Though pertinent to other areas of research, this issue becomes most important under field conditions in which the researcher may be required to use unproven measures and often loses control of such factors as measurement protocols, sample composition, treatment implementation, and other sources of extraneous variance that affect the ability of the measures to detect changes or differences of interest. In such circumstances it is quite likely that irrelevancies will creep into the measures and degrade their responsiveness, making treatment effects difficult or impossible to detect even under otherwise favorable conditions (Boruch & Gomez, 1979).

The issue with which measurement sensitivity is concerned is the responsiveness of the measured value of a variable to a change or difference on the underlying construct of interest (Aiken, 1977). Formulated in terms of classical measurement theory, $X_i = T_{o_i} + T + E_i$, where X_i is the measured value for an individual, T_{o_i} is the portion of the true score that represents the individual's baseline value on the particular measure, T is the portion of the true score that represents change or difference (treatment effect), and E_i is error, that is, everything else that contributes to the individual's score. A sensitive measure is one in which a change or difference on the underlying construct produces a proportionately large T and, hence, a readily detectable increment in the measured value, X_i.

Measurement sensitivity can thus be represented as a signal to noise ratio in which the signal is the change or difference on the construct that is to be detected and noise is everything else that contributes to the measured value. A variety of factors can be seen to influence measurement sensitivity in this formulation. In the first place, the component of the measure representing the signal (T) may not adequately reflect the actual change on the underlying construct. For example, the measure may show ceiling or floor effects so that a change on the construct is not proportionately represented in the measure. Or, the scale units of the measure may be too coarse to reflect the change on the construct, just as a balance scale with one pound weights is insensitive to differences of a few ounces.

In addition, even if the T component is fully responsive to the underlying change, the noise reflected in the measure may obscure it. The component of the measure reflecting

This work was supported in part by Grant 80-IJ-CX-0036 from the National Institute of Justice, Office of Research and Evaluation Methods.

From Mark W. Lipsey, "A Scheme for Assessing Measurement Sensitivity in Program Evaluation and Other Applied Research," Psychological Bulletin, 1983, 94, 152-165.

the true baseline value (T_{o_i}) may itself be quite large and overshadow the change component; that is, the measure may primarily emphasize "traits" instead of "states." In psychology this is most apt to occur with measures designed to index stable individual differences rather than changes in individual performances (Carver, 1974). Natural subject heterogeneity then contributes a substantial "noise" to a measure being used to detect changes.[1] This source of measurement noise is typically handled at the level of experimental design; for example, with blocking, covariates, or repeated measures.

Finally, the error term itself (E_i), of course, contributes noise to the measure. Error can be divided into two categories, measurement error and experimental error. Measurement error refers to the intrinsic unreliability and invalidity of the measurement instrument. Every measurement operation varies somewhat from application to application even under "exact" replication and, additionally, responds to certain stable but extraneous influences. Experimental error, on the other hand, is introduced through procedural variation in administering measures, uncontrolled measurement conditions, and careless experimental technique.

Measurement sensitivity thus depends on the properties of the measure, the conditions of measurement, and the research design. Though the issues of measurement sensitivity are intimately related to the issues of experimental design and statistical power (cf. Cleary, Linn, & Walster, 1970; Levin & Subkoviak, 1977; Sutcliffe, 1980), the measurement issues are more general. Factors from the experimental design and implementation represent some, but not all, of the factors that contribute to the variance in the measured values.

In applied research, good texts (e.g., Posavac & Carey, 1980; Riecken & Boruch, 1974) acknowledge the importance of measurement sensitivity and advise the researcher to give careful consideration and, if necessary, separate study, to the proposed measures before they are used in an evaluation study. Little practical guidance for designing a measurement study is offered, however. It is the purpose of this article to show how an analysis of the variance components of a measure, modeled after Cronbach's generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), can provide a practical basis for an assessment of the sensitivity of proposed measures, a diagnosis of their deficiencies, and an exploration of possible modifications in measurement operations and research design that might compensate for those deficiencies. This framework also allows the researcher to investigate the relationships among measurement sensitivity, reliability (or generalizability), and statistical power.

The Measurement Assessment Study

Following Cronbach et al. (1972), we suggest that the first step in a measurement assessment study is careful consideration of the measurement application that is planned in the subsequent substantive study. The candidate measurement instruments or procedures for use in that study must be selected and the conditions of measurement determined. Each potential source of insensitivity of the measures to the treatment effects of interest must then be identified and considered for examination in the measurement study. That is, any aspect of the measurement procedure that may vary in the application study, such as setting, occasion of measurement, or individual differences among subjects, should be defined as a separate measurement facet in the measurement study.

The basic framework for a measurement assessment study is an analysis of variance design in which the important measurement facets are systematically varied so that their effects on the candidate measures can be examined. Inclusion of two or more occasions of measurement (repeated measures), for instance, allows an assessment of conventional reliability. It might also be desirable to include different observers or raters, different settings, and so forth, as factors in the analysis of variance (ANOVA) design. This part of the measurement study follows Cronbach et al. (1972) very closely.

---

[1] Note that one observer's "noise" is another observer's "signal." If attention centers on individual differences, T_{o_i} is the signal of interest, everything else is error. When T is of interest, T_{o_i} is extraneous.

Direct examination of the overarching issue of measurement sensitivity, however, requires that the measures be assessed in terms of their ability to detect a change or difference under the circumstances of interest. Such an effect can be provided in surrogate form through the use of criterion groups, samples of subjects chosen a priori to differ on the construct of interest to the study. The use of criterion groups to provide a "known" difference against which to assess the characteristics of measures is, of course, a well established psychometric technique (Cronbach & Meehl, 1955).

Criterion Groups

In the physical sciences it is customary to check the accuracy and sensitivity of test instruments by applying them to samples of known composition to confirm that the measurement results are appropriate. The measurement assessment strategy proposed here takes an analogous approach by using criterion groups to provide a standard for judging measurement adequacy. Criterion groups are chosen to represent a difference of at least the same order of magnitude as the minimal treatment effect deemed worthwhile to detect in the application study. Appropriate criterion groups should also be representative of the relevant population in terms of heterogeneity and similarity of pertinent characteristics, and they should be measured under circumstances comparable to those anticipated for the application study. In all cases, of course, the determination of criterion groups must be independent from the measures to be tested on those groups.

If achievement in an educational program is of interest, for example, average progress from one grade year to the next under ordinary school conditions might be chosen as a reasonable standard of a practical amount of change. Researchers and program staff might agree that if a treatment program produced an effect equivalent to the normal year to year developmental and educational change, the treatment effect was of practical benefit. In this case, the specific criterion groups might be a third-grade class and a fourth-grade class, with their difference in achievement level representing an effect size to which the proposed measures must be sensitive if they are to be useful for treatment evaluation. By beginning with a contrast judged to have practical significance, one can put measured effects in better perspective. Carver (1975) observed, for example, that the effect of one grade year of development and learning accounted for only 2%–3% of the variance on some standardized reading tests. A special reading program that produced that much effect on one of those tests thus might be judged of considerable practical significance even though the effect size would be small by the usual statistical standards.

Other examples of possible criterion groups can be readily produced though, in general, they will be specific to the context of application. For example, in a mental health setting, criterion groups might be defined on the basis of therapists' nominations of the greatest successes and failures in their caseloads. Alternatively, categorical client groups might be used, for example, short term outpatients versus chronic patients in a day treatment program. Income maintenance and other welfare programs might form criterion groups from applicants needy enough to meet eligibility requirements and those whose resources clearly exceed requirements. A health care program might use its initial diagnostic screening to identify more and less severely impaired patients.

Inevitably, the choice of criterion groups involves considerable judgment and even a degree of arbitrariness on the part of the researcher. The guiding principle, however, is that the contrast between groups should represent a difference that has clear practical significance when judged in the context of the subject population of interest. One check on the appropriateness of the criterion groups might be whether expert practitioners can agree that the groups represent a "better" and a "worse" status with regard to the condition to be treated. Additionally, the criterion group contrast can be compared with typical effect sizes for the treatment at issue if sufficient information from other studies is available (cf. Sechrest & Yeaton, 1981).

Analyzing Variance Components in a Measurement Study

We are interested in examining the relations among the measurement terms defined by the expression $X_t = T_{o_t} + T + E_t$, which

we must now understand to represent a set of potential measurement models, with E_i representing not only residual random error but whatever measurement facets are assumed to contribute systematic error plus interactions among the various measurement components. In order to accomplish this, an analysis of variance study is designed in which the criterion group contrast (as a surrogate treatment effect) is one factor and the measurement facets thought to pose threats to measurement sensitivity are varied to constitute additional factors. A great variety of such designs is possible (cf. Brennan, 1980; Cronbach et al., 1972), although, as will be shown shortly, fully crossed split-plot designs provide a general and informative framework for measurement assessment studies.

For a population of measured values, the measurement terms of interest are most conveniently expressed as components of the total measurement variance. Components of variance can be estimated for data in an ANOVA format by using an established technique often referred to as *components of variance analysis* (Cronbach et al., 1972; Dwyer, 1974; Gaebelein, Soderquist, & Powers, 1976; Halderson & Glasnapp, 1972; Hays, 1973; Searle, 1971; Vaughan & Corballis, 1969). Although an example will be worked out here, the reader should consult the references above for more complete details on the procedure.

Suppose that only one measurement facet is of interest—a likely choice is differences due to occasion of measurement since that comparison reflects classically defined reliability. The researcher designs the measurement assessment study as a split-plot ANOVA with two criterion groups of subjects, each of which is measured on two occasions. In this case, the measurement components defined by the expression $X_i = T_{o_i} + T + E_i$ can be put in the form of an ANOVA model linking the observed scores with the appropriate population parameters as follows:

$$X_{ijk} = \mu + \pi_{i(j)} + \gamma_j + o_k + o\gamma_{jk} + o\pi_{ik(j)} + \epsilon_{ijk},$$

where

X_{ijk} = the score of the ith individual in the jth level of γ and the kth level of o;

μ = the grand mean of the population, which, when added to the subject constant, $\pi_{i(j)}$, equals T_{o_i}, the individual's true baseline value;

$\pi_{i(j)}$ = the subject constant associated with each individual, i, nested under level γ_j;

γ_j = the effect of the jth level of the criterion contrast (a surrogate for the treatment of interest);

o_k = the effect of the kth level of occasion of measurement, part of the total measurement error;

$o\gamma_{jk}$ = the effect of the interaction between the occasion of measurement and the criterion contrast, part of the total measurement error;

$o\pi_{ki(j)}$ = the effect of the interaction between occasion of measurement and the subject constant, part of the total measurement error;

ϵ_{ijk} = the random error component specific to this individual score and assumed independent of all other components; in this design, ϵ_{ijk} cannot be estimated separately from $o\pi_{ki(j)}$.

Table 1 summarizes the design with C_j representing two criterion groups and O_k representing two different occasions of measurement (test–retest). The researcher will want to generalize to other similar occasions and subjects, so these are analyzed as random effects, whereas the criterion contrast, which was deliberately chosen, is analyzed as a fixed effect. In general, the criterion contrast will always be a fixed effect, whereas levels of the measurement facets should be sampled and treated as random effects, a mixed-model ANOVA.[2]

Working from the mean square values calculated from sample data and the formulas for the expected mean squares, $E(MS)$, as

[2] In practice, of course, it may prove difficult to sample randomly from a universe of measurement facets; indeed, it may even prove difficult to adequately define the universe to be sampled (cf. Cronbach et al., 1972, pp. 376–380). Where such sampling does not actually take place, the researcher must decide whether it is tenable to assume that the variability across the selected levels of each measurement facet can be generalized beyond those specific levels. If not, the facet must be treated as a fixed effect variable and appropriate adjustments must be made when calculating variance components.

Table 1
Analysis Summary of the Illustrative Measurement Assessment Design

Source	Associated variance component estimate	*df*	*MS*	*E*(*MS*)[a]	*E*(*MS*) formulations as simultaneous equations
Between subjects					
Criterion contrast (C)	$\hat{\sigma}^2_\gamma$	$(J-1)$	MS_c	$\sigma^2_\varepsilon + \sigma^2_{o\pi} + I\theta^2_{o\gamma} + K\sigma^2_\pi + IK\theta^2_\gamma$	$MS_c = \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_{o\pi} + I\hat{\theta}^2_{o\gamma} + K\hat{\sigma}^2_\pi + IK\hat{\theta}^2_\gamma$
Subjects within groups (S)	$\hat{\sigma}^2_\pi$	$J(I-1)$	MS_s	$\sigma^2_\varepsilon + \sigma^2_{o\pi} + K\sigma^2_\pi$	$MS_s = \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_{o\pi} + K\hat{\sigma}^2_\pi$
Within subjects					
Occasions (O)	$\hat{\sigma}^2_o$	$(K-1)$	MS_o	$\sigma^2_\varepsilon + \sigma^2_{o\pi} + IJ\sigma^2_o$	$MS_o = \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_{o\pi} + IJ\hat{\sigma}^2_o$
O × C	$\hat{\sigma}^2_{o\gamma}$	$(K-1)(J-1)$	$MS_{o\gamma}$	$\sigma^2_\varepsilon + \sigma^2_{o\pi} + I\theta^2_{o\gamma}$	$MS_{o\gamma} = \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_{o\pi} + I\hat{\theta}^2_{o\gamma}$
O × S; residual	$\hat{\sigma}^2_E$	$J(I-1)(K-1)$	MS_E	$\sigma^2_\varepsilon + \sigma^2_{o\pi}$	$MS_E = \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_{o\pi}$
Total	$\hat{\sigma}^2_T$	$IJK-1$			

Note. This example assumes a nonadditive model, thus $\hat{\sigma}^2_\varepsilon$ and $\hat{\sigma}^2_{o\pi}$ are confounded and must be solved as a single composite term, here designated as $\hat{\sigma}^2_E$.
[a] C represents a fixed effect; O and S are random effects.

given in standard texts (e.g., Kirk, 1968, who also gives the algorithm for more complex designs), each variance component can be separately estimated by redesignating the *E*(*MS*) formulations as sample estimates and solving them as simultaneous equations, shown on the right in Table 1. The resulting estimates of each parameter of the *E*(*MS*) formulations represent the sample variances of interest directly for random effect variables and, for fixed effect variables, produce the appropriate variances when each is weighted by $(K-1)/K$, where K equals the number of levels of the variable at issue[3] (see Searle, 1971; Vaughan & Corballis, 1969).

With the variance components defined, the total measurement variance can be expressed as the sum of the separate variance components. For the example of Table 1:

$$\sigma^2_x = \sigma^2_\gamma + \sigma^2_o + \sigma^2_\pi + \sigma^2_{o\gamma} + \sigma^2_{o\pi} + \sigma^2_\varepsilon.$$

The signal/noise ratio defined here as measurement sensitivity is thus[4]

$$ES_m = \frac{\sigma^2_\gamma}{\sigma^2_o + \sigma^2_\pi + \sigma^2_{o\gamma} + \sigma^2_{o\pi} + \sigma^2_\varepsilon}. \quad (1)$$

[3] The parameters of the *E*(*MS*)s have the form $\theta^2_\alpha = \Sigma\alpha^2/(K-1)$, where α is the parameter for the treatment effect and K is the number of treatment levels compared. For random effects where the K levels in the design are a sample from a population, θ^2_α represents the sample variance directly. For fixed effects, however, the K treatment levels represent the entire population and a population variance of the form $\Sigma\alpha^2/K$ is needed. In the latter case, multiplying θ^2_α by $(K-1)/K$ produces the desired form and this adjustment must be made to each estimate of fixed effect parameters. Terms representing the interaction of fixed effect variables with random effect variables are also adjusted in the present article, although there is some dispute regarding that practice (e.g., Dwyer, 1974). Such adjustment is appropriate under the customary ANOVA assumption that $\sum_{j=1}^{J} \alpha\beta_{jk} = 0$ for all k, where J is the number of treatment levels for α and K is the number of treatment levels for β (Gaebelein et al., 1976; Searle, 1971, chap. 9). The parameters involving fixed effects are designated with θs in Table 1; those corresponding to random effects are presented in σ^2 form.

[4] Although estimates of the ratio itself are likely to be biased as population estimates in ways not easy to correct (Glass & Hakstian, 1969; Olkin & Pratt, 1958), we echo the sentiments of Vaughan and Corballis (1969): "The risk of bias may not be a serious objection . . . particularly if large samples are used, and if one adopts the view that even a biased estimate is better than none" (p. 205).

Equation 1 essentially defines an effect size ratio for the magnitude of the criterion contrast under the particular measurement circumstances of interest. It is not dissimilar from the most common effect size index, proportion of total variance accounted for, lacking only σ^2_γ in the denominator to make it total variance instead of irrelevant variance. Using the procedures above, an estimate from sample values can be made for each of the variance parameters in Equation 1. By carefully examining and manipulating those variance components, it is possible to gain considerable insight into the strengths and weaknesses of the candidate measures.

The ES_m Numerator

The numerator of the ES_m ratio represents a direct estimate of the response of the candidate measure to the surrogate treatment effect built into the design with the criterion groups. With various measures taken on the same samples under the same conditions, ES_m can be compared for different measures and combinations of measures to determine which are more responsive to the criterion contrast. The factors that restrict the criterion contrast variance in ES_m are very fundamental aspects of measurement. For instance, if the measures are not valid for the treatment effect of interest they will be unresponsive to the criterion group difference. Or, the measure's response to the criterion group difference may be limited if the scale units of the measure are too coarse for the magnitude of the difference represented by the criterion contrast or if the measure has a ceiling or floor to its response that is reached by one of the criterion groups.

The ES_m Denominator

The ES_m denominator in a measurement study provides information about the background noise that may obscure the treatment effect signal that the measures are intended to detect. The variance components of the denominator can be divided into three overlapping groups—those relating to measurement facets (occasions in our example), those relating to persons, and the residual error. Inspection of the terms in each of these groups provides some general diagnosis of any unusual measurement features.

Large interaction variances involving measurement facets alert the researcher to capricious characteristics of the measures, that is, results that differ considerably when such measurement circumstances as occasion, setting, rater, and so on, differ. The variance components involving between-persons differences and the residual error term are especially important because these relate directly to subject heterogeneity, reliability and generalizability, and statistical power. Each of these topics requires a more thorough discussion.

Subject Heterogeneity

The $\hat{\sigma}^2_\pi$ variance component is a particularly important one for purposes of measurement assessment. It represents the individual differences variability among persons on the measures of interest. It is only when the measurement study includes repeated measures that this variance component can be separated from the residual error term. In study designs that make comparisons between groups, for example, treatment versus control groups, $\hat{\sigma}^2_\pi$ will be part of the error term against which treatment effects are assessed.

Carver (1974, 1975) has argued that many standardized psychometric tests are developed in ways that maximize their response to stable individual differences and minimize their response to externally induced changes, thus decreasing their sensitivity to treatment effects. Measures that have large between-persons variance relative to the criterion contrast may be poor choices for investigation of treatment effects because the individual differences variance may obscure legitimate effects.

If such measures must be used, the researcher who performs an appropriate measurement study can be alerted to the problem and take corrective action. The subsequent application study itself might be made a repeated-measures design, for instance, so that the between-persons variance can be removed from the error term used for testing the treatment effect. Alternatively, various blocking variables or covariates could be tried in an effort to control subject heterogeneity statistically (Myers, 1979, chap. 6). Indeed, promising blocking factors can be incorpo-

rated into the measurement study so that a direct examination can be made of the extent to which they alleviate the problem of subject heterogeneity. The Blocks × Treatment interaction may also be examined for its utility in reducing the residual error.

Reliability and Generalizability

If the same measure is applied to the same persons on two occasions and the results are analyzed in terms of the variance components from an ANOVA, the reliability of the measure is given by the following intraclass correlation (Winer, 1971, p. 283):

$$\rho = \frac{\sigma^2_\pi}{\sigma^2_\pi + \sigma^2_E/k_o}, \qquad (2)$$

where $\sigma^2_E = \sigma^2_\epsilon + \sigma^2_{o\pi}$ and k_o = number of occasions aggregated. Cronbach's generalizability theory has extended this formulation to encompass generalizability of the results of measurement across any measurement facet or set of facets, with conventional reliability representing only the case of generalizability across temporal occasion (Brennan, 1980; Cronbach et al., 1972). For example, in a design that varied both raters (R) and occasions (O), the coefficient for generalizability across both these facets would have the following form:

$$G = \frac{\sigma^2_\pi}{\sigma^2_\pi + \sigma^2_{r\pi}/k_r + \sigma^2_{o\pi}/k_o + \sigma^2_E/k_rk_o}$$

More generally, the generalizability coefficient is determined by the ratio of the between-subjects variance to the weighted sum of the variance components representing interactions between subjects and the measurement facets (including the residual error term) over which generalizability is being examined. A useful feature of the generalizability coefficient for present purposes is that the variance components can be adjusted to represent different measurement circumstances, and the generalizability re-estimated. In particular, when the ks in the formulas given above are set equal to one, the coefficient estimates the generalizability under conditions of one rater, occasion, and so on. Setting the appropriate k to another value estimates generalizability for a measure aggregated over k raters, occasions, and so on.

Since the measurement study defined here uses a generalizability study format, adding only a criterion contrast, the resulting variance components provide appropriate estimates of the terms in the generalizability coefficient. Comparing the denominator of the ES_m in Equation 1 with Equation 2, it can be seen that the variance component representing residual error and the confounded interaction of occasions and subjects carries information about the reliability of the measure. More generally, the relevant components are those representing interactions of subjects with measurement facets, including the residual error term. If those components are large relative to the between-subjects variance, it indicates a problem of poor reliability with consequent degradation of the sensitivity of the measure to the criterion contrast. By dividing the appropriate variance components by some $k > 1$, the researcher can estimate the effect of aggregating raters, occasions, and so on upon the total ES_m. That is, the researcher can estimate the decrease in noise and, hence, the enhancement of measurement sensitivity that will result from combining measures into composites (Cascio, Valenzi, & Silbey, 1980) or averaging over multiple occasions or multiple raters (Epstein, 1980; Green, Nguyen, & Attkisson, 1979).

Statistical Power

The ability of a research study to detect a treatment effect depends on the statistical power of the design. Statistical power is a function of the significance criterion (alpha level), sample size, and effect size, that is, the population magnitude of the effect to be detected on the chosen measure. A convenient way to examine statistical power within the ANOVA framework defined here is to use Cohen's (1977) tables, which are built around an effect size index, f, defined as follows:[5]

$$f = \sqrt{\frac{\sigma^2_\tau}{\sigma^2_\xi}},$$

[5] Cohen's effect size index f is related to the frequently used phi parameter as $f = \phi/n$ and to the noncentrality

where σ^2_τ = population variance for treatment effect and σ^2_ξ = common variance of populations involved in the statistical comparison.

These population parameters are estimated from sample values in order to determine the statistical power of the comparisons of interest. The variance components estimated from the measurement study ANOVA provide appropriate values. The criterion contrast, σ^2_γ (estimated by $\hat{\sigma}^2_\gamma$), corresponds in a straightforward way to σ^2_τ. The σ^2_ξ term is estimated by the error term against which the treatment is appropriately tested in the design of interest. If the measurement study has adequately represented the important measurement and design facets, σ^2_ξ for the planned application study can be approximated by combining the appropriate variance components from the denominator of the ES_m ratio. For example, if the planned study will be a group comparison rather than a repeated measures design, $\hat{\sigma}^2_\pi$ must be added to $\hat{\sigma}^2_E$ to estimate σ^2_ξ. Furthermore, variation associated with each of the measurement facets that will not be controlled or statistically removed in the planned study must also be included in the estimate of σ^2_ξ. Where judges, occasions, and so forth are to be aggregated in the planned study, the associated variance is divided by the appropriate k to determine its contribution to the estimate of σ^2_ξ.

To take a somewhat extreme example, suppose the application study was planned as a simple comparison between treatment and control groups. Suppose further that the study conditions were such that the outcome measures would be collected on randomly varying occasions of the sort represented by O_k in Table 1. Under these circumstances, Cohen's effect size index, f, would be estimated by

$$f = \sqrt{\frac{\hat{\sigma}^2_\gamma}{\hat{\sigma}^2_\pi + \hat{\sigma}^2_{o\gamma} + \hat{\sigma}^2_{o\pi} + \hat{\sigma}^2_\epsilon}}$$

Examining these variance components helps the researcher determine what modifications in the design and measurement procedures might yield the greatest increases in statistical power. A large $\hat{\sigma}^2_\pi$ term indicates that the design might profit from some statistical or design control of subject heterogeneity. Large interaction variances of measurement facets with subjects and/or large residual error indicates that improvements in reliability would be beneficial, for example, by aggregating measures or occasions. The effects of these various modifications on power can be assessed by removing or adjusting (e.g., by $1/k$) the appropriate variance components in the denominator of the f index. Of course, attention should also be given to sample size and the level at which alpha is set, since these are the other major determinants of statistical power.

Adjusting the Criterion Group Contrast

The criterion contrast that is designed into a measurement study represents only a single (surrogate) treatment effect. Often it will be desirable to examine the implications of a treatment effect that is smaller or larger than the one available for the measurement study. For example, it may be of interest to consider the effects of incomplete implementation of the treatment, that is, not all experimental subjects receiving the full treatment as intended and, perhaps, some controls receiving the treatment when it was not intended (Sechrest & Redner, 1979). Such treatment degradation is not uncommon under field conditions and has been shown to sharply decrease statistical power (Boruch & Gomez, 1979).

One approach to adjusting the criterion contrast variance, $\hat{\sigma}^2_\gamma$, is to look at the mean difference between the two criterion groups, call it $\bar{c}$. Then $\bar{c}$ estimates the average increment that the treatment is expected to add

parameter, λ, as $f = \lambda/nk$. Note that Cohen's f is a signal/noise ratio very similar to the ES_m defined here for measurement sensitivity. Indeed, under some circumstances, they will both be estimated, with identical variance components. The ES_m ratio, however, includes all variance components irrelevant to the treatment effect whereas Cohen's f includes only those that represent variance within the treatment populations defined by the design. Some of the irrelevant components that degrade overall measurement sensitivity may be removed from the variance appropriate to statistical power through judicious selection of design and measurement features. Measurement sensitivity and the effect size issue in statistical power are thus intimately related but are not identical concepts (Boruch & Gomez, 1979; Sutcliffe, 1980).

to the treated group, as assumed when the criterion groups were chosen. We can then suppose that the effect of treatment degradation is to decrease the $\bar{c}$ by a certain average proportion, p. Thus the degraded treatment effect is $p\bar{c}$. This is a relatively easy alteration to think about. For example, if every treated person changed only 80% as much as would be expected, p equals .8; if the controls changed 20% as much as expected for experimentals because they got unplanned treatment, p also equals .8; if 80% of the experimentals got the full treatment but 20% received none, p also equals .8; and so forth.

If the treatment effect is reduced to $p\bar{c}$, the new criterion contrast becomes $p^2\hat{\sigma}^2_\gamma$. Thus the $\hat{\sigma}^2_\gamma$ variance component is adjusted by p^2 as the difference between treatment group means changes by proportion p.[6] The researcher can therefore explore the effects of treatment degradation on the various measurement sensitivity issues by choosing meaningful values of p and making appropriate modifications in $\hat{\sigma}^2_\gamma$. This procedure can also be used to adjust the criterion group contrast in circumstances where the researcher is unable to obtain criterion groups that differ to the degree desired to represent the minimal treatment effect expected in the planned application study. While more representative, the adjusted criterion contrast in such cases still has a readily understandable relationship to the mean difference between the original groups.

Some Examples

Two examples from program evaluation data are presented here. They illustrate the application of the measurement assessment procedures described in this paper and the nature of the results that can be expected.

Juvenile Delinquency

One illustrative data set comes from the author's evaluation studies of juvenile delinquency prevention programs (e.g., Lipsey, Cordray, & Berger, 1981). These data comprise recidivism measures on 1,069 juveniles arrested at four different police stations in Los Angeles County. The primary measure is the number of offenses recorded in police records during a 6-month period subsequent to or prior to a marker offense drawn from police logs during a specified interval. The number of recidivistic offenses during a fixed time period is a common outcome measure for juvenile delinquency programs (cf. Wright & Dixon, 1977).

The cases in these data can be divided into two very reasonable criterion groups. One group of juveniles was 'reprimanded and released' by the police because their offenses and past records were judged minor. The other group was judged by the police to be more serious offenders and was referred to the probation department for formal action. The difference between these two groups is substantial and meaningful to law enforcement and justice system personnel. Indeed, a delinquency prevention program that had an effect on the behavior of the delinquents that was as much as half the difference between these criterion groups could justifiably claim that it was achieving a result of practical significance.

Police station of arrest was included as a blocking factor in the measurement study design to explore the possibility that it might reduce the variability among subjects with regard to their arrest records. To provide a repeated measures comparison, one variable counted offenses during the 6-month 'prior' period and another counted offenses during the 6-month 'recidivism' period. The distributions were normalized by transforming the values as $\sqrt{x} + \sqrt{x+1}$ (Freeman & Tukey, 1950) and scores for each observation period were separately standardized to force the test–retest means to be equal (the mean difference between O_1 and O_2 being irrelevant to the measurement issues). The variance components for the various factors were determined using the appropriate $E(MS)$ formulations for a mixed model split-plot design with one blocking factor and the mean squares calculated from sample data. The

---

[6] As a sample variance, $\hat{\sigma}^2_\gamma$ has the form $\Sigma(\bar{x}_j - \bar{\bar{x}})^2/(J-1)$, where $\bar{\bar{x}}$ equals the grand mean and $\bar{x}_j$ equals the mean for each treatment. Since there are only two groups in our criterion comparison, $\hat{\sigma}^2_\gamma$ where we expect a group difference of $\bar{c}$ is simply $\bar{c}^2/2$, assuming equal sized groups (i.e., $(\bar{c}/2)^2 + (\bar{c}/2)^2$). When the treatment effect is reduced to $p\bar{c}$, the new criterion contrast becomes $p^2\bar{c}^2/2$, that is, $p^2\hat{\sigma}^2_\gamma$.

estimates of the variance components that resulted are shown in Table 2.

The criterion contrast, though large and meaningful in practical terms, contributed less than 4% of the total variance on the delinquency measure used. The corresponding signal/noise ratio (ES_m) was .04. This delinquency measure was clearly very insensitive to group differences of the sort that might arise in program evaluation studies. Such insensitivity has particularly serious consequences for statistical power. Suppose a treatment effect roughly one-half the size of the criterion contrast of Table 2 ($\hat{\sigma}^2_\gamma \times .5^2$) is expected in a simple treatment versus control group comparison. Under these circumstances, Cohen's (1977) effect size index, f, is .10, that is, the square root of $(.25\hat{\sigma}^2_\gamma/(\hat{\sigma}^2_T - \hat{\sigma}^2_\gamma))$. The n required to detect that criterion contrast 90% of the time at the $\alpha = .05$ level is over 500 in each group. The evaluation studies of delinquency prevention programs reviewed by Wright and Dixon (1977) typically used a two-group design and had a median group n of 80. Most of those studies, therefore, had little chance of detecting a treatment effect of worthwhile magnitude even if one were present.

Under such circumstances, the researcher may wish to explore design modifications that might enhance statistical power. For example, removing $\hat{\sigma}^2_\pi$ and $\hat{\sigma}^2_\rho$ from the denominator of Cohen's f will determine the power with the between subjects variance removed statistically, as with a repeated measures design. Similarly, the contribution of the blocking factor, police stations, can be assessed by removing its variance component and interaction terms involving it from what would otherwise be the total error variance used in the analysis.

To more clearly diagnose the deficiencies of the delinquency measure, the relative magnitude of the variance components $\hat{\sigma}^2_\pi$ and the residual error, $\hat{\sigma}^2_{o\pi} + \hat{\sigma}^2_\epsilon$ must be examined. The between-subjects within-groups component, $\hat{\sigma}^2_\pi$, indicates the relative amount of uncontrolled subject heterogeneity which, if large, could overshadow the treatment effect. The blocking factor, police stations, reduced the subject heterogeneity an appreciable amount but, even so, almost 20% of the total variance remained in that category. The problem presented by subject heterogeneity, however, was comparatively minor when compared to the unreliability of the delinquency measure, as indicated by the residual error term, $\hat{\sigma}^2_{o\pi} + \hat{\sigma}^2_\epsilon$, which constituted almost 70% of the total measurement variance. Using Equation 2 to calculate the reliability for this measure (with $\hat{\sigma}^2_\rho + \hat{\sigma}^2_{\gamma\rho} + \hat{\sigma}^2_\pi$ as the between-persons variance) gave a coefficient of .26, well below conventional research standards. Thus the insensitivity of the delinquency measure stemmed primarily from its extremely low reliability. Rather than pursuing design enhancements, the researcher would be better served by investigating alternate measures, aggregation of multiple measures, and other such approaches to improving the measurement system itself. Some simple algebraic manipulations based on Equations 2 and 3 would permit a calculation of the minimal reliability the measure must have in order to support a given level of statistical power in any of various research designs.

Table 2
Variance Components for Recidivism Measure

Source	Associated variance components estimate	Raw value	Proportion of total
Between subjects			
Criterion contrast (C)	$\hat{\sigma}^2_\gamma$	3.65	.037
Police station (P)	$\hat{\sigma}^2_\rho$	5.60	.056
C × P	$\hat{\sigma}^2_{\gamma\rho}$	0.00	.000
Ss w/in groups (S)	$\hat{\sigma}^2_\pi$	19.47	.195
Within subjects			
Occasions (O)	$\hat{\sigma}^2_o$	.00	.000
O × C	$\hat{\sigma}^2_{o\gamma}$	.85	.009
O × P	$\hat{\sigma}^2_{o\rho}$	.28	.003
O × C × P	$\hat{\sigma}^2_{o\gamma\rho}$	.20	.002
O × S; residual	$\hat{\sigma}^2_{o\pi} + \hat{\sigma}^2_\epsilon$	69.89	.699
Total	$\hat{\sigma}^2_T$	99.94	1.000

Note. C and P represent fixed effects; O and S are random effects.

Language Development

The second example illustrates the generality of the measurement assessment frame-

Table 3
Variance Components Expressed as a Proportion of Total Variance for the Language Measures

Measure	Blocks (B) $\hat{\sigma}^2_\beta$	Subject constant[a] (S) $\hat{\sigma}^2_\pi$	Criterion contrast (C) $\hat{\sigma}^2_\gamma$	B × C $\hat{\sigma}^2_{\beta\gamma}$	S × C[a] $\hat{\sigma}^2_{\gamma\pi}$	Residual error[a] $\hat{\sigma}^2_\varepsilon$	ES_m
Peabody (Raw Scores)	.01	.68	.14	.00	.06	.12	.16
Peabody (Scaled Scores)	.00	.45	.01	.00	.18	.36	.01
ITPA Auditory Reception	.01	.61	.13	.00	.08	.17	.15
ITPA Auditory Association	.01	.64	.18	.00	.06	.11	.22
ITPA Verbal Expression	.01	.49	.22	.00	.10	.19	.28
ITPA Grammatical Closure	.00	.63	.16	.00	.07	.14	.19
ITPA Auditory Memory	.00	.75	.09	.00	.05	.10	.10
NSST Receptivity	.00	.56	.16	.00	.09	.19	.19
NSST Expressivity	.00	.63	.14	.00	.08	.15	.16
Elicited Imitation	.01	.69	.12	.00	.06	.12	.14
WISC Verbal	.00	.99	.00	.00	.00	.01	.00

Notes. ITPA = Illinois Test of Psycholinguistic Ability; NSST = Northwestern Syntax Screening Test; WISC = Wechsler Intelligence Scale for Children. All test values analyzed as raw scores except where otherwise indicated.
[a] Partially confounded variances; estimated using midpoint of possible range under the constraints of the $E(MS)$ equations (see Footnote 8).

work. A components of variance analysis can be applied to virtually any data that can be appropriately put in analysis of variance format, and will yield estimates of at least some of the variance components corresponding to the factors defined in the design. If one of those factors can be interpreted as a criterion contrast, some useful measurement sensitivity information can be gleaned even if specific measurement facets are not included in the design. The data for this example consisted of scores on nine standardized language tests or subtests for a group of about 700 language disordered children in the Los Angeles County Schools special education program. Each child in this group was tested at entry into the program and retested approximately 2½ years later. Though each of the measures was supposed to index a somewhat different aspect of language performance, a factor analysis showed that, for this population, they shared so much common variance as to seem virtually interchangeable (Schery, 1981).[7]

The test–retest comparison was treated as a criterion group contrast. That is, it was judged that 2½ years of development and learning produce an improvement in performance great enough to be of practical significance, even for these language disordered children who progress slowly. Thus any measures useful for program evaluation should, at the least, be capable of detecting this criterion difference and, in fact, should be sensitive enough to detect less substantial differences.

Table 3 shows the variance components (as proportions of total variance) for each of the nine individual measures in the language battery. Also, for purposes of comparison, Table 3 shows raw scores versus scaled scores for one measure (Peabody) and includes a verbal IQ measure. The variance components were calculated using the $E(MS)$ formulations for a fixed effect, repeated measures design with a blocking factor. The blocking factor, sex, was included to examine its potential for reducing the large between-subjects variance expected on these measures and was treated as a fixed effect variable. Two variance components are separately confounded with $\hat{\sigma}^2_\varepsilon$ in this design, but the simultaneous equations so constrain them that order of magnitude approximations could be easily derived by taking the midpoint between the highest possible value and the lowest possible value for each.[8]

[7] The loadings on the first principal component ranged from .75 to .89, with a mean of .82.

[8] No adequate general solution exists for disentangling such confounding. In this particular case, an approximation was possible because of the arithmetic of the specific situation. For example, on the Peabody raw scores the crucial equations were $\hat{\sigma}^2_\varepsilon + 2\hat{\sigma}^2_\pi = 428$ and $\hat{\sigma}^2_\varepsilon + \hat{\theta}^2_{\gamma\pi} = 67$. Since no variance can be negative, both $\hat{\sigma}^2_\varepsilon$ and $\hat{\theta}^2_{\gamma\pi}$ must range somewhere between zero and 67. Given that range for $\hat{\sigma}^2_\varepsilon$, then $\hat{\sigma}^2_\pi$ clearly must range be-

Though this design includes no explicit measurement facets and does not follow the form illustrated in Table 1, it nonetheless provides useful information about the sensitivity of the measures.

1. The criterion contrast accounted for about 15% of the total variance for the typical language measure in Table 3, an ES_m of .18. This is a relatively large effect, as it should be for a 2½-year period of learning and development. The criterion contrast was considerably smaller, however, when adjusted for shorter time periods that were more likely to represent true program effects. For example, a treatment effect equivalent to the average half year of growth, roughly one-fifth the mean difference between the criterion groups, constituted less than 1% of the total variance on most of the measures (e.g., $.2^2 \times 15\%$). The sensitivity of the various individual measures to the criterion contrast ranged considerably. At one extreme, the verbal IQ measure had an ES_m of .0, indicating, as would be expected, that it was completely insensitive to language development. At the other end of the spectrum, one subtest of the Illinois Test of Psycholinguistic Ability (ITPA) had an ES_m of .28, with 22% of the measurement variance responding to the criterion contrast. The results on the Peabody Picture Vocabulary test were particularly interesting with regard to measurement sensitivity. When analyzed in raw score form, the ES_m was a respectable .16. When the scores were scaled according to the age norms for normal children, however, the scale unit for these developmentally slow children became much too coarse to reflect their language gains.

2. By far, the largest variance component on these measures was the between-subjects variance, $\hat{\sigma}^2_{\tau}$. Indeed, it so overshadowed the criterion contrast variance, itself based on a large criterion difference (in practical terms), that there can be little doubt that these measures were oriented toward individual differences measurement. The verbal IQ measure represented an extreme: Virtually all of its variance was individual differences. For all measures, the blocking factor, sex, was clearly useless as a means to control a nontrivial portion of the subject heterogeneity.

3. Although the repeated measures portion of the design represented a test–retest interval of 2½ years, it can be examined as an indicator of reliability, albeit a lower bound at best. The residual error terms and the components representing interaction with between-subjects differences were relatively small, indicating that the measures can be expected to be quite reliable. Indeed, if the appropriate variance components for the language measures are entered into Equation 2, ignoring the long test–retest interval, the resulting reliability coefficients are on the order of .74, quite respectable under the circumstances. The language measures provide quite a contrast with the delinquency measure used in the previous example. The insensitivity of the delinquency measure was shown to result from the large residual error component, that is, unreliability, while subject heterogeneity was relatively modest. The language measures reversed that pattern—reliability was high but large subject heterogeneity on the measures reduced the overall sensitivity to treatment effects.

4. The statistical power that can be expected using the language measures depends on the study design and the measure chosen. To take the most extreme cases, a study in which the researcher expected a program effect equivalent to ½ year of language development and change (one-fifth the difference between criterion group means) and used the least sensitive measure (ITPA Auditory Memory) and a simple treatment versus control group design, would require well over 1,000 subjects in each group to have .90 power to detect the target contrast at the .05 alpha level. (The denominator for Cohen's f in this case combines all variance components except the criterion contrast itself.) Using the most powerful measure (ITPA Auditory Association) in a repeated measures design, only about 120 subjects per group are needed to have the same statistical power. (In this case, the subject constant variance is removed from the denominator of Cohen's f). Note that the most sensitive measure, in terms of ES_m (ITPA Verbal Expression) was not the most powerful in this design because

tween (428 − 67)/2 and 428/2. Whatever the actual values, we can be sure that $\hat{\sigma}^2_{\tau}$ and $\hat{\theta}^2_{\tau\tau}$ are small relative to $\hat{\sigma}^2_{\tau}$.

more of its extraneous variance was concentrated in relatively uncontrollable residual error rather than subject heterogeneity, which lends itself more readily to statistical control.

Conclusions

Measurement sensitivity is a crucial measurement property for research directed at the detection of treatment effects, particularly in applied contexts, where treatments are often weak and measurement conditions may be uncontrollably noisy. This article has suggested that, under such circumstances, a researcher should perform a preliminary measurement assessment study in order to determine the characteristics of the candidate measures for the specific application that is planned. The measurement study accomplishes two purposes. First, it alerts the researcher to a situation in which the proposed measures are not adequate for detecting the expected treatment effect. More constructively, it provides the basis for an exploration of various modifications in the measurement and design regimen that might potentially alleviate any deficiencies in measurement sensitivity. For example, if low statistical power is considered in isolation, the researcher may feel that the only available options are to increase the number of subjects in the design or to relax the alpha level set for statistical significance. When the same problem is considered within the measurement study framework, other strategies and combinations of strategies are more readily discerned, for example, selection of a more sensitive measure, aggregation of several measures to improve reliability, or judicious use of covariates or blocking variables. These are important alternatives, particularly under field conditions, where there may be firm limits on the number of subjects available for the research.

Even if the advantages of a measurement assessment study are acknowledged, it might appear that such a study would be impractical to accomplish as a routine part of applied research. In fact, the essential ingredients for a minimal measurement study are often available with little extra effort. Many circumstances will offer definable groups that differ at a practical level on the construct of interest and hence provide possible criterion groups. Indeed, the ubiquity of such 'nonequivalent' groups in applied research is evidenced by the attention that has been given to the special problems of design and analysis they pose (cf. Cook & Campbell, 1979). A repeated-measures comparison can often be made using archival data, as in the examples in this article, or using a double pretest prior to treatment implementation. Less desirable, but still useful, is the strategy of making the measurement study an adjunct to the application study rather than a preliminary. This can be done, for example, simply by adding a criterion group to the primary design so that the criterion–control group contrast constitutes a measurement study and the treatment–control group comparison tests the treatment of interest.

The two examples from program evaluation data presented here illustrate the features of a measurement study analysis and the kinds of information it yields. Perhaps more important, those examples demonstrate the very great difference measurement sensitivity can make in the detection of program effects under field conditions. Commonly used measures in the areas of juvenile delinquency prevention and special education showed little response to program effects of a modest but worthwhile order of magnitude. A researcher who has not carefully considered the expected effect size and the sensitivity of the measures under the conditions of application could easily conduct an application study that had virtually no chance of finding effects even if they were present. A preliminary measurement assessment study provides researchers with some assurance that they are not looking through the wrong end of the telescope when searching for treatment effects.

References

Aiken, L. R. Note on sensitivity: A neglected psychometric concept. *Perceptual and Motor Skills*, 1977, *45*, 1330.

Boruch, R. F., & Gomez, H. Measuring impact: Power theory in social program evaluation. In L. Datta & R. A. Perloff (Eds.), *Improving evaluations*. Beverly Hills: Sage, 1979.

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), *Criterion-referenced measure-*

ment: The state of the art. Baltimore: Johns Hopkins University Press, 1980.

Carver, R. P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist,* 1974, *29,* 512–518.

Carver, R. P. The Coleman report: Using inappropriately designed achievement tests. *American Educational Research Journal,* 1975, *12*(1), 77–86.

Cascio, W. F., Valenzi, E. R., & Silbey, V. More on validation and statistical power. *Journal of Applied Psychology,* 1980, *65,* 135–138.

Cleary, T. A., Linn, R. L., & Walster, G. W. Effect of reliability and validity on power of statistical tests. In E. G. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology,* San Francisco: Jossey Bass, 1970.

Cohen, J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1977.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally, 1979.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, *52,* 281–302.

Dwyer, J. H. Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin,* 1974, *81,* 731–737.

Epstein, S. The stability of behavior: II. Implications for psychological research. *American Psychologist,* 1980, *35,* 790–806.

Freeman, M. F., & Tukey, J. W. Transformations related to the angular and the square root. *Annals of Mathematical Statistics,* 1950, *21,* 607–611.

Gaebelein, J. W., Soderquist, D. R., & Powers, W. A. A note on variance explained in the mixed analysis of variance model. *Psychological Bulletin,* 1976, *83,* 1110–1112.

Glass, G. V., & Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal,* 1969, *6,* 403–414.

Green, R. S., Nguyen, T. D., & Attkisson, C. C. Harnessing the reliability of outcome measures. *Evaluation and Program Planning,* 1979, *2,* 137–142.

Halderson, J. S., & Glasnapp, D. R. Generalized rules for calculating the magnitude of an effect in factorial and repeated measures ANOVA designs. *American Educational Research Journal,* 1972, *9,* 301–310.

Hays, W. L. *Statistics for the social sciences.* New York: Holt, Rinehart & Winston, 1973.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences.* Belmont, Calif.: Brooks/Cole, 1968.

Levin, J. R., & Subkoviak, M. J. Planning an experiment in the company of measurement error. *Applied Psychological Measurement,* 1977, *1,* 331–338.

Lipsey, M. W., Cordray, D. S., & Berger, D. E. Evaluation of a juvenile diversion program: Using multiple lines of evidence. *Evaluation Review,* 1981, *5,* 283–306.

Myers, J. L. *Fundamentals of experimental design.* Boston: Allyn and Bacon, 1979.

Olkin, I., & Pratt, J. W. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics,* 1958, *29,* 201–211.

Posavac, E. J., & Carey, R. G. *Program evaluation: Methods and case studies.* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Riecken, H. W., & Boruch, R. F. (Eds.), *Social experimentation: A method for planning and evaluating social intervention.* New York: Academic Press, 1974.

Schery, T. K. Selecting assessment strategies for language-disordered children. *Topics in Language Disorders,* 1981, *1,* 59–73.

Searle, S. R. *Linear models.* New York: Wiley, 1971.

Sechrest, L., & Redner, R. Strength and integrity of treatments in evaluation studies. Criminal Justice Evaluation Reports. Washington, D.C.: Law Enforcement Assistance Administration, 1979.

Sechrest, L., & Yeaton, W. H. Empirical bases for estimating effect size. In R. F. Boruch, P. M. Wortman, & D. S. Cordray (Eds.), *Reanalyzing program evaluations.* San Francisco: Jossey-Bass, 1981.

Sutcliffe, J. P. On the relationship of reliability to statistical power. *Psychological Bulletin,* 1980, *88,* 509–515.

Vaughan, G. M., & Corballis, M. C. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin,* 1969, *72,* 204–213.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1971.

Wright, W. E., & Dixon, M. C. Community prevention and treatment of juvenile delinquency: A review of evaluation studies. *Journal of Research in Crime and Delinquency,* 1977, *14,* 35–67.

VII

ANALYZING AND INTERPRETING EVALUATION DATA

Once evaluations have been implemented, the next important step is analyzing and interpreting the data. Some of the issues of concern to evaluators in this stage of the evaluation process are assessing the reliability and validity of data, the type and extent of error variance, the potential threats to accurate interpretation of the data, and the appropriate analysis strategy to employ. While it is impossible to give coverage in this section to all of the relevant data analysis issues, we highlight four of the most important ones to the evaluation community.

Cohen discusses the costs of Type I and Type II errors and proposes various ways to minimize their effects. Many of her suggestions can be employed even after data are collected. Ways to minimize Type I errors (false positives) include using theory to guide data analysis, examining data plots for unusual distributions, minimizing the number of significance tests performed, using lower alpha levels whenever possible, including replications from an independent sample, and not drawing conclusions about effects without testing for significance. In minimizing Type II errors (false negatives), Cohen suggests conducting power analysis before beginning a study, increasing the sample size, removing extraneous sources of variability in independent and dependent variables, increasing the effect size, using the most powerful analytic procedure availabe, and increasing the alpha level. Since the 12 rules that Cohen discusses are relatively easy to understand and implement, this article should be useful to evaluators in their efforts to analyze and interpret valid data bases.

A common problem experienced by evaluation researchers is the differential attrition of subjects from comparison and intervention groups. This is especially salient in evaluations of medical interventions in which subjects in control groups often cross over to intervention groups (e.g., from medical to surgical treatment). Yeaton, Wortman, and Langberg propose a procedure, using worst case assumptions, to estimate the effect of such attrition on the evaluator's ability to detect intervention effects if any exist. By employing this procedure, the quality of inferences made by the evaluator can be enhanced. While the primary focus of the article concerns the evaluation of medical technologies, the authors also discuss general issues in the estimation of attrition effects as well as limitations of this estimation approach.

In most evaluation research, classical statistical tests are commonly used to test theories, hypotheses, and program effects. Unfortunately, however, many

researchers misuse these tests. Sawyer and Peter critically examine the use of statistical tests and conclude that, in many studies, researchers misinterpret and overvalue the results of statistical tests. They propose three means to improve the use of these tests and four research strategies that provide additional critical information about the research issues under study. To improve the interpretation of significance tests, they suggest viewing them as "tests against the null hypothesis" rather than as "tests of significance." In addition, they suggest that researchers, before data analysis begins, specify the effect size to be considered meaningful. They believe that emphasizing the size and substantive significance of results, as opposed to the p values, is more meaningful. The four research strategies for augmenting significance testing are replication, Bayesian hypothesis testing, meta-analysis, and strong inference. This article will provide even the experienced evaluator with valuable information about the appropriateness of common data analysis and interpretation practices.

The final article in this section highlights the importance of examining suppressor effects in conducting data analyses. Lipton and Smith use delinquency research to explicate this issue and report that the failure to examine such effects often leads to false conclusions. In our reading of the recent evaluation literature, we identified few evaluators who thoroughly assessed the effects of possible suppressor relationships among variables, perhaps increasing the likelihood of both Type I and Type II errors.

The four articles in this part of the *Annual* address some of the key issues involved in analyzing and interpreting evaluation data. This stage of the evaluation process is the critical linchpin between data collected by evaluators and use of these data by consumers. While appropriate data analysis and interpretation are not the only factors related to the utilization of evaluation findings, they are key variables in the equation, since evaluation data need to be relevant to the people who are apt to use them.

29

To Be or Not To Be

Control and Balancing of Type I and Type II Errors

Patricia Cohen

In the process of interpreting the results of evaluation and other research studies we frequently encounter a dilemma with which our formal training in methodology and statistics has not equipped us to cope. What should be done about findings which just miss formal statistical significance and/or "findings" which, although not really anticipated, pop up as "significant" amid that mass of such tests which we (or rather our computer programs) have routinely performed? Our clinical or substantive self believes that there are many interesting relationships between variables (group differences, correlations) which we would be remiss to overlook in our efforts to sanctify our examination of phenomena by passing a suitably low significance test criterion. On the other hand, our hard-nosed, skeptical scientist self recognizes the real damage to the development of a scientific field done when many "facts" are only tricks played on us by the accidental characteristics of a particular sample.

These alternative errors, as we all learned in our first methodology courses, are respectively Type II error (β) or failing to find an effect or difference in our sample which exists in the population and Type I error (α) "finding" an effect in the sample which does not exist in the population. Even those of us who do not recall being taught that there is an inverse relationship between Type I and Type II errors will probably have experienced the dilemma in the course of data analysis. This dilemma becomes concrete whenever the decision is faced about what to do about that "finding" which is associated with a $p = .06$.

From one point of view, the tradition of scientific evaluation of evidence has well-equipped us for the task of evaluating the risk of generalizing from a given random sample. Our standards are relatively high – we typically tolerate "positive" conclusions only when, on the evidence, we have less than a one in 20 chance of being wrong. Even when we are aware of violating the assumptions underlying the statistical model we are rarely likely to go as high as a 7% risk when we use a 5% criterion. Even this kind of increased risk can be controlled by the simple expedient of lowering the nominal criterion α value.

A much more serious problem is the control of experiment or investigation-wise Type I error. In this day of multivariate investigation it is commonplace to examine literally hundreds of effects. If 5% of those which are nonexistent in the population are attri-

From Patricia Cohen, "To Be or Not To Be: Control and Balancing of Type I and Type II Errors," *Evaluation and Program Planning*, 1982, 5(3), 247-253.

butable to blind fate, how are we to distinguish them from the 5 or 10% of effects which are really present in the population (Shine, 1980)? No "creaming" of the "most significant," highest t or lowest p effects will serve to reliably discriminate the real from the not real. And once having observed an interpretable effect which reaches the α criterion in a study, even the most sophisticated researcher will have difficulties in admitting that it may be due to chance. For example, a seasoned investigator and text author responded recently to expressed concern that one of his unanticipated findings, picked from the context of dozens or perhaps hundreds of tests, might not be true for the population. He asserted that he really couldn't see what could have gone wrong in the study which would have produced an erroneous finding. This assertion is analogous to assuming that, having flipped a coin 5 times, getting 5 tails *necessarily* means either the coin or the flipping method is biased. The case in question is most parallel to seizing upon the 1 coin in 32 that has yielded 5 tails.

Many of us have had the humbling experience of interpreting an unanticipated "significant" finding, only to find that it is in the direction opposite to that implied by the interpretation. Under these circumstances, it is usually possible to find an equally plausible interpretation of the observed direction of the effect. Unfortunately, the ground rules of our publication game insist that we should provide a theoretical framework for our findings without requiring that we report whether that theoretical framework preceded or postdated the analysis. Thus it is nearly impossible to distinguish between possibly serendipitous findings and those in which we may have more faith because the theory preceded the finding. In any case, it is well established (Tversky & Kahneman, 1976) that we human beings have a strong tendency to overgeneralize from limited personal experience and other small samples, a tendency which the acquisition of a Ph.D. does not curb. Once we have found it, a "finding" is likely to gain our allegience to a degree unjustified by the numbers.

Type II errors are perhaps even more painful. It is (justifiably) difficult to get "negative" results published (Cohen, 1962; Rosenthal, 1979). Reviewers may be aware that most studies have relatively low power, or chances of finding nontrivial effects which *do* exist in the population (see e.g., Cohen, 1962; Cohen, 1977; Freiman, et al., 1978). Therefore studies with no positive findings may be seen as among the "unlucky" half of investigations that will fail to find real population effects. Furthermore, no study is totally without flaw in the eyes of outside reviewers. Failures to detect phenomena in which reviewers are likely to believe may easily be attributed to these flaws.

Since, in general, both our personal professional advancement and the advancement of knowledge in a substantive field require the publication of research findings, Type II errors are a real problem. Furthermore, many of us are in the business of developing and evaluating service systems. In this field Type II errors can be a professional tragedy, since they can lead to the abandonment of effective service modes or components.

For all these reasons the following didactic summary of means by which these errors of inference can be minimized is offered.

MINIMIZING TYPE I ERROR

1. Use a theory to guide you: don't "fish." One way to understand the role of a priori theory in avoiding Type I errors is by seeing it as an informal Bayesian procedure. Formal Bayesian statistical procedures take into account the size of relationship (correlation, regression, mean difference, proportional change) which is most probable on the basis of the evidence and beliefs which antedate the investigation at hand. Then current findings are combined with a priori belief to produce a new best estimate of the population effect. This procedure is a marvelously rational way of building sound estimates although, for reasons not relevant here, they have not as yet made major inroads into most areas of behavioral data analyses. When theory is specified before data are collected one is in a quasi-Bayesian situation in which it is appropriate to have greater confidence in those findings which are predicted by the theory and therefore reinforce prior expectations (Overall, 1969).

Obviously theory construction and selection of the central tests of that theory should precede data collection. However, even after the data are collected one need not abandon theoretical efforts. Avoid letting the computer create your theory for you. Computer programs which select variables from independent variable pools are a snare and a delusion. Their "findings" are a snare because the algorithms merely choose the variables which contribute most in the sample, regardless of the fact that their contributions are typically not significantly greater than those of other variables which may be more theoretically revealing. They are a delusion because the printed "significance" level is vastly overstated (Cohen & Cohen, 1975; Shine, 1980).

2. Practice good housekeeping of your data sets. The full distribution of every variable and a number of bivariate plots and tables should be inspected prior to more complex analysis. It is all too easy to overlook outliers in one's data which may be attributable to

coding or punching errors or failure to appropriately recode, as when missing data are given a numerical value. Even very few outliers can produce "significant findings" if extreme enough. Real effects involving extreme cases need careful attention lest the resulting effects such as mean differences, correlations or regressions are mistakenly understood as reflecting the entire sample. Extremely non-normal distributions which result from the presence of outliers are most likely to lead to serious bias in reported α when parametric procedures are applied. It is also sometimes desirable to transform certain variables. It is perhaps the fear of not being able to identify these problems which has led some researchers to prefer simple bivariate to multivariate analysis. It may be true that investigations employing many variables run a greater risk that individual distributions will have received insufficient attention. Conscientious attention to these details is at least as important in multivariate analysis as in bivariate analysis.

3. Minimize the number of significance tests performed per study. Try to use a single test for each substantive issue. Resist the temptation to search for subgroups in which some hypothesized relationship holds when it is not significant on the a priori most appropriate larger group. Combine measures that should be related to the dependent variable by virtue of the same theoretical construct. This combination may be accomplished by creating summary variables or by treating variables in sets (Cohen & Cohen, 1975; Cohen, 1982). In either case, a single overall test of statistical significance may be employed for inference about the presence of a given relationship. Experimentwise error may also be reduced by the use of protected t tests, and/or Bonferroni or other multiple comparison procedures. These are the formal methods that take into account the multiple tests performed within the same study. The protected t test approach requires that before subsidiary single variables are subjected to significance testing the overall summary multiple degree of freedom test must have met the statistical significance criterion. Protected tests need not be confined to the classic multiple group analysis of variance situation, but may also be employed with sets of variables in multiple regression analysis (Cohen & Cohen 1975, chap. 4) and with multiway contingency tables for which a significant χ^2 value on the entire set is a precondition for the comparison of particular cells or combinations of cells.

In the classic analysis of variance model the multiple comparison procedures take into account the number of tests being performed and thus control the experimentwise Type I error rate. An alternative simple procedure which is available in studies of virtually any design is the Bonferroni test (Dunn, 1961). An investigator using this method simply divides the tolerable experimentwise α criterion by the number of tests to be performed and uses the result as the per-test criterion. A variation on this procedure takes into account the fact that not all tests are equally important to the purposes of the study and divides the significance criterion unequally so as to give more power to the most central issues and less power (more stringent tests) to peripheral issues.

4. Use lower levels of α whenever you can afford them. Under circumstances in which you have plenty of statistical power to find effects of the magnitude which are of interest, the simplest way to decrease the rate of false positive findings is to use a more stringent criterion, such as $p < .01$ rather than .05, or .001 rather than .01. The best way to know that these circumstances prevail is to do a formal power analysis. (See Cohen, 1977, for a full presentation of power analysis methods). The most likely circumstances in which power is more than adequate are either when data are available on a very large number of cases (perhaps in the thousands) or when only large effects are relevant. The former case is fairly common in evaluation studies employing record data from large service systems. The latter case would include trials of risky or expensive services for which only very large positive outcomes would justify future adoption. In such a case the fact that one has low power to detect small effects does not matter.

5. Include replications and partial replications whenever possible. Nothing should reassure one more that an unpredicted finding is real than cross-validation on an independent sample. When the sample one has at hand is large enough for statistical power not to be too serious an issue it may be divided into one hypothesis generating half and another half sample in which generated hypotheses can be validated. When researchers follow a program of research on related issues there are often opportunities to carry overlapping variables in such a manner as to produce opportunities for replication. Unanticipated findings should be considered only suggestive until cross-validation in one's own hold-out sample or replication by means of other data in the public domain has been secured. Alternatively, a literature search may produce relevant investigations into the same issue which have been by-products of other studies. When the variables in question can be subsumed within a theoretical framework one may search for partial replication in the form of alternate measures of the same constructs.

Thus, data analysis and publication policy that makes a clearcut distinction between findings which were central to an investigation and imbedded in a priori theory and those which were not seems indicated. The latter group are then explicitly to be seen as hypotheses in need of testing in new data.

6. Don't make conclusions about differences or ef-

fects which have not been tested for significance. Research reports in which a number of hypotheses are evaluated statistically often also discuss other "findings" which are not subjected to such evaluation. Readers who focus on the substantive issues may well take these conclusions to be as warranted by the evidence as findings for which an appropriate level of statistical reliability has been demonstrated. This kind of conclusion may be especially common in circumstances in which some overall effect has been tested and found to meet criteria but appropriate follow-up comparisons have not been made. However, they also tend to be common in circumstances in which large numbers of statistics (especially means or proportions) are presented in tabular form and then discussed without formal statistical evaluation as well as in circumstances in which popular computer program packages do not include the required test.

MINIMIZING FALSE NEGATIVES: TYPE II ERROR

An unfortunate characteristic of every method of decreasing Type I error is that is will inevitably increase the risk of Type II error. There is no way to lower the risk of "finding" unreal effects without lowering the probability of finding real effects. However, fortunately the reverse is not true. Except for the last, none of the following methods result in an increase in false positive findings. The reason for this is that the false negative β risk depends on three parameters: the size of the effect in the population, the size of the sample, and α. One can therefore take steps to maximize the first two without jeopardizing the validity of one's conclusions.

1. Carry out a power analysis before beginning a study. For readers whose memory of power analysis may be somewhat vague it may be worthwhile to review the basic power analysis procedure. As stated, statistical power depends on only three elements: the size of the population effect you are looking for, the size of the random sample you plan to examine, and the selected statistical significance criterion. The effect size may be expressed in a variety of metrics as appropriate to the particular test you plan to apply—a standardized difference between means, a proportion of variance, or a difference between proportions. This effect size can be estimated from the related literature, or it may be determined as the minimum effect which would be of substantive importance, or one may use conventional values suitable to the substantive field (Cohen, 1977). One next selects a significance criterion (typically .05 or .01) and the sample size intended for the investigation. Armed with this material, power may be determined by using a standard source such as Cohen (1977), frequently by no more complicated a procedure than a table look-up.

One practical way to improve the chances of getting positive findings when they exist is to avoid carrying out studies in which one has a poor chance of detecting such effects. If the most closely relevant literature has failed to reveal a significant effect do not assume that by improving a study's methods and measures you will be able to show the effect without increasing the sample size. Similarly, if an effect which would definitely be large enough to be of substantive interest would stand a poor chance of being revealed in your study, *don't do it.*

2. Increase the sample size—almost any way you can. It is usually the case that getting X amount of data on $2n$ cases is much better than getting 2X on n. The latter will inevitably yield both more α errors and more β errors than the former, analytic strategy being equal. The typical analysis of published studies has found them to have had about a 50% chance of detecting medium-sized population effects (Brewer, 1972; Cohen, 1962; Overall, Hollister & Dalal, 1967). This suggests that reporting bias has led to many unpublished studies in which examined effects were not significant (Lane & Dunlap, 1978; Rosenthal, 1979). Surely, it is the case that given the total cost of research an improvement in sample size should produce a large return on the invested costs!

One frequently misunderstood issue is the consequence of unequal group size in multiple group studies on the power to detect real effects, 1-β. It *is* the case that power for detecting between group differences tends to be greatest when group sizes are equal if the overall number of subjects is fixed. That does *not* mean that one should cut down one's observations to the lowest common n. Equal n studies were once favored, not so much on power grounds but because the resulting analysis made for simple computation. Now that most analyses are carried out on computers no such restraint on the total sample size is warranted and it is usually wisest to forego equality of subsample in favor of overall sample maximization.

3. Remove extraneous sources of variability, especially in your dependent variable and primary independent variables. One way to accomplish this is by restricting the population studied in terms of these extraneous variables. The best candidates for restriction by sampling procedure are those variables which are not easy to measure or are too numerous to measure within the study. Typically these may be unknown environmental factors, but in experimental studies they may also include genetic factors and a great variety of unmeasured common history variables. Sampling may be limited to specified populations, sibs or in the case of animal studies, litter-mates. The diffi-

culty with selecting samples and procedures in ways which limit the number of extraneous variables which effect dependent variables *or* independent variables but not both is that one inevitably also restricts one's ability to generalize from findings. Furthermore, if restricting the population also serves to lower the sample size a power analysis should be done to confirm that the exchange is worth the price.

The second way of removing extraneous (uncorrelated) effects from independent variables (IVs) or dependent variables (DVs) is by measuring these variables and taking them into account in the analysis. There are certain classes of variables which are especially likely candidates for these analyses—the so-called "lurking" variables (Joiner, 1981). One type of lurking variable is variation in the source of data: informant, rater, observer, or instrument effects. Of course, it is not always the case that these will be correlated with IV only or DV only—they may frequently be correlated with both. In that case their removal will not necessarily increase your power to find a statistically significant effect but may only improve the validity of your inferences from the findings. Time related variables are another major subset of lurking variable. Secular (i.e., time) trends over the course of data collecting and diurnal or seasonal effects may be present far more often than we think to look for them. Age is another major lurking variable that we are much more likely to routinely investigate.

There seems to be some considerable confusion over the role of matched subject designs and in particular the effect of these designs on statistical power. The advantage of such designs is that one may control for a number of difficult-to-measure individual differences simultaneously by the matching procedure. The disadvantage is that one is still left with within-group differences which may contribute to dependent variable variance. In order to remove the effects of these variables from dependent variables some form of statistical partialling is required. If the process of matching has caused us to lose many potential subjects from the "pool" with consequent smaller study n it is probably unwise to match. (We omit discussion of other problems of matching such as regression to different means and unrepresentativeness because they are not relevant to power issues.) Perhaps the strongest case for matched subject designs can be made in those cases in which sample size is fixed by the high per subject cost for inclusion in the study. In such a case, a power analysis may reveal a matched subject design to be most powerful.

4. Increase the effect size. Although the effect size in the population may be conceived of as fixed, the sample effect size may be increased in several ways, including the following: (a) Maximize the variance in your major independent variables. Just as restriction of range (variance) of an independent variable produces smaller effects on the dependent variable, so choosing a sample or treatment with a large range (variance) on the independent variable will produce larger effect sizes. When the total sample size is fixed, an over-representation of extreme groups, including omission of cases at the center of the distribution will be most powerful (Abrahams & Alf, 1978; Alf & Abrahams, 1975; Feldt, 1961). Indeed, such over-representation is the rationale behind most plans for sample stratification, in which extreme groups are often over-represented to provide sufficient n for reasonably powerful comparisons. (b) Improve the reliability of your measure. By classic reliability theory one's variables are made up of true and error components. Error is, by definition, uncorrelated with other error and with true components. Therefore, it minimizes and obscures effects by contributing to a variable's variance without contributing to its covariance. Improving construct measurement is one of the easiest and cheapest ways to improve power in many studies. Since reliability generally increases as a function of the number of items, attention to this issue in the planning phase of a study can produce substantial returns at the cost of a few minutes or even seconds per subject (Cleary & Linn, 1969; Cleary, Linn & Walster, 1970). (c) Improve or enlarge the generalizability of your measures (Cronbach, Gleser, Nanda, Rajaratnam, 1972). It is sometimes tempting to get preciously specific in one's measurement and related theory. As a rule, independent variables will have larger effect sizes when they are more general rather than very specific. Thus, overall IQ will tend to show larger relationships with dependent variables than will any one of its subtests and social class will tend to have larger effect sizes than education. This rule also tends to be true of dependent variables, so that overall symptomatology is often more predictable than specific syndromes. (d) Proximal variables will have larger effect sizes than distal ones. By proximal variables we would include both closeness to the dependent variable in time and closeness in terms of subject matter. Thus, my status this week will generally predict my status next year better than will last year's status. Attitude about some issue or behavior will predict the parallel behavior better than will some other attitude or trait. This assertion is quite obvious, but it deserves consideration when power is being assessed and measures are being selected. (e) Combine items or tests which are meant to get at the same issues. Regardless of one's data analytic technique it is generally true that one "uses up" one degree of freedom for each additional independent variable. Even more important, inclusion of correlated predictors tends to increase the standard error of each IV and therefore lower power. Investigators are often reluctant to combine items on the grounds that they

don't know which will turn out to be more important and therefore have no clear cut a priori optimal weighting scheme. However, there is very good news at hand. Within a very substantial range of optimal weights it "don't make no nevermind" (Dawes & Corrigan, 1974; Wainer, 1976); weighting variables equally will work just about as well as the optimal scheme and frequently better than a sample derived scheme on cross-validation. Therefore, one may often account for about the same variance while spending a single *df* as one would with several (say *w*) *df* and one's effect size per variable will be approximately *w* times as large and more stable. (f) Don't squander true variance by combining adjacent categories or scores on a variable. In contrast to the previous rule in which we talk about combining several items with little loss in total variance and a substantial gain in power, here we are talking about combining scores on a single variable before carrying out analysis. One typical situation is that involving a Likert scaled item. An investigator may think that the respondent or rater differences between "sometimes" and "often" are probably not very reliable and decide to combine these categories. Or in the interests of analytical ease a variable may be dichotomized at its median or some other value. Whenever it is the case, as it nearly always is, that *on the average* subjects scored on the higher of two combined categories are in fact higher on the true variable being imperfectly measured, the resulting effect size will be lower than that employing the ungrouped data. In some cases the consequential loss of power will be minor, but it need not be. In the extreme case in which a normally-distributed continuous variable is reduced to a dichotomy the consequent effect size (r^2) will be less than two-thirds as large as the continuous variable if it is cut at the mean and only 44% as large if it is cut one *s.d.* away from the mean (Cohen, 1982, Note 1). The consequent increase in β may well be devastating. Dichotomizing the dependent variable as well as the independent variable results in further substantial power loss.

5. Use the most powerful available data analytic procedure. It is hard to give general rules for which no exceptions can be found. However, it seems safe to say that when assumptions are even approximately met parametric procedures tend to be more powerful than nonparametric procedures. When distributional assumptions are grossly violated, nonparametric procedures *may* be more powerful than parametric procedures. A sound alternative is to consider appropriate transforms for such troublesome variables (see Cohen & Cohen, 1975; Chap. 6) and then proceed to analysis with parametric statistics. In cases in which variables are strictly categorical, log-linear analysis and simpler contingency table-based analysis should have optimal power and most appropriate significance tests. However, when a variable's ordered categories are grouped to produce large enough cell frequencies for categorical analysis there will often be a considerable loss of power compared to investigation of the fully-measured variable. In addition, methods that fail to take into account the ordering of categories will tend to be much less powerful than those that do. This is again because of the additional degrees of freedom required in the representation of such variables in the categorical model.

6. Increase α. Clearly this is something of a last resort since it will increase one's chance of erroneously rejecting the null as well as the chance of correctly doing so. However, there are circumstances in which it may be appropriate to raise one's criterion to $\alpha = .10$ or even higher. When it is not possible to increase one's sample size (because of the paucity of the population) or not advisable to delay conclusions until a sufficient sample can be amassed, one may decide to increase α in order to have a reasonable power to detect a theoretically or practically significant effect. Such a circumstance may prevail, for example, in assessing alternative medical treatments of rare conditions. Another circumstance in which one's threshold for accepting a difference may be low is when a choice must be made between otherwise equally desirable alternatives.

SUMMARY

Twelve rules have been given by means of which one can improve the reliability of one's inferences from data. One may minimize the number of claimed findings not actually true of the population by using theory to guide analysis, conscientiously inspecting raw data, minimizing the total number of tests and employing procedures which take into account the multiple comparisons of variables, lowering α, cross-validating and not interpreting nonsignificant results. Unfortunately each of these procedures exposes one to some increase in risk of missing real effects. Fortunately, each of these methods may be employed after data have been collected even if not anticipated in the study's planning.

One way of keeping β low in one's study is to conduct power analyses and abandon or modify plans for studies with insufficient power. Other means of increasing power need to be built into the study from the beginning, namely increase in N, sampling to remove extraneous variance, maximizing IV variance by study design and inclusion of reliably measured, general and proximal variables in the study. However, several steps to maximize power can be taken even after data are collected. When sufficient and relevant items are avail-

able data reduction methods may be employed to produce measures of greater reliability and generalizability. One should use the most powerful data analytic techniques suitable to the data, avoid lumping of discriminated categories and statistically remove irrelevant variance from study variables, especially dependent variables. Finally, if no other means of increasing the power to detect meaningful effects are feasible, it may be appropriate to increase one's significance criterion to weight the risks of Type I and Type II error in keeping with their social or scientific costs.

REFERENCE NOTE

1. Cohen, J. *The cost of dichotimization*. Unpublished manuscript, New York University, 1982.

REFERENCES

ABRAHAMS, N. M., & ALF, E. F. Relative costs and statistical power in the extreme groups approach. *Psychometrika*, 1978, *43*, 11-17.

ALF, E. F., & ABRAHAMS, N. M. The use of extreme groups in assessing relationships. *Psychometrika*, 1975, *40*, 563-572.

BREWER, J. K. On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal*, 1972, *9*, 391-401.

CLEARY, T. A., & LINN, R. L. Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, 1969, *22*, 49-55.

CLEARY, T. A., LINN, R. L., & WALSTER, G. W. Effect of reliability and validity on power of statistical tests. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological Methodology*. San Francisco: Jossey-Bass, 1970.

COHEN, J. The statistical power of abnormal and social psychological research: A review. *Journal of Abnormal and Social Psychology*, 1962, *65*, 145-153.

COHEN, J. Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology*. New York: McGraw-Hill, 1965.

COHEN, J. & COHEN, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, N.J.: Erlbaum Associates, 1975.

COHEN, J. *Statistical power analysis for the behavioral sciences*. (Rev. ed.). New York: Academic Press, 1977.

COHEN, J. Set correlation as a general multivariate data analytic method. *Multivariate Behavioral Research*, 1982, *16*(3).

CRONBACH, L. J., GLESER, G., NANDA, H., & RAJARATNAM, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.

DAWES, R. M., & CORRIGAN, B. Linear models in decision making. *Psychological Bulletin*, 1974, *81*, 95-106.

DUNN, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 1961, *56*, 52-64.

FELDT, L. S. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, 1961, *26*, 307-316.

FREIMAN, J. A., CHALMERS, T. C., SMITH, H., & KUEBLER, R. R. The importance of Beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine*, 1978, *229*, 690-694.

JOINER, B. L. Lurking variables: Some examples. *The American Statistician*, 1981, *35*, 227-233.

LANE, D., & DUNLAP, W. Estimating effect sizes: Bias results from the significance criteria in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 107-112.

OVERALL, J. E. Classical statistical hypothesis testing within the context of Bayesian Theory. *Psychological Bulletin*, 1969, *71*, 285-292.

OVERALL, J. E., HOLLISTER, L. E., & DALAL, S. N. Psychiatric drug research. *Archives of General Psychiatry*, 1967, *16*, 152-161.

ROSENTHAL, R. The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 1979, *86*, 638-641.

SHINE, L. C. The fallacy of replacing an a priori significance level with an a posteriori significance level. *Educational and Psychological Measurement*, 1980, *40*, 331-335.

TVERSKY, A., & KAHNEMAN, D. The belief in the law of small numbers. *Psychological Bulletin*, 1976, *71*, 105-111.

WAINER, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, *83*, 213-217.

WOODWARD, J. A., & OVERALL, J. E. The significance of treatment effects in ordered category data. *Journal of Psychiatric Research*, 1977, *13*, 169-177.

30

Differential Attrition

Estimating the Effect of Crossovers on the Evaluation of a Medical Technology

William H. Yeaton, Paul M. Wortman,
and Naftali Langberg

Among the various ways that research designs are comprised, perhaps the most troublesome is the differential attrition of subjects from comparison groups during the actual conduct of an evaluation or experiment (Cook and Campbell, 1979). For example, Boeckmann's reanalysis (1981) of the New Jersey negative income tax

AUTHOR'S NOTE: *The work on this report was supported by a grant from the National Center for Health Services Research (HS-04848-01). The authors wish to acknowledge the helpful comments of Dr. Charles Reichardt and two anonymous reviewers on earlier versions of this paper. Requests for reprints should be sent to Dr. William H. Yeaton, Center for Research on the Utilization of Scientific Knowledge, University of Michigan, Box 1248, Ann Arbor, Michigan, 48106.*

From William H. Yeaton, Paul M. Wortman, and Naftali Langberg, "Differential Attrition: Estimating the Effect of Crossovers on the Evaluation of a Medical Technology," *Evaluation Review*, 1983, 7(6), 831-840.

experiment (Watts and Rees, 1977) suggests that differential attrition from the experimental and control groups of both black and Spanish-speaking minorities could account for differences otherwise attributable to the intervention. Likewise, Wortman (1978) has argued that the process of differential attrition was a plausible explanation of the negative effects found in McCord's (1978) 30-year follow-up of the relationship of counseling to subsequent delinquency in a randomized controlled trial.

As a result of this differential attrition process, a well-conceived randomized experimental design may drift toward a quasi-experimental design model with all of its inferential limitations (see, for example, Special Report, 1982). Despite the creative application of statistical procedures to adjust for the resulting nonequivalence between groups (Kenny, 1975; Magidson, 1977; Reichardt, 1979), there is no satisfactory statistical solution to the lack of adherence to the original design protocol.

The most common recommendation of methodologists is to analyze the data from randomized experiments according to the original assignment (Riecken and Boruch, 1974) or "intention to treat" (Peto et al., 1976). This approach is a tradeoff that preserves the design at the expense of a biased estimate of the treatment effect.

THE PROBLEM OF CROSSOVERS

In the assessment of medical technologies, researchers will often confront situations in which specific techniques are preferred by patients due to their association with secondary outcomes that are intrinsically desirable. In instances such as these, it will be particularly difficult to maintain the design protocol. This precise situation confronts researchers interested in the evaluation of the potential benefits of coronary artery bypass graft surgery (CABGS) for patients with coronary heart disease (Wortman, 1981).

Given the consistent finding that angina is relieved in patients receiving surgery (Special Report, 1981), it is ethically problematic to withhold a potential benefit from the medical group. Accordingly, any efforts to compare survival rate between patients who are operated on and those adhering to a medical regimen are greatly complicated by the fact that substantial percentages of medical patients cross over to the group of patients who receive surgery. Furthermore, the migration is

unidirectional since it is not possible for patients to cross over from the surgical to the medical groups once they have received surgery.

In a synthesis of results from 25 controlled trials of CABGS (Wortman and Yeaton, 1983), the crossover rates from the medical to the surgical group were found to be quite sizeable, ranging up to 45.0% in randomized controlled trials, with a mean rate of 21%. Compounding this problem is the systematic nature of the attrition. It is typically those medical patients with the worst prognosis, that is, those with the most severe angina and imminent danger of heart attacks who become crossovers (Murphy, Hultgren, Detre, Thomsen, and Takaro, 1977: 1470). As these researchers have noted:

> Medical nonadherers are frequently assumed to be treatment failures. Although this result was true in approximately 54 per cent of our patients (unstable angina, 19 per cent, and progressive angina, 35 per cent), patient or physician preference prevailed in some cases.

ESTIMATING THE EFFECT OF CROSSOVERS

Any inferences that are made from controlled trials of CABGS must weigh the extent to which patients in the medical group have crossed over to receive surgery. Clearly, the effect of the loss of the most ill patients from the medical group is to raise the survival rate in the medical group. Whether crossovers are dropped from the study or included, as most evaluation methodologists recommend, the extent of the increase is not readily apparent (Wortman, 1981).

In fact, the most common biostatistical research practice is to consider crossovers as an endpoint, that is, as no longer in the study, at the time they receive surgery. This would bring the medical group survival rate closer to that for the surgical group, if one assumes that surgery is beneficial, an assumption consistent with the data. Consequently, the crossover problem will underestimate any potential benefit due to surgery. If one follows the recommendation to include crossovers in their originally assigned group, then the effectiveness of surgery will also be underestimated (again assuming it is beneficial). Neither method then can overcome the effects of differential attrition and treatment diffusion to produce an unbiased estimate of effect.

The worst case approach uses a general strategy of determining the maximum degree of influence attributable to a particular factor and

thus the factor's likelihood of contributing to the difference obtained. In this way it resembles the sensitivity analysis used by economists in cost-effectiveness and cost-benefit analyses (Weinstein and Stason, 1982) to ascertain the consistency of conclusions under various conditions such as extreme and intermediate values of the discount rate in determining present costs. If conclusions are preserved under worst case assumptions, one can place considerable confidence in the validity of the results. If conclusions are contingent upon the values assumed, one can judge "up front" the plausibility of the value that is needed to maintain consistent conclusions.

Though the very nature of crossovers makes it impossible to determine exactly the effect of such attrition on survival rate, it is possible to estimate the maximum influence that crossovers would have on the mean survival rate of the medical group. By calculating this maximum influence attributable to crossovers, researchers would be in a greatly improved position for defending inferences about differences between medical and surgical groups. Since this procedure is likely to overadjust for attrition, it could provide convergent evidence if it agrees with the more traditional estimates, in their direction if not their magnitude.

This estimation procedure necessitates some very specific assumptions, however. We will assume that only those patients in the medical group with the worst prognosis cross over to receive surgery, and that these patients are considered as an endpoint at the time they cross over. With regard to the distribution of survival rates of medical patients, this assumption implies that the tail of the distribution is truncated at precisely the point that will eliminate the exact percentage of patients who cross over. This means that the area under the distribution curve that is eliminated will coincide with the percentage of patients who cross over. We further assume a distribution of composite measures of health status that reflects the probability of survival for medical patients after the time patients in the surgical group have received CABGS. We also hypothesize that the measure is standardized normal (so that the mean equals zero and standard deviation equals one), allowing us to use standard formulae to calculate the mean of the truncated distribution. Though truncated, standard normal distributions are also employed by economists (Stromsdorfer and Farkas, 1980), they are commonly used to correct statistically for attrition bias in ANOVA and regression models (Hansman and Wise, 1979) rather than to form a basis for a worst case solution.

Formally, if f is a density function defined by

$$f(x) = \frac{e^{-x^2/2}}{SQ\ RT(2\pi)}.$$

where x assumes any real number value, then for any given percentage p of crossovers, the area under the normal curve yielding this percentage p can be found by integrating the normal curve density function from minus infinity to that point z on the x-axis which yields p as the result of the integration:

$$\int_{-\infty}^{z} \frac{e^{-x^2/2}}{SQ\ RT(2\pi)} = p.$$

The mean of interest (the mean of the truncated distribution) will be:

$$\left(\frac{1}{1-p}\right)\int_{z}^{\infty} \frac{xe^{-x^2/2}}{SQ\ RT(2\pi)} = \left(\frac{1}{1-p}\right)\left[\frac{-e^{-x^2/2}}{SQ\ RT(2\pi)}\right]_{z}^{\infty}$$

$$= \frac{e^{-z^2/2}}{(1-p)\ SQ\ RT(2\pi)}.$$

Given various crossover rates p, one can use a table of standard normal deviates to determine the corresponding z-value on the abscissa. These two constants can then be substituted into the above result for the mean of the truncated distribution to ascertain the magnitude of shift of the mean.

Examples

For purposes of illustration, several p values and the corresponding means of interest are displayed below. When plotted, the relationship appears to be essentially linear.

p = .01	mean = .03
p = .05	mean = .11
p = .10	mean = .19
p = .15	mean = .27
p = .20	mean = .35
p = .21	mean = .36
p = .25	mean = .42

Thus, the crossover rate of 21% (when p = .21) found in the authors' synthesis of the results of controlled trials of CABGS (Wortman and Yeaton, 1983 would be associated with a mean shift of .36 (a 36% increase in the standardized mean value), the maximum change attributable solely to crossovers.

In some instances it will not be necessary to translate shifts calculated in standardized units to their equivalents in unstandardized terms. Measures of effect size (Glass, McGaw, and Smith, 1981) are calculated by dividing mean differences by an appropriate standard deviation and thus are directly comparable to results generated from mean shifts in the standard normal distribution. For example, given an effect size of .50 and a crossover rate of 21%, one can determine the maximum effect of crossover by simply adding .36, the mean shift, to .50, to adjust for the underestimated outcome measure. By comparing this adjusted value to the original value, one can estimate the extent to which a difference between groups is likely to be due to crossovers.

In other cases there will be no immediately obvious standard deviation value by which one can standardize results, but reasonable estimates may be available. For example, in the medical research cited above, survival rates were assumed to be reflected in the distribution of composite measures of health status, and these measures might be used to produce a standard deviation. Another measure of variability might be the standard deviation of the survival means of other similar studies. For example, given a medical group mean survival percent of 65 and a standard deviation of 10 found from a composite index of health status, the maximum effect of a 21% crossover rate would be 3.6 (.36 times 10). Therefore, the survival rate would be increased to 68.6 as a result of crossovers.

In practice, means and standard deviations are available after crossovers have occurred, and researchers will be interested in determining the adjusted mean before the effect of crossovers. In this case, one simply subtracts the product of the percent shift and the standard

deviation from the given mean. As an illustration, again assume a medical group mean survival percent of 65 and a standard deviation of 10 percent, values determined after crossovers. The adjusted mean would be 61.4 (65 −.36(10)), thus allowing the researcher to conclude that the difference between medical and surgical groups would be underestimated by a maximum of 3.6 percentage points as a result of crossovers, assuming that the mean in the surgical group is greater than the mean in the medical group. While in the case of attrition due to crossovers it is obvious that an adjustment must be made in the control group measure, the practice is consistent with the identification of distortion in research that uses historical (Sachs, Chalmers, and Smith, 1982) and other nonrandomized controls (Meier, 1978).

GENERAL COMMENTS

These findings suggest that high crossover rates can substantially increase the mean of the distribution of the control group of medical patients in which crossovers have been eliminated. Consequently, the benefit attributable to surgery would be substantially underestimated in controlled trials of CABGS. While from a statistical point of view mean shifts between 20% and 50% would be considered between small and medium effect sizes (Cohen, 1977), innovative surgeries typically produce modest benefits (Gilbert, McPeek, and Mosteller, 1977) that assume importance through their implementation with large groups of patients. For example, the evidence from randomized controlled trials suggests a benefit of CABGS of less than 5% (Wortman and Yeaton, 1983). Fortunately, the ability to detect these modest benefits is enhanced considerably by the above estimation technique, since the degree to which crossovers may alter a group mean and thus underestimate differences between groups can be determined easily.

Of course, the relationship between the rate of crossovers and the shift in mean survival rate found in actual reports of CABGS will not follow the idealized relationship described above. Distributions may only approximate the normal, and variances will change as a function of the range of diagnostic severity of patients in the medical group. To the degree that the distribution is negatively skewed or the variance is large, the shift in the mean will increase. Also, not all medical patients will cross over at the same point in time, as we have assumed in our calculations.

When there is a high incidence of crossovers early in the follow-up period of a controlled trial, the degree of bias attributable to crossovers will be maximal. The longer the delay period before patients begin to cross over from the medical group, the closer one approaches the case of an intact control group. Furthermore, we have assumed that only the worst medical patients cross over, an assumption not likely to be true in actual practice. However, the closer the mix of crossovers approximates the case in which only the worst medical patients cross over, the closer the mean shift will approximate the maximum shift shown in this report.

The problem of crossovers in the assessment of effectiveness of CABGS is illustrative of the differential attrition process that plagues evaluation research. The "solution" presented here is applicable to those instances in which the differential attrition process selects subjects or patients in the same manner as they were selected in this report. Specifically, if a differentially attrited subgroup of persons is homogeneous on some measure(s) of status (such as health in this report, occupation in the McCord (1978) study, and ethnicity in the Watts and Rees (1977) volume) that correlates with the outcome variable in question, then the findings of this report are relevant. Obviously, the degree of direct relevance will depend on the match of the groups resulting from the differential attrition process to the pertinent assumptions upon which our estimates are based: attrition of only worst case persons from one of the comparison groups and the shape of the relevant distributions. Other potentially important factors such as the strength of correlation between the status and the outcome measures may compensate for departures from worst case assumptions, however.

While the emphasis of this report has been on the accurate interpretation of research results in studies plagued by differential attrition, the findings may also be used in planning studies. Briefly, if one is armed with knowledge from past studies with regard to the expected rate of crossovers, precise estimates can be made of the degree to which the magnitude of difference between groups is likely to be altered. Accordingly, sample sizes can be either increased or decreased to reflect smaller or larger differences between groups, thus enhancing the power of experiments or diminishing their expected costs.

Despite the shortcomings associated with idealized data, the relationship between crossover rates and survival presented in this report will allow researchers to estimate more accurately the potential influence of crossovers, and thus to improve the quality of their inferences. Given the uncertainties in interpreting the results from flawed research studies, it is

important that investigators acknowledge the potential bias caused by such "threats to validity" (Campbell and Stanley, 1966). These problems are much more common in the applied field studies characteristic of program evaluation and medical technology assessment. Worst case assumptions can provide a bound for a treatment's impact and indicate the extent to which the estimate of effect is sensitive to bias.

REFERENCES

BOECKMANN, M. E. (1981) "Rethinking the results of a negative income tax experiment," pp. 341-355 in R. F. Boruch, P. M. Wortman, D. S. Cordray, and Associates (eds.) Reanalyzing program evaluations: Policies and Practices for Secondary Analysis of Social and Educational Programs. San Francisco: Jossey-Bass.

CAMPBELL, D. T. and J. C. STANLEY (1966) Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally.

COHEN, J. (1977) Statistical Power Analysis for the Behavioral Sciences (Rev. Ed.) New York: Academic Press.

COOK, T. D. and D. T. CAMPBELL (1979) Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

GILBERT, J. P., B. McPEEK, and F. MOSTELLER (1977) "Progress in surgery and anesthesia: benefits and risks of innovative therapy," in J. P. Bunker, B. A. Barnes, & F. Mosteller (eds.) Costs, Risks, and Benefits of Surgery. New York: Oxford Univ. Press.

GLASS, G. V, B. McGAW, and M. L. SMITH (1981) Meta-Analysis in Social Research. Beverly Hills, CA: Sage.

HAUSMAN, J. A. and D. A. WISE (1979) Attrition bias in experimental and panel data: the Gary income maintenance experiment. Econometrica 47: 455-473.

KENNY, D. A. (1975) "A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design." Psych. Bulletin 82: 345-362.

MAGIDSON, J. (1977) "Towards a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: a general alternative to ANCOVA." Evaluation Q. 1: 399-420.

MEIER, P. (1978) "The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine." pp. 3-15 in J. M. Tanur, et al. Statistics: A Guide to the Unknown. Berkeley: Holden-Day.

McCORD, J. (1978) "A thirty-year follow-up of treatment effects." American Psychologist 33: 284-289.

MURPHY, M. L., H. N. HULTGREN, K. DETRE, J. THOMSEN, and T. TAKARO (1977) "Special correspondence: a debate on coronary bypass." New England J. of Medicine 297: 1470.

PETO, R., M. C. PIKE, P. ARMITAGE, N. E. BRESLOW, D. R. COX, S. V. HOWARD, N. MANTEL, K. McPHERSON, J. PETO, and P. G. SMITH (1976) "Design and analysis of randomized clinical trials requiring prolonged observation of each patient: introduction and design." British J. of Cancer 34: 585-612.

REICHARDT, C. S. (1979) "The statistical analysis of data from nonequivalent group designs," in T. D. Cook and D. T. Campbell (eds.) Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally College Publishing Company.

RIECKEN, H. W. and R. F. BORUCH (1974) Social Experimentation: A Method for Planning and Evaluating Social Intervention. New York: Academic Press.

SACHS, H., T. C. CHAIMERS, and H. SMITH, JR. (1982) "Randomized versus historical controls for clinical trials." The American J. of Medicine 72: 233-240.

Special Report (1981) "National Institutes of Health Consensus Development Conference Statement." New England J. of Medicine 304: 680-684.

Special Report (1982) "The anturane reinfarction trials: reevaluation of outcome." New England J. of Medicine 306: 1005-1008.

STROMSDORFER, E. W. and G. FARKAS (1980) "Methodology," pp. 32-41 in E. W. Stromsdorfer and G. Farkas (eds.), Evaluation Studies Review Annual, Vol. 5. Beverly Hills, CA: Sage.

WATTS, H. and A. REES (1977) "The New Jersey Income Maintenance Experiment," Vol. 2. New York: Academic Press.

WEINSTEIN, M. C. and W. B. STASON (1982) "Cost-effectiveness of coronary artery bypass surgery." Circulation 66: III-56-III-66.

WORTMAN, P. M. (1978) "Differential attrition: Another hazard of follow-up research." American Psychologist 33: 1145-1146.

WORTMAN, P. M. (1981) "Randomized clinical trials," pp. 41-60 in P. M. Wortman (ed.), Methods for evaluating health services. Beverly Hills, CA: Sage.

WORTMAN, P. M. and W. H. YEATON (1983) "Synthesis of results in controlled trials of coronary artery bypass graft surgery," in R. J. Light (ed.), Evaluation Studies Review Annual, Vol. 8. Beverly Hills, CA: Sage.

31

The Significance of Statistical Significance Tests in Marketing Research

Alan G. Sawyer and J. Paul Peter

Two basic types of empirical evidence used in hypothesis testing in marketing research are observations of covariation and observations of differences between groups. This evidence usually consists of sample data, and the acceptability of the evidence is based almost inevitably on classical statistical significance tests. However, a review of marketing research texts and a variety of marketing research articles leads to the conclusion that, in both theory and practice, the logic of statistical significance testing is sometimes misinterpreted in the marketing literature. Perhaps because of this misinterpretation, marketing researchers may seriously overvalue the role of classical inferential statistics in the research process.

The purpose of our article is to examine the interpretation and value of statistical significance testing and to offer recommendations to improve the quality of hypothesis testing in marketing research. Although the issues we discuss pertain directly to data from experimental research, most of these issues also apply to correlational and other data. Many examples, including several from the marketing literature, illustrate our recommendations. We hope to persuade more marketing researchers to follow their lead. Though we are not arguing against using classical inferential statistics for what they were designed to do, we are concerned with the tendencies to endow them with capabilities they do not have and to utilize them as the sole approach to analyzing research data.

These problems are not exclusive to research in marketing. Writers in psychology (e.g., Bakan 1966; Lykken 1968), sociology (e.g., Henkel 1976; Selvin 1957), and education (e.g., Carver 1978) have argued that these misinterpretations also pervade their disciplines. In fact, tests of statistical significance seem to be relied upon and often misused in all the social sciences. In comparison, perhaps because of the more highly developed theory, more reliable measurement techniques, and greater opportunity to control nuisance variables, researchers in the physical sciences often forego inferential statistical tests and instead focus directly on the data themselves. Although marketing phenomena may not lend themselves to this approach (Peter 1983) and the theory, measurement, and research procedures in marketing may never develop sufficiently for us to follow the analytical practices of the physical sciences, researchers should be more aware of the limitations of most inferential statistics and the value of augmenting them with other information and other research approaches.

INTERPRETATION OF STATISTICAL SIGNIFICANCE TESTS

For a proper interpretation of the meaning of a statistically significant result, the assumptions of the classical statistical significance testing model must be understood. A primary assumption is that the null hypothesis (e.g., no difference between treatment effects, no association between variables) is true and any observed differences

From Alan G. Sawyer and J. Paul Peter, "The Significance of Statistical Significance Tests in Marketing Research," *Journal of Marketing Research*, 1983, 20, 122–133.

or associations are the result of sampling error. For example, a statistically significant mean difference at $p \leq .05$ tells us that if we sampled many pairs of groups from the same hypothetical population, we would expect to get a difference as large as the observed result or larger with no more than 5% of the groups as the result of sampling error, given that the null hypothesis is true. In general terms, a statistically significant result is one which occurs rarely if the null hypothesis is true.

Many writers in social science have commented on the failure of researchers and textbook writers to interpret statistical significance correctly. In a recent summary of much of this work, Carver (1978) discusses three common misinterpretations, all three of which can be found in the marketing literature. These three misinterpretations are that a statistically significant result indicates (1) the probability that the results occurred because of chance, (2) the probability that the results will be replicated in the future, and (3) the probability that the alternative hypothesis is true. A fourth misinterpretation involves confusion about the role of sample size and the level of statistical significance.

The Probability of the Null Hypothesis

The first misinterpretation is to view a p-value as the probability that the results occurred because of sampling error or chance fluctuations. For example, $p = .05$ is interpreted to mean that there is a probability of only .05 that the results were caused by chance. However, this interpretation is completely erroneous because (1) the p-value was calculated by assuming that the probability is 1.0 that any differences were the result of chance and (2) the p-value is used to decide whether to accept or reject the idea that the probability is 1.0 that chance caused the mean difference. A p-value of .05 means that, *if* the null hypothesis is true, the odds are 1 in 20 of getting a mean difference this large or larger and the odds are 19 in 20 of getting a smaller mean difference. However, *there is no way in classical statistical significance testing to determine whether the null hypothesis is true or the probability that it is true.* As Cronbach and Snow (1977, p. 52) explain:

> A p value reached by classical methods is not a summary of the data. Nor does the p value attached to a result tell how strong or dependable the particular result is . . . . Writers and readers are all too likely to read .05 as p(H/E), "the probability that the Hypothesis is true, given the Evidence." As textbooks on statistics reiterate almost in vain, p is p(E/H), the probability that this Evidence would arise if the (null) hypothesis is true. Only Bayesian statistics yield statements about p(H/E).

The Probability of Results Being Replicated

A second misinterpretation is that the p-value represents the confidence a researcher can have that a given result is reliable or can be replicated. Basically, this argument is that the complement of the p-value yields the probability that a result is replicable or reliable, e.g., 1 − .05 = .95 probability that results can be replicated. This misinterpretation probably comes from a notion that a statistically significant difference in sample means suggests that the samples are from different hypothetical populations and future samples drawn from these different hypothetical populations will therefore yield equivalent results. However, *nothing in classical statistical significance testing says anything about the probability that the same results will occur in future studies.* Replicating results is a function of how exactly the method is repeated, and some aspects, such as the time of measurement, clearly cannot be identical to those of the original study.

The Probability of Results Being Valid

The third and most serious misinterpretation of classical statistical significance testing is that it directly assesses the probability that the research (alternative) hypothesis is true. For example, a p-value of .05 is interpreted to mean that its complement, .95, is the probability that the research hypothesis is true. Related to this misinterpretation is the practice of interpreting p-values as a measure of the degree of validity of research results, i.e., a p-value such as $p < .0001$ is "highly statistically significant" or "highly significant" and therefore much more valid than a p-value of, say, .05. Again, such a practice is inappropriate. Although it is true, for example, that the greater the difference between group means the greater the chance of obtaining a small p-value, and it is true that such a result may be rarer given the null hypothesis, a statistically significant result cannot properly be construed as a more valid result for at least two reasons.

First, a statistical test is not a complete test of a research hypothesis. Instead it examines only one of many possible operationalizations of a research hypothesis. Thus, it is improper to infer that the research hypothesis is valid without testing and support from a representative sample of operationalizations. Second, a variety of threats to drawing valid inferences are not addressed by statistical tests (Cook and Campbell 1979). Theoretically, the researcher's job is to eliminate or at least to render implausible all of the alternative explanations before accepting the research hypothesis. However, this task is no small matter given the variety of theories and method factors that can be offered to explain any empirical result. When these variables along with possible higher order interactions are considered, the task becomes even more difficult (see Cronbach and Snow 1977). In any event, *rejection of the null hypothesis at a predetermined p-level supports the inference that sampling error is an unlikely explanation of results but gives no direct evidence that the alternative hypothesis is valid.*

Sample Size and the Probability of the Research Hypothesis

A fourth common misinterpretation about statistical testing involves the relationship between sample size and

level of statistical significance. If a given relationship is found to be statistically significant at a given confidence level, it is sometimes implied that more confidence should accompany this result if the study had a large sample size rather than a small one. Rosenthal and Gaito (1963) report that such a conclusion was very prevalent among the research psychologists they surveyed. However, such a conclusion is false. Larger samples do reduce likely sampling error because their estimates more closely approximate the population parameters, but it should also be clear that differences in the amount of sampling error are included explicitly in the computation of statistical significance tests. Thus, there should not be a bias against statistically significant results obtained from properly selected small samples.

Moreover, because effect size is a measure of the strength of the relationship and large effects are more likely to be replicated than small ones, researchers should have more confidence in the study with the smaller sample. Meyer (1974) demonstrates this fact with a Bayesian analysis of binominal data with results for different sample sizes. Meyer's results stem from the simple fact that smaller effect sizes are considered statistically significant with larger sample sizes and that, though a larger sample size helps to reduce sampling error and the resulting higher statistical power of a classical inferential statistic increases the probability of a rejection of the null hypothesis, it does not necessarily increase the probability of a *valid* rejection.

THE VALUE OF STATISTICAL SIGNIFICANCE TESTS

The value of statistical significance testing is severely restricted because it does not accomplish what researchers often want or, perhaps in some cases, assume that it does. Several factors detract from the value of statistical tests. First, the process of statistical hypothesis testing is hardly objective given the many subjective decisions made by the researcher. Second, exact null hypotheses are very rarely true in the population, and researchers typically are very biased against the null hypothesis in their testing procedures. Third, classical statistical significance tests are often uninformative without various descriptive statistics and other inferential tests such as confidence intervals. Finally, classical statistics offer no direct evidence about individual behavior.

The Subjectivity of Statistical Tests

Perhaps a major factor contributing to the perceived value of statistical significance tests is the illusion that they are completely objective. Though such tests are mathematical and precise[1] and provide "a formal and nonsubjective way of deciding whether a given set of data shows haphazard or systematic variation" (Winch and Campbell 1969, p. 143), one should not infer that they are objective tests. The reason is that whether a given statistical significance level is obtained is strongly influenced by subjective decisions by the researcher. As Bakan (1966, p. 426) points out, "the probability of rejecting the null hypothesis is a function of five factors: Whether the test is one or two-tailed, the level of significance, the standard deviation, the amount of deviation from the null hypothesis and the number of observations." The researcher clearly controls the first, second, and fifth factors and can influence the third and fourth. Thus, many obtained results which are not statistically significant can become so by such methods as (1) increasing the sample size, (2) increasing the reliability of the measures, (3) changing *post hoc* the acceptable level of statistical significance (i.e., from .01 to .05), (4) changing from a two-tailed to a one-tailed test, and (5) obtaining better control over nonmanipulated variables. *Because researchers make many subjective decisions that greatly influence the probability of rejecting the null hypothesis, it is misleading to consider the process of statistical significance testing as objective solely because of the objectivity of the mathematics.*

A methodological paradox in social science research relates to the illusion of objectivity (Meehl 1967). Methodological improvements such as increased control, more precise measurement, and a greater number of observations make it *easier* for the social scientist to reject the null hypothesis (and claim support for the alternative hypothesis), whereas such improvements make it more difficult for the physical scientist to reject the null hypothesis. The reason for this paradox is that, in the physical sciences, theory is often used to predict point values and, if used at all, statistical significance tests evaluate the difference between the value predicted by the theory and the value found in the data. In contrast, most social science theories are not developed sufficiently to make point predictions and instead statistical significance tests are used to test all other values against the null hypothesis of zero. Meehl suggests that the use of statistical significance testing in social science thus makes it very difficult to not reject the null hypothesis and the involved theory.

Research Bias Against the Null Hypothesis

Classical statistical significance tests are set up under the assumption that the null hypothesis is true. Such an assumption is, in fact, almost always false, and much well intended marketing research practice is biased against the null hypothesis. First, null hypotheses of no treatment effect or no relationship are almost always false because, in the population, few behavioral variables ever

[1]Some marketing researchers have tried to quantify problems with collected data other than sampling errors (e.g., Brown 1967; Lipstein 1975; Mayer 1970). Though not optimistic about the ability to quantify these many other types of errors, we applaud the effort because it helps point out the obvious limitations of a statistic that precisely quantifies what very often is the least serious of the many threats to accurate estimation (see, for example, Assael and Keon 1982).

have *exactly* a zero mean difference between two groups or an exactly zero correlation with each other. For example, Meehl (1967) reported that 91% of pairwise associations among 45 variables in a sample of 55,000 people were statistically significant, and Bakan (1966) failed to find any statistically insignificant relationships among many tests in a sample of 60,000. Given sufficiently high statistical power, one would expect virtually *always* to conclude that the exact null hypothesis is false. It is no wonder that "statistical significance" has occurred often in recently published marketing research because these studies typically have relatively high statistical power (Sawyer and Ball 1981). We find it frightening to consider how much of the conventional wisdom in marketing is based on little evidence other than statistical significance.

Researchers and publication practices are biased against the null hypothesis. Researchers inevitably expect to reject the null, and publication practices overwhelmingly favor studies which achieve this objective. Greenwald (1975a) describes how researchers are unlikely even to try to publish results of an empirical study that failed to reject the null, and journals are even less likely to accept the few statistically insignificant-result studies that are submitted. In an extensive review, Glass, McGaw, and Smith (1981) determined that "findings reported in journals are, on the average, one-third standard deviation more disposed toward the favored hypothesis of the investigators than findings reported in theses and dissertations" (p. 67).

Such a selection bias toward submitting and/or publishing only statistically significant results leads to the fear that "file drawers" are full of statistically insignificant studies and that the published ones are the only ones that attain conventional statistical significance. Using measures of effect size, Rosenthal (1979) demonstrates how to incorporate the possibility of "file drawer" support for the null hypothesis into calculations of possible Type I error and concludes that, "when the number of studies available grows large or the mean directional Z (effect size) grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out" (p. 640). In contrast, with a small sample of statistically significant studies, relatively few "filed" studies with "insignificant" results would have to exist to yield a net statistically insignificant conclusion. For example, according to Rosenthal, 15 studies with an average effect size of $Z = .50$ have a combined Type I error rate of $p = .026$, but, if there were as few as six other studies showing a mean effect size of .00, the overall set of results would be judged statistically insignificant (i.e., $p > .05$).

After one rechecks the calculations (Rosenthal 1969),[2] the typical reaction to a failure to reject a null hypothesis is to blame the failure on something wrong such as a weak manipulation, a small sample size, or unreliable measurement (McGuire 1973). Even when several failures to reject a null hypothesis are reported in the literature, researchers still cling to the alternative hypothesis as the most likely (e.g., Cartwright 1973). Our suspicion that marketing researchers suffer from a similar bias is based on our inability to recall any instances in which it is widely agreed that a previously hypothesized relationship does not hold. Apparently, *results from statistical significance tests are perceived to be valuable when they support the favored hypothesis but are commonly discounted when they support the null.*

The Need for Descriptive Statistics

A major problem in the use and reporting of classical statistical significance tests is that they commonly appear to dominate or even substitute for the data themselves. Frequently, tables of F-values are discussed before or instead of such vital descriptive statistics as means and confidence intervals. Such priority is clearly misinforming as well as misinformed. *The major results of any empirical study, regardless of whether the prime purpose is description, prediction, or explanation, are the descriptive statistics that indicate the nature and size of any obtained effects.* As Sawyer and Ball (1981) summarize, statistical significance tests do not say anything about the size or importance of an effect. Lower Type I error probabilities do not necessarily imply a larger effect. A very small effect can be statistically significant with a sufficiently large sample; conversely, a sizable effect can be judged statistically insignificant with a very small sample. Effect size can be measured in many ways including R^2, ω^2, and other estimates of the ratio of explained to total variance; alternatively, various expressions of the standardized mean difference between groups such as Z or Cohen's (1977) d values can be used (see also Rosenthal and Rubin 1982).

Statistics and Individual Behavior

A final point that is occasionally overlooked about the value of statistical significance tests is that they focus on aggregate central tendencies and reflect little about individual behavior. One interesting way to illustrate this point is to consider Cohen's (1977) U descriptive statistic which measures the percentage overlap between two distributions. Even with a reasonably large effect size, a large percentage of individuals in two groups will often be essentially similar or ordered contrary to the direction of the overall group mean. For example, Cohen states that a difference as large as .8 of a standard deviation is relatively "large" for much social science research. Even

[2]Lest the reader doubt this, we ask the following question: After having calculated, for example, an F-value that suggests your favored research hypothesis is statistically significant, how likely are you to recheck your figures, make sure your computer format statement was correct, etc.? Alternatively, how many hours have you spent checking and rechecking data that failed to attain statistical significance? Interestingly, Rosenthal (1969) observed that when computational errors occur, nearly three-fourths of those errors are in the direction of the researcher's hypothesis.

with such a difference, however, 52.6% of the two populations are overlapped. Thus, though marketing researchers frequently conclude that, for example, new product adopters are different from nonadopters in a certain way, it is almost always erroneous to conclude that *all* adopters are different from *all* nonadopters in that way and, in most instances, wrong to infer even that *most* adopters are different in a given way. Although this is often the type of conclusion researchers want to draw, a statistical significance test alone does not justify such a conclusion.[3]

RECOMMENDATIONS

We offer several recommendations designed to address the problems discussed and to strengthen hypothesis testing in marketing. First we present three considerations for improving the use of classical statistical significance tests against the null hypothesis. We then describe and illustrate four research perspectives that provide valuable additional information about research questions: replication, Bayesian hypothesis testing, meta-analysis, and strong inference.

Tests Against the Null Hypothesis

We do not recommend as do some writers (e.g., Carver 1978; Henkel 1976) that classical statistical significance testing be discarded. Statistical significance testing is a useful "act of discipline" (Cronbach and Snow 1977) to sort out findings that may be worthy of more attention. However, marketing researchers should become more aware of the limited value of classical statistical significance tests. We offer three recommendations for improved practice in the use of classical statistical tests against the null hypothesis.

First, we support Kish's (1959) recommendation of two decades ago that the phrase "test against the null hypothesis" be substituted for the ambiguous and potentially misleading phrase "test of significance" to avoid miscommunication about the proper meaning of statistical tests. Furthermore, though results may be "statistically significant" they should not be reported as "significant" or "highly significant" which suggests that they are valid or important or provide a measure of effect size. Researchers also should avoid the misleading impression of precision or objectivity by reporting the exact statistical significance level to the fourth decimal place.

Second, because point null hypotheses are of limited value, a range rather than a point null hypothesis should be employed if possible. A range null hypothesis requires a decision in advance of data collection about the lowest effect size that will be considered to be of consequence or nontrivial. Any result within the range of effects smaller than the specified minimum would be judged as evidence that fails to reject the null hypothesis. Even if the decision is an arbitrary one, such a practice can lead to more meaningful use of tests against the null hypothesis because the range constituting the null hypothesis then becomes a respected alternative instead of a "straw man." At the very least, point null hypotheses should be replaced by a directional hypothesis; a theory that cannot generate at least a directional prediction is unworthy of the term "theory." As Meehl (1978, p. 825) forcefully argues, "It is always more valuable to show approximate agreement of observations with a theoretically predicted numerical point value, rank order, or function form, than it is to compute a 'precise probability' that something merely differs from something else."

By recommending use of directional hypotheses, we simply mean that investigators should make their expectations explicit to both themselves and others instead of following the traditional practice of stating hypotheses in the null form. However, we do not want to appear to support the practice of using one-tailed tests to prove that, for example, a t-value of 1.69 is "significant." Such emphasis on p-values gives them undue importance and diverts attention from effect size estimates. Furthermore, the tentativeness of any marketing theory ought to be recognized explicitly by more conservative two-tailed statistical tests.

Third and most important, empirical results should be described and analyzed such that the size and substantive significance of obtained effects are emphasized and not merely the p-values associated with the resulting test statistics. Presenting appropriate descriptive statistics such as means, variances, confidence intervals, contrast estimates, and estimates of total variance accounted for by a given variable before any inferential statistics can help achieve the goal of a more complete description of results. Estimates of the power of an employed statistical test to detect an effect of a chosen size can help the reader to understand more fully the nature of the obtained results and to judge the precision of the chosen inferential statistical test. Reporting statistical power is especially important when the statistical analysis does not reject the null hypothesis (Sawyer and Ball 1981).

Replication

More value should be placed on replication in marketing research. We stated before that statistical significance testing does not provide evidence about the replicability of obtained results. Science, however, depends on replication (cf. Lykken 1968; Smith 1970). If a result is replicated sufficiently, statistical significance tests are unimportant. As Stevens (1971, p. 440) stated:

> In the long run, scientists tend to believe only those results that they can reproduce. There appears to be no better option than to await the outcome of replications. It is probably fair to say that statistical tests of significance, as they are so often miscalled, have never convinced a scientist of anything.

[3]Perhaps more value would be placed on the insights from studies of individual behavior (e.g., Bettman 1974; Krugman 1971) if marketing researchers were concerned less with statistical inference tests than with the data themselves and descriptions of them.

Tversky and Kahneman's (1971) results indicate that research scientists are overly confident about the future replicability of a research result which favors the alternative hypothesis. Brown and Gaulden (1980) and Leone and Schultz (1980) have cited the dearth of replication in marketing research. Perhaps our field would not hold replication in such low regard if we were properly less naïve and smug about the interpretation and value of tests against the null hypothesis.

In early stages of research on a given set of hypotheses, replications which come as close as possible to the original study can be valuable for determining the nature and extent of effects. However, even more valuable as well as more efficient than *exact* replications are *balanced* replications. Balanced replications combine exact replications as control conditions with other conditions which manipulate additional substantive and/or methodological variables (see Carlsmith, Ellsworth, and Aronson 1976).

In several recent marketing studies researchers have used replication and statistical analysis of survey data in a manner similar to several of our recommendations. Dodson, Tybout, and Sternthal (1978) used economic utility and self-perception theories to predict and test a series of hypotheses about brand switching after purchasers either used a media-distributed coupon, bought during a cents-off deal, or redeemed a cents-off package coupon. Self-perception theory made successful (and unintuitive) predictions that repeat purchase probability would decrease, not increase, after a purchase with a media-distributed coupon. Results were replicated successfully over two product classes. Although the authors carefully conducted statistical tests of the data, they properly placed emphasis on the data and effect magnitudes.

Bagozzi (1978) similarly used theory from a variety of sources to generate several specific hypotheses about salesforce performance and satisfaction. Bagozzi carefully replicated his results across test and validation subsamples of two different samples of salespeople which differed in terms of experience and need for planning and motivation. The analysis also properly emphasized estimates of effect size such as beta coefficients and R^2. Ryans and Weinberg's (1979) analysis of determinants of territory sales response shares many of the aforementioned qualities, as do Della Bitta, Monroe, and McGinnis' (1981) replicated experiments about different ways to advertise a price reduction. Finally, Eskin and Baron's (1977) series of replicated field experiments which factorially manipulated both price and advertising expenditures is an excellent example of how replications can strengthen confidence in the external validity of a given result—especially when the result is unanticipated such as the price-advertising interaction effect they found in three of four experiments. Eskin (personal communication, 1982) has recently gathered information on about 40 experiments with retail advertising and pricing that further replicate the results of Eskin and Baron.

Bayesian Hypothesis Testing

In applied problems, when replications are not possible before a decision must be made, the use of Bayesian statistics instead of classical statistics is highly advisable. However, Bayesian statistics ought not to be confined to applied decision problems. Bayesian analysis affords several advantages in theoretical research that may not be appreciated by many marketing researchers.

Unlike classical statistical significance testing, the Bayesian approach does estimate a continuous likelihood of p(H/E) and does not necessitate a dichotomous decision that the null hypothesis is either completely false or true (Edwards, Lindman, and Savage 1963; Iverson 1970). The Bayesian approach directly compares the null and alternative hypotheses and allows one to consider more fully the possibility that the null hypothesis is true. Because the posterior distribution may be influenced by the subjective prior probabilities of an individual researcher, some researchers may reject Bayesian statistics for scientific analyses of theoretical propositions. However, as discussed before, classical statistical tests are not free from subjective decisions that can influence results. Bayesian statistics at least force the researcher to specify clearly in the prior distribution any subjectivity that enters the analysis, and allow a determination of the effects of subjective choices on the final conclusions (Iverson 1970). Furthermore, the subjective nature of prior probability estimates can be reduced by adopting a prior distribution which is essentially "flat" or insensitive in the most likely region of effect and which does not favor one extreme over another (Phillips 1973).

Greenwald (1975a,b,c) has demonstrated the greater flexibility of Bayesian hypothesis testing for making a decision between two relevant and feasible hypotheses in theoretical research, and how the Bayesian approach can provide more useful information than classical statistical significance tests when one is analyzing a series of replications. Greenwald (1975c) cited as one example the research of Layton and Turnbull (1975), who conducted two nearly identical experiments which manipulated two independent variables. They found only one small main effect in the first experiment and no statistically significant effects in the second experiment. Layton and Turnbull concluded that, given the results, they were "left with no alternative but to consider these studies *inconclusive* regarding the effects of the experimental manipulations" (p. 178).

Greenwald disputed Layton and Turnbull's conclusion and suggested that reliance on classical statistical tests was to blame for their failure to conclude something from the data of more than 400 subjects in two well-conducted experiments. In his Bayesian reanalysis, Greenwald first defined the minimum effect sizes that the experiments were able to detect. Then, for the first experiment, he formulated a flat prior probability distribution that was not subjectively biased in favor of either the null or alternative hypotheses. He next computed a likelihood

function and a posterior probability distribution for each effect from the data and tested each of the hypotheses in terms of the odds computed from the posterior distribution. The same analysis was performed on the data from the second experiment except that the posterior distributions from the first experiment were used as the priors for the second analysis. The final posterior odds *in favor of the null hypothesis* were 7.8 to 1 for one independent variable and 23.3 to 1 for the other. Greenwald thus concluded that the chances of obtaining results supportive of the alternative hypotheses for either effect were very low. Whereas the original classical statistical analysis resulted in a decision that the findings were inconclusive, Greenwald's Bayesian statistical analysis led to a more definitive conclusion that the effects of the variables in question were likely very small and that, if one wanted to test the likelihood of a null hypothesis, it was much more probable than the alternative.

Unfortunately, only a few published studies in marketing research have employed Bayesian statistics to test hypotheses. An excellent recent example of the advantage of the Bayesian over the classical approach in applied marketing research is discussed by Blattberg (1979) and Ginter et al. (1981). Banks (1965) also gives an extensive example (which was taken from Schlaifer's 1961 textbook), as does Roberts (1963). One exception to the non-utilization of Bayesian hypothesis testing in marketing is Levitt's (1972) reanalysis of his hypotheses about source credibility in industrial selling with Bayesian statistics. Levitt's Bayesian analysis helped to describe better the experimental results without the typical marketing research use of an insignificant classical statistical test as a barrier to examining the data for any valuable information (Zeisel 1955). More marketing researchers ought to use the Bayesian approach.

Meta-Analysis

Researchers' undue reliance on classical statistical tests is illustrated in many literature reviews. Traditional literature reviews often focus on counting the number of studies in a given area which do and do not find statistically significant relationships or differences. However, this approach ignores many of the issues we have raised and can result in misleading conclusions. As Meehl (1978) states, "When a reviewer tries to 'make theoretical sense' out of such a table of favorable and adverse significance test results, what the reviewer is actually engaged in, willy-nilly or unwittingly, is meaningless substantive constructions on the properties of the statistical power function, and almost nothing else" (p. 823). An alternative approach for summarizing previous empirical studies is *meta-analysis* (Glass, McGaw, and Smith 1981; Houston, Peter, and Sawyer 1983).

Meta-analysis involves a quantitative review of a research question and focuses on the obtained effect sizes in previous studies on the topic. In a meta-analysis one attempts to obtain all previous empirical studies pertaining to the research question, including if possible both published and unpublished work. The researcher using meta-analysis seeks general conclusions while searching for methodological conditions and substantive variables that might measurably moderate any observed main effects. To the extent that included studies are of varied quality, study characteristics ought to be coded as well as possible so that the size and direction of any effects of study quality can be assessed in the meta-analysis. A variety of quantitative criteria (including statistical tests) have been suggested for summarizing results. However, Glass, McGaw, and Smith (1981) and Rosenthal (1978) emphasize descriptive statistics—such as the mean effect size across a set of studies. This approach is useful not only for summarizing previous research findings but also for disentangling conflicting results and conclusions where the conflict has arisen from some studies showing statistical significance and others failing to do so.

An excellent recent example of this systematic approach to literature review is Hyde's (1981) meta-analysis of previous studies of whether males or females are superior in terms of several dimensions of cognitive ability. Authors of previous qualitative literature reviews had concluded that differences in various abilities were "well-established." However, Hyde found very small effect sizes. Hyde suggested that traditional literature reviews based simply on the number of studies yielding statistically significant results may have misleadingly communicated the impression that the moderately consistent statistically significant sex differences were large when in fact they explained only from 1 to 4% of the variance and averaged less than .5 of the population standard deviation. Hyde concluded that, "of course, a small effect might still be a important one. But at least the reader would have the option of deciding whether a statistically significant effect was large enough to merit further attention, either in teaching or in research" (p. 900).

A marketing meta-analysis that focused on effect size was Clarke's (1976) review of research assessing the duration of advertising effects on sales. Clarke's award-winning meta-analysis made an impact because his prime focus was on three substantive questions: how long do advertising effects last; do other variables interact with those effects; and, if so, how do these interacting variables affect advertising carryover? Clarke analyzed 69 studies, including some for which the effects of advertising were not statistically significant. This meta-analysis yielded several important insights not available from a more traditional qualitative literature review (e.g., Pollay 1979). First, the results indicated that the estimate of the duration of advertising effect was contingent upon the data interval. Shorter intervals (weekly, monthly, or bimonthly) indicated shorter estimates of the duration of advertising effects than longer data intervals (quarterly, annually). Perhaps most important, Clarke was able to conclude that, contrary to past beliefs, advertising effects are likely to last for no more than three to nine months and not years. Clarke summarized by stating that, although he had to make some subjective decisions in

order to produce comparable model specifications, "In isolation, none of the papers gives a satisfactory answer to the question of how long advertising affects sales. By putting them together, as has been done here, one achieves greater confidence in the result" (p. 355).

Several meta-analyses in marketing research have been reported recently. Yu and Cooper (1983) analyzed the effects of several variables on survey response rates after examining 497 response rates from 93 research studies. One conclusion was that, as would be expected intuitively, personal and telephone interviews obtained higher rates of response than mail surveys. However, Yu and Cooper's meta-analysis was able to estimate the size of that and other effects as well as support their presence. Sudman and Bradburn (1974) performed an extensive meta-analysis which investigated the influence of 46 independent variables on response effects. Other recent meta-analyses in marketing research include investigations of 37 multiattribute attitude model studies (Farley, Lehman, and Ryan 1981), four studies examining the Howard-Sheth theory of buyer behavior (Farley, Lehman, and Ryan 1982), 28 studies of price perception (Monroe and Krishnan 1983), and seven studies of the relationship of information search and prior product experience to familiarity (Reilly and Conover 1983).

A systematic meta-analysis can go beyond traditional literature reviews which focus on statistical significance and, in fact, can give a more objective and sometimes different description of results. For example, Rousseau and Redfield's (1980) meta-analysis of the effects of cognitive-level questions on achievement test scores revealed an average effect size of a half of a standard deviation, whereas a traditional analysis of the same literature indicated no effect (Winne 1979). Cooper and Rosenthal (1980) conducted an experiment in which 39 professional researchers analyzed seven studies in either a traditional qualitative manner or with a meta-analysis. The researcher subjects were asked to focus on the average effect size in terms of a *Z*-score and the average statistical probability of such an effect. Even with this relatively small number of studies to review, the qualitative reviewers formed much different and much less correct impressions about the presence and nature of the relationship between the two variables addressed in the seven studies. Finally, in addition to affording potentially greater objectivity, the use of effect size measures in meta-analysis can suggest point values or ranges that can be compared in subsequent empirical research.

Strong Inference

Although rigorous meta-analyses may increase the likelihood that point value or range predictions can be formulated such that a test of a given theory or hypothesized explanation can go beyond rejections of the null hypothesis, few areas in marketing and consumer research are amenable to such precision at the present time (see Houston, Peter, and Sawyer 1983). However, some situations may at least allow a sorting out of the best currently available theoretical explanation or model from several alternatives.

Platt (1964) advocates strong inference as a useful procedure to augment conventional tests against the null hypothesis. This approach involves comparing competing hypotheses with each other where support for one hypothesis (theory) implies rejection of others. The process of strong inference includes the following steps: (1) devising alternative hypotheses, (2) devising a crucial experiment (or several of them) with alternative possible outcomes each of which will, as nearly as possible, exclude one or more of the hypotheses or explanations, (3) carrying out the experiment so as to get a clean result, (4) recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain, and so on. Though the approach sounds simple, much ingenuity clearly is needed to implement this research strategy. However, several examples of the approach are reported in the marketing and consumer behavior literature.

An excellent example of strong inference in hypothesis testing is the investigation of the low-ball technique by Cialdini et al. (1978). The authors observed that automobile sales dealers induce final compliance by getting customers to decide initially to purchase at a lower price and then to retain that compliance when the price advantage is removed. Cialdini et al. used a strong inference design and the results supported an explanation that initial commitment creates a resistance to change in future behavior but not necessarily a more positive attitude. At least as important in terms of strong inference is the fact that the results also rejected the plausibility of the other three explanations of the obtained low-ball effect. Burger and Petty (1981) further refined the conclusions of Cialdini et al. with a strong inference experiment which supported an explanation that an unfulfilled obligation to the person requesting the behavior, not a commitment to the initial target behavior, is responsible for the effectiveness of the low-ball technique.

Another strong inference design directly confronted the Fishbein belief-evaluation multiattribute attitude model with the adequacy-importance approach (Bettman, Capon, and Lutz 1975). This study examined how role-playing subjects formed attitudes toward fictitious brands from given attribute information. The authors used within-individual analyses of variance and ω^2 estimates of explained variance to classify individuals on the basis of how attribute information was utilized. Their research revealed that the multiplicative model was by far the best description of the individuals' information processing and that the Fishbein model was superior to the adequacy-importance model.

Even if use of a strong inference design to test alternative theories is not feasible, one may at least be able to compare a sample result with the value predicted by a given theory or model instead of simply testing whether the result is statistically significantly different from zero. In addition, the predictions of competing or alternative

models can be compared with each other (Armstrong 1979). One such system of statistical analysis is Jöreskog and Sörbom's (1978) maximum likelihood estimation of structural equations to test causal models involving unobservable variables (Bagozzi 1980; Bentler 1980). This approach requires explicit specification of the complex interrelationships among measured and unobservable variables and thus strong theory is needed. Sawyer and Page (1983) summarize how various measures of effect size can augment statistical tests of the fit of sample data to theoretical models.

LIMITATIONS OF OUR RECOMMENDATIONS

We have argued for practices and priorities which differ from current conduct and reporting of empirical research in marketing. Statistics should be used to illuminate rather than obscure data, and we hope that our recommendations can help to achieve this goal. However, we also recognize that there are limitations and problems with any type of hypothesis testing and our recommendations are no exception. In this section we briefly review some of these problems.

We have argued for increased use and reporting of *descriptive statistics* in marketing research. Though such reporting conflicts with the limited space in journals, space constraints should not prevent the inclusion of sufficient information for replication and/or inclusion of the study in a subsequent meta-analysis. If journal space constraints preclude the complete description of a study's results, perhaps the journal could require and store pertinent method information, data, and statistics to aid inquiring researchers. We acknowledge, however, that even simple descriptive statistics can sometimes be misleading. For example, averaging many individuals who exhibit "all-at-once" learning patterns, albeit at a varying number of trials, would result in the incorrect conclusion that individuals learn at a gradual rate (Baloff and Becker 1967).

There are several difficulties in the quantification, interpretation, and generalization of effect size measures. Some such measures estimate the ratio of explained to total variance. In quantifying the amount of explained variance (such as R^2 or ω^2), researchers must realize that total variance is increased by measurement and treatment unreliability, heterogeneous subjects, and poorly controlled research procedures (O'Grady 1982; Sechrest and Yeaton 1981a,b). Experimental researchers also can influence the amount of explained variance by restricting or magnifying the manipulation of an independent variable. Independent variables which are qualitative or categorical present particular interpretation problems. Such variables often have no conceptually meaningful or practically important characteristics in common within or across studies; the number of "levels" of such variables is infinite and any estimates of the "size" of their effects are very difficult to interpret. Finally, the problems of the influence of individual characteristics of particular studies and manipulations within a study make it very difficult to generalize effect sizes meaningfully or to compare them across a set of different studies (such as in a meta-analysis). However effect sizes are estimated, these descriptive statistics are more generalizable if the levels of the independent variable are a random subset of all levels of interest (Glass and Hakstian 1969) and orthogonal to other independent variables (Green, Carroll, and DeSarbo 1978; LaTour 1981a).

Fortunately, other measures of effect size are available. Rosenthal (1978) discusses the advantages and disadvantages of nine relatively simple methods of summarizing results including three estimates of effect size. These methods include adding *t*-test statistics, *Z*-values, and weighted *Z*-values. LaTour (1981a,b) recommends the use of a contrast estimate to quantify effect size because it eliminates many of the problems of explained variance estimates. Glass, McGaw, and Smith (1981, p. 102) recently concluded that, "The findings of comparative experiments are probably best expressed as standardized mean differences between pairs of treatment differences." Most of these methods that do not estimate the proportion of explained variance seem most appropriate for simple research designs and are difficult to use and interpret with more complex designs (Glass and Hakstian 1969).

Some limitations of our other recommendations should also be noted. Though we believe that *replication research* is very important, recognition for conducting replications seems to be lacking in marketing research. Also, it is very unlikely that all sources of variance in research involving human subjects can be specified or controlled. Thus, replications can never exactly duplicate prior research conditions, and different results may be obtained. Such conflicting results can lead to confusion rather than consensus. Of course, confusion is better than the acceptance of a single result as conclusive, and subsequent meta-analyses may be able to determine the source of the conflict in results.

We believe *Bayesian hypothesis testing* is useful, but also recognize that researchers need to have a high degree of mathematical sophistication to understand and apply the approach. It is clearly not an approach which is amenable to canned computer programs and is thus difficult for researchers to use.

In addition to the problem of meaningfully comparing effect sizes, a *meta-analysis* often encounters other formidable obstacles. One problem is the search for a census of studies including the unpublished ones that are likely to have smaller effect sizes. For studies that are available, information is often insufficient for calculating effect sizes and study authors must be contacted. Unfortunately, it is also often difficult to obtain sufficiently detailed descriptions of the study method and to code these study characteristics so that their effects can be assessed in the meta-analysis. Small samples and confounded study characteristics make it difficult to disentangle main effects across studies, as well as complex

interactions. An opposite problem is that, if all surveyed studies use the same procedure, the effect of that method cannot be assessed (e.g., Cartwright 1973). One important outcome of a meta-analysis might be a specification of types of studies that would fill a void and allow an examination of the effects of variables that cannot currently be meaningfully evaluated.

It should be obvious that a meta-analysis, though quantitative, depends on many subjective researcher decisions and affords much opportunity for disagreement. Perhaps because the publication of a meta-analysis carries an aura of finality, researchers very commonly disagree about the many decisions involved in a meta-analysis and, hence, challenge the conclusions. For example, Stanley and Benbow (1982) challenged Hyde's (1981) meta-analysis of gender differences in quantitative ability and Weinberg and Weiss (1982) disputed some of the analysis decisions in Clarke's (1976) meta-analysis of advertising carryover as well as the statistical validity of his conclusions.[4]

Finally, though *strong inference designs* are superior to test against the null hypothesis, often theories are incommensurable and hence cannot be confronted empirically. In addition, even strong inference designs can obtain conflicting results. For example, Mazis, Ahtola, and Klippel (1975) compared four formulations of multiattribute attitude models and concluded that the adequacy-importance model yielded better predictions than the Fishbein model. This conclusion conflicts with the findings of Bettman, Capon, and Lutz (1975).

Though the preceding discussion is by no means a complete list of limitations, the problems noted should serve as a reminder of one critical fact: *there is no universal approach to hypothesis testing which can guarantee a meaningful empirical test or offer fully objective analysis and description of results*. Some approaches are better than others for particular problems. As we have illustrated, biases in choosing an approach and the decisions made in implementing it have an extremely important influence on conclusions from the data. Thus, if possible, researchers ought to use multiple approaches to testing hypotheses and reporting the results.

SUMMARY

Several issues related to the interpretation and value of statistical significance testing are reviewed. Although properly applied statistical significance tests are useful aids in drawing inferences and for signalling relationships which need further study, they are not sufficient for falsifying hypotheses or judging research results. Despite the fact that many of these ideas have been discussed previously, many researchers, including those in marketing, continue to ignore them. Attention should be placed on the data themselves and their descriptions. Instead of relying solely on classical inferential statistics, researchers should make added use of replication, Bayesian statistics, meta-analysis, and strong inference to provide more meaningful examination of theoretical questions in marketing research.

[4]Though the statistical models involved in the exchange between Weinberg and Weiss and Clarke (1982) are very sophisticated, the arguments pertain to important basic ideas discussed in this article about statistical power, whether failure to reject the null hypothesis implies that the null hypothesis is true, and the need for testing results against theoretically based point value predictions instead of merely comparing results against a zero point null hypothesis.

REFERENCES

Armstrong, J. Scott (1979), "Advocacy and Objectivity in Science," *Management Science*, 25 (May), 423–38.

Assael, Henry and John Keon (1982), "Nonsampling vs. Sampling Errors in Survey Research," *Journal of Marketing*, 46 (Spring), 114–23.

Bagozzi, Richard P. (1978), "Salesforce Performance and Satisfaction as a Function of Individual Difference, Interpersonal, and Situational Factors," *Journal of Marketing Research*, 15 (November), 517–31.

——— (1980), *Causal Models in Marketing*. New York: John Wiley & Sons, Inc.

Bakan, David (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66 (December), 423–37.

Baloff, Nicholas and Selwyn Becker (1967), "On the Futility of Aggregating Individual Learning Curves," *Psychological Reports*, 20, 183–91.

Banks, Seymour (1965), *Experimentation in Marketing*. New York: McGraw-Hill Book Company.

Bentler, P. M. (1980), "Multivariate Analysis with Latent Variables: Causal Modeling," in *Annual Review of Psychology*, Vol. 31, M. R. Rosenzweig and L. W. Porter, eds. Palo Alto, CA: Annual Reviews.

Bettman, James R. (1974), "Toward a Statistics for Consumer Decision Net Models," *Journal of Consumer Research*, 1 (June), 71–80.

———, Noel Capon, and Richard J. Lutz (1975), "Cognitive Algebra in Multi-Attribute Attitude Models," *Journal of Marketing Research*, 12 (May), 151–64.

Blattberg, Robert C. (1979), "The Design of Advertising Experiments Using Statistical Decision Theory," *Journal of Marketing Research*, 16 (May), 191–202.

Brown, Rex V. (1967), "Evaluation of Total Survey Error," *Journal of Marketing Research*, 4 (May), 117–27.

Brown, Stephen W. and Corbett F. Gaulden, Jr. (1980), "Replication and Theory Development," in *Theoretical Developments in Marketing*, C. W. Lamb, Jr. and P. M. Dunne, eds. Chicago: American Marketing Association, 240–3.

Burger, Jerry M. and Richard E. Petty (1981), "The Low-Ball Compliance Technique: Task or Person Commitment?," *Journal of Personality and Social Psychology*, 40 (March), 492–500.

Carlsmith, J. Merrill, Phoebe C. Ellsworth, and Elliot Aronson (1976), *Methods of Research in Social Psychology*. Reading, MA: Addison-Wesley.

Cartwright, Dorwin (1973), "Determinants of Scientific Progress: The Case of Research on the Risky Shift," *American Psychologist*, 28 (March), 222–31.

Carver, Ronald P. (1978), "The Case Against Statistical Significance Testing," *Harvard Educational Review*, 48 (August), 278–399.

Cialdini, Robert B., John T. Cacioppo, Rodney Bassett, and John A. Miller (1978), "Low-Ball Procedure for Producing Compliance Commitment then Cost," *Journal of Personality and Social Psychology*, 36 (May), 463–76.

Clarke, Darral G. (1976), "Econometric Measurement of the Duration of Advertising Effect on Sales," *Journal of Marketing Research*, 13 (November), 345–57.

——— (1982), "A Reply to Weinberg and Weiss," *Journal of Marketing Research*, 19 (November), 592–4.

Cohen, Jacob (1977), *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cook, Thomas D. and Donald T. Campbell (1979), *Quasi-Experimentation: Design and Analysis for Field Settings*. Chicago: Rand-McNally.

Cooper, Harris M. and Robert Rosenthal (1980), "Statistical Versus Traditional Procedures for Summarizing Research Findings," *Psychological Bulletin*, 87 (May), 442–9.

Cronbach, Lee J. and R. E. Snow (1977), *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington.

Della Bitta, Albert J., Kent B. Monroe, and John M. McGinnis (1981), "Consumer Perceptions of Comparative Price Advertisements," *Journal of Marketing Research*, 18 (November), 416–27.

Dodson, Joe A., Alice M. Tybout, and Brian Sternthal (1978), "Impact of Deals and Deal Retraction on Brand Switching," *Journal of Marketing Research*, 15 (February), 72–81.

Edwards, Ward, Harold Lindman, and Leonard J. Savage (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70 (May), 193–242.

Eskin, Gerald J. and Penny H. Baron (1977), "Effects of Price and Advertising in Test-Market Experiments," *Journal of Marketing Research*, 14 (November), 499–508.

Farley, John U., Donald R. Lehman, and Michael J. Ryan (1981), "Generalizing from 'Imperfect' Replication," *Journal of Business*, 54 (October), 597–610.

———, ———, and ——— (1982), "Patterns in Parameters of Buyer Behavior Models: Generalizing from Sparse Replication," *Marketing Science*, 1 (Spring), 181–204.

Ginter, James, Martha Cooper, Carl Obermiller, and Thomas Page (1981), "The Design of Advertising Experiments: An Extension," *Journal of Marketing Research*, 18 (February), 120–3.

Glass, Gene V. and A. Ralph Hakstian (1969), "Measures of Association in Comparative Experiments: Their Development and Interpretation," *American Educational Research Journal*, 6 (May), 403–14.

———, Barry McGaw, and Mary Lee Smith (1981), *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications.

Green, Paul E., J. Douglas Carroll, and Wayne S. DeSarbo (1978), "A New Measure of Predictor Variable Importance in Multiple Regression," *Journal of Marketing Research*, 15 (August), 356–60.

Greenwald, Anthony G. (1975a), "Consequences of Prejudice Against the Null Hypothesis," *Psychological Bulletin*, 82 (January), 1–19.

——— (1975b), "Does the Good Samaritan Parable Increase Helping? A Comment on Darley and Batson's No-Effect Conclusion," *Journal of Personality and Social Psychology*, 32 (October), 578–83.

——— (1975c), "Significance, Nonsignificance, and Interpretation of an ESP Experiment," *Journal of Experimental Social Psychology*, 11 (March), 180–91.

Henkel, Ramon E. (1976), *Tests of Significance*. Beverly Hills, CA: Sage Publications.

Houston, Michael J., J. Paul Peter, and Alan G. Sawyer (1983), "The Role of Meta-Analysis in Consumer Behavior Research," in *Advances in Consumer Research*, Vol. 10, R. P. Bagozzi and A. M. Tybout, eds. Ann Arbor, MI: Association for Consumer Research.

Hyde, Janet Shibley (1981), "How Large Are Cognitive Gender Differences?: A Meta-Analysis Using ω^2 and d," *American Psychologist*, 36 (August), 892–901.

Iverson, Gudmund R. (1970), "Statistics According to Bayes," in *Sociological Methodology*, Edgar F. Borgatta and George W. Bohrnstedt, eds. San Francisco: Jossey-Bass, 185–99.

Jöreskog, Karl G. and Dag Sörbom (1978), *LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*, Version IV, Release 2. Chicago: National Educational Resources, Inc.

Kish, Leslie (1959), "Some Statistical Problems in Research Design," *American Sociological Review*, 24 (June), 328–38.

Krugman, Herbert E. (1971), "Brain Wave Measures of Media Involvement," *Journal of Advertising Research*, 11 (February), 3–9.

LaTour, Stephen A. (1981a), "Effect Size Estimation: A Commentary on Wolf and Bassler," *Decision Sciences* (January), 136–41.

——— (1981b), "Variance Explained: It Measures Neither Importance nor Effect Size," *Decision Sciences* (January), 150–60.

Layton, R. D. and B. Turnbull (1975), "Belief, Evaluation, and Performance on an ESP Task," *Journal of Experimental Social Psychology*, 11 (March), 166–79.

Leone, Robert P. and Randall L. Schultz (1980), "A Study of Marketing Generalizations," *Journal of Marketing*, 44 (Winter), 10–18.

Levitt, Theodore (1972), "Industrial Purchasing Behavior: A Bayesian Reanalysis," *Journal of Business Administration*, 4 (Fall), 79–81.

Lipstein, Benjamin (1975), "In Defense of Small Samples," *Journal of Advertising Research*, 15 (February), 33–40.

Lykken, David T. (1968), "Statistical Significance in Psychological Research," *Psychological Bulletin*, 70 (September), 151–9.

Mayer, Charles (1970), "Assessing the Accuracy of Marketing Research," *Journal of Marketing Research*, 7 (August), 285–91.

Mazis, Michael, Olli T. Ahtola, and R. Eugene Klippel (1975), "A Comparison of Four Multi-Attribute Models in the Prediction of Consumer Attitudes," *Journal of Consumer Research*, 2 (June), 38–52.

McGuire, William J. (1973), "The Yin and Yang of Progress in Social Psychology: Seven Koan," *Journal of Personality and Social Psychology*, 26 (June), 446–56.

Meehl, Paul E. (1967), "Theory Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, 16 (June), 103–15.

——— (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," *Journal of Consulting and Clinical Psychology*, 46, 806–84.

Meyer, Donald L. (1974), "Statistical Tests and Surveys of Power: A Critique," *American Educational Research Journal*, 11 (Spring), 179–88.

Monroe, Kent B. and R. Krishnan (1983), "The Effect of Price

on Subjective Product Evaluations: A Synthesis of Outcomes," in *Advances in Consumer Research*, Vol. 10, R. P. Bagozzi and A. M. Tybout, eds. Ann Arbor, MI: Association for Consumer Research.

O'Grady, Kevin E. (1982), "Measures of Explained Variance: Cautions and Limitations," *Psychological Bulletin*, 92 (November), 766–77.

Peter, J. Paul (1983), "Some Philosophical and Methodological Issues in Consumer Research," in *Marketing Theory: The Philosophy of Marketing Science*, Shelby D. Hunt, ed. Homewood, IL: Richard D. Irwin.

Phillips, Lawerence D. (1973), *Bayesian Statistics for Social Sciences*. London: Thomas Nelson.

Platt, John R. (1964), "Strong Inference," *Science*, 146 (October 16), 347–53.

Pollay, Richard W. (1979), "Lydiametrics: Applications of Econometrics to the History of Advertising," *Journal of Advertising History*, 1, 3–18.

Reilly, Michael D. and Jerry N. Conover (1983), "Meta-Analysis: Integrating Results from Consumer Research Studies," in *Advances in Consumer Research*, Vol. 10, R. P. Bagozzi and A. M. Tybout, eds. Ann Arbor, MI: Association for Consumer Research.

Roberts, Harry V. (1963), "Bayesian Statistics in Marketing," *Journal of Marketing*, 27 (January), 1–4.

Rosenthal, Robert (1969), "Interpersonal Expectations: Effects of the Experimenter's Hypothesis," in *Artifact in Behavioral Research*, Robert Rosenthal and Ralph L. Rosnow, eds. New York: Academic Press, 181–277.

——— (1978), "Combining Results of Independent Studies," *Psychological Bulletin*, 85 (December), 185–93.

——— (1979), "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin*, 86 (March), 638–41.

——— and John Gaito (1963), "The Interpretation of Levels of Significance by Psychological Researchers," *Journal of Psychology*, 55, 33–8.

——— and Donald B. Rubin (1982), "Comparing Effect Sizes of Independent Studies," *Psychological Bulletin*, 92 (September), 500–4.

Rousseau, E. W. and D. L. Redfield (1980), "Teacher Questioning," *Evaluation in Education, An International Review Series*, 4, 51–2.

Ryans, Adrian B. and Charles B. Weinberg (1979), "Territory Sales Reponse," *Journal of Marketing Research*, 16 (November), 453–65.

Sawyer, Alan G. and A. Dwayne Ball (1981), "Statistical Power and Effect Size in Marketing Research," *Journal of Marketing Research*, 18 (August), 275–90.

——— and Thomas J. Page, Jr. (1983), "Incremental Goodness of Fit Indices in Structural Equation Models in Marketing Research," paper presented at the AMA special Conference on Causal Modeling, Sarasota, FL, March 2.

Schlaifer, Robert (1961), *Introduction to Statistics for Business Decisions*. New York: McGraw-Hill Book Company.

Sechrest, Lee and William Yeaton (1981a), "Empirical Bases for Estimating Effect Size," in *Reanalyzing Program Evaluations: Policies and Practices*, R. F. Boruch, P. M. Wortman, and D. S. Cordray, eds. Ann Arbor: University of Michigan Institute for Social Research.

——— and ——— (1981b), "Estimating Magnitudes of Experimental Effects," unpublished manuscript, University of Michigan Institute for Social Research, Ann Arbor.

Selvin, Hanan C. (1957), "A Critique of Tests of Significance in Survey Research," *American Sociological Review*, 22 (October), 519–27.

Smith, N. C., Jr. (1970), "Replication Studies: A Neglected Aspect of Psychological Research," *American Psychologist*, 25 (October), 970–5.

Stanley, Julian C. and Camilla P. Benbow (1982), "Huge Sex Ratios at Upper End," *American Psychologist*, 37 (August), 972.

Stevens, S. S. (1971), "Issues in Psychophysical Measurement," *Psychological Review*, 78 (September), 426–50.

Sudman, Seymour and Norman M. Bradburn (1974), *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

Tversky, Amos and Daniel Kahneman (1971), "Belief in the Law of Small Numbers," *Psychological Bulletin*, 76 (August), 105–10.

Weinberg, Charles B. and Doyle L. Weiss (1982), "On the Econometric Measurement of the Duration of Advertising Effects on Sales," *Journal of Marketing Research*, 19 (November), 585–91.

Winch, Robert F. and Donald T. Campbell (1969), "Proof? No. Evidence? Yes. The Significance of Tests of Significance," *American Sociologist*, 4 (May), 140–3.

Winne, P. H. (1979), "Experiments Relating Teacher's Use of Higher Cognitive Questions to Student Achievement," *Review of Educational Research*, 49, 13–50.

Yu, Julie and Harris Cooper (1983), "A Quantitative Review of Research Design Effects in Response Rates to Questionnaires," *Journal of Marketing Research*, 20 (February), 36–44.

Zeisel, Hans (1955), "The Significance of Insignificant Differences," *Public Opinion Quarterly*, 17 (Fall), 319–21.

32

Explaining Delinquent Involvement

A Consideration of Suppressor Effects

Wendy L. Lipton and M. Dwayne Smith

A common goal of delinquency theories has been the search for underlying etiological factors which motivate or impel youths to engage in delinquent behavior. Much effort has been expended in the development of these theories, resulting in a discipline rich in theoretical perspectives. Indeed, the study of delinquency can never be accused of being an atheoretical pursuit; in fact, choosing from among a wide variety of theoretical orientations constitutes a major decision for most researchers.

In recent years, a movement has occurred away from the development of new delinquency theories toward the testing of existing theories. In many respects, this trend is long overdue; while theoretical formulations have been numerous, carefully devised empirical tests of specific theories have been all too rare (Gibbons, 1976).

The purpose of this empirical scrutiny has been directed toward not only establishing a preference for one theory over another, but also for providing a reliable set of variables that may be used to predict delinquent behavior. Unfortunately, the results of most empirical studies of delinquent behavior have been less than satisfactory. Regardless of the particular theory, or even synthesis of theories, being tested, few studies have accounted for more than 30 percent of the variance in delinquent activities among various samples of youths (cf. Jensen, 1972; Elliott and Voss, 1974; Conger, 1976; Hepburn, 1976; Minor, 1977; Cernkovich, 1978; Linden, 1978; Aultman and Wellford, 1979;

From Wendy L. Lipton and M. Dwayne Smith, "Explaining Delinquent Involvement: A Consideration of Suppressor Effects," *Journal of Research in Crime and Delinquency*, 1983, 20(2), 199–213.

Johnson, 1979; Shover et al., 1979; Wiatrowski et al., 1981).[1]

It may be that these results should not be that disappointing. Schuessler (1971) has warned that the validity of a model cannot necessarily be ascertained from its predictive efficiency (i.e., its R^2). Further, several researchers have posited that the search for deep-seated etiologies of delinquency may be in vain because, for the vast majority of youths, delinquency is much more of a spontaneous, situational act than a long-term pattern of behavior (e.g., Matza, 1964; Briar and Piliavin, 1965; Gold, 1970).

Nevertheless, most delinquency researchers assume that there *are* underlying factors that somehow determine one's propensity toward delinquent behavior, and that empirical models which account for these factors should provide statistical results indicative of more than moderate (at best) predictive power. The inability to demonstrate this adequately has apparently taken its toll. Lotz (1979), for instance, has provided evidence which seems to indicate the development of a new pessimism among delinquency researchers. A component of this pessimism specifies that it may be impossible to understand delinquency or, consequently, to provide ameliorative programs to deal with this phenomenon.

Before abandoning a central pursuit of the discipline, however, it may be worthwhile to consider (and reconsider) some of the difficulties which plague delinquency research. While many of these difficulties are common to all social research efforts, there are some with particular ramifications for the study of delinquency (Hirschi and Selvin, 1967). The purpose of this report is to discuss one area which has heretofore been neglected—the statistical issue of suppression—and to demonstrate how this issue has both statistical *and* theoretical implications for the development of models of delinquent behavior.

THE ISSUE OF SUPPRESSION

Whether a model of delinquency is an operationalization of a specific theory or a compilation of elements from several perspectives, the central focus of the research is on the dependent variable in the model. Since the goal is to predict delinquent involvement, variables which do not display statistically significant and/or substantive correlations with the delinquency measure tend to be excluded from the models.

Herein lies a possible problem—we have often failed to recognize or test for the possibility of suppressor effects. Our inability to detect relationships between certain factors and delinquency may be due to the fact that we are leaving out key variables which serve as suppressors in the relationship between delinquency and the variable(s) in question. A consideration of the etiologies of our predictor variables could lead to a discovery of additional delinquency correlates (and ultimately higher R^2s), particularly if our predictors are serving as suppressor variables. This report offers an empirical example of how we might approach the issue of suppression in delinquency

research by considering the relationships between social class, educational experiences, and delinquent involvement.

A Note on Suppression

The phenomenon of suppression is usually ignored in statistical discussions and remains an ambiguous concept to many social scientists. Since this statistical oddity serves as the focus of this study, a few explanatory notes may be in order.

The concept of suppression has been classically described as a situation in which a given independent variable serves to weaken or conceal a relationship between two other variables (Rosenberg, 1968:65). The statistical explanation of suppression is perhaps most broadly presented by Cohen and Cohen (1975:87-91). Their discussion describes three types of suppression: classical, net, and cooperative.

Classical suppression: Where the correlation between the dependent variable (Y) and one independent variable (X_2) is equal to zero, its correlation with another independent variable (X_1) is greater than zero, and the correlation between X_1 and X_2 does not equal zero. "In spite of its zero correlation with Y, X_2 increases the variance accounted for in Y by 'suppressing' some of the variance in X_1 that is irrelevant to Y" (p. 87).

Net suppression: In this situation, all correlations in the equation are positive. Although the correlation between Y and X_2 is positive, "the function of X_2 in the multiple correlation and regression is primarily in suppressing a portion of the variance in X_1 that is irrelevant to (uncorrelated with) Y" (p. 89).

Cooperative suppression: Where "the independent variables are mutually enhancing . . . and each variable accounts for a *larger* proportion of the Y variance in the presence of the other than it does alone" (p. 91). This situation occurs when the independent variables correlate negatively with each other and positively with Y (or vice-versa).

A suppressor effect is detectable through a comparison of zero-order correlations and multivariate (standardized) regression coefficients. Typically, a comparison of these coefficients would reveal the value of the regression coefficient to be between the range of zero and the value of the zero-order correlation; however, if a suppressor effect exists, the result of controlling for additional predictors will be to reverse the sign of the coefficient and/or to *increase* the value of the coefficient such that it goes outside the expected range defined in the bivariate situation. Such a finding would be in contrast to what would normally be expected in hierarchical regression analyses, where the inclusion of additional predictors would be expected to *decrease* the effect of the exogenous variable(s) (see Alwin and Hauser, 1975).

In sum, when variations in suppressor variables are held constant, we are able to obtain a truer picture of the effects of an independent variable on the specified dependent variable.

AN EXAMPLE OF SUPPRESSION

One of the more controversial issues in the delinquency literature concerns the relationship between social class and delinquent behavior. Regardless of the failure of many contemporary studies to detect a relationship between class and delinquency, criminologists have been hesitant to remove such a powerful variable from their models. This is understandable in light of the fact that the relationship between class origin and life-cycle experiences has been well documented (e.g., Kerckhoff, 1972; Elder, 1974).

The weight of the current evidence would seem to suggest that such a relationship does exist (Hindelang et al., 1979; Braithwaite, 1981), although it is not nearly as simple nor direct as had been assumed. Still, a large number of studies, particularly those utilizing self-report measures, have failed to detect a relationship between social class and delinquent behavior (Tittle et al., 1978). There are several possible explanations for this failure, not the least of which is the difficulty in measuring social class in a society where class distinctions are becoming increasingly blurred. However, the inability of previous studies to uncover an association between social class and delinquency may have been due to their failure to consider variables which could serve as suppressors in this relationship. It is possible that social origins *do* affect an adolescent's propensity to commit delinquent acts, and that the inability to justify this claim empirically derives from a failure to control for factors which operate to suppress the class effects.

The Educational Arena

The task of searching for suppressors becomes one of determining factors which are related to both the dependent and independent variables. Since educational experiences are among the most consistently documented correlates of both delinquency and social class, it would seem plausible to consider school-related variables as possible suppressors in this relationship.

The school and delinquency. A review of the delinquency literature reveals a long history of theroretical and empirical inquiries regarding the relationship between school experiences and delinquent behavior (see Phillips and Kelley [1979] for a review of this literature). One method of operationalizing educational experiences stems from Hirschi's (1969) conceptualization of what he terms "attachment to school." According to Hirschi, academic performance and mental ability are both related to delinquent involvement, but a failure in academic pursuits does not necessarily lead to the commission of delinquent acts. He argues that a better indicator of propensity to delinquency is the degree of an adolescent's attachment to school; i.e., the extent to which the adolescent is affectionally tied to his teachers and considers education important.

A number of research efforts have supported Hirschi's contention regarding attachment to school. Low delinquency has been reported as being

associated with such factors as positive attitudes toward school (Frease, 1972), perceived importance of education (Krohn and Massey, 1980), and number of hours spent on homework (Polk and Halferty, 1966). Conversely, high rates of delinquency have been shown to covary with a lack of commitment to school and adult values (Polk and Halferty, 1966) and to alienation from the educational sphere (Elliott and Voss, 1974). Thus it would seem that the greater the adolescent's interpersonal and behavioral attachment to the educational sphere, the less likely he or she will be to engage in delinquent behavior.

Educational experiences and social class. The literature regarding the relationship between social class and educational experiences clearly suggests that the lower the level of the adolescent's class origins, the less likely he or she is to succeed (or even *want* to succeed) in school. There is an extensive literature regarding the relationship between social class and educational aspirations and attainments, and a number of researchers have noted a strong association between family status and preparation for, attitudes toward, and success within the educational arena (Becker, 1952; Clausen, 1968; Sewell et al., 1969; Alexander et al., 1978).

Insofar as children internalize the values and motives of their parents early in the socialization process, and since the family is largely responsible for preparing the child for entry into the educational sphere, then a relationship could be expected between the degree to which adolescents are attached to school and the social class of their parents.

The dilemma. If educational experiences are related to both class and delinquency, then it would seem logical to assume that attachment to school would serve as some sort of link between the two. The question, however, is "what kind of link?" As Tittle et al. (1979) note in their work,

> . . . there does seem to be an empirical relationship between class origin and academic performance in high school. There also seems to be a consistent and strong association between academic performance and delinquency. . . . Therefore, it should follow that there would be a strong class origin/deliquency association, but of course, our paper shows that in general such a relationship has not been demonstrated. Either the origin/performance or the performance/delinquency association is in error or some rather complex interactions are involved which need to be sorted out empirically. (670)

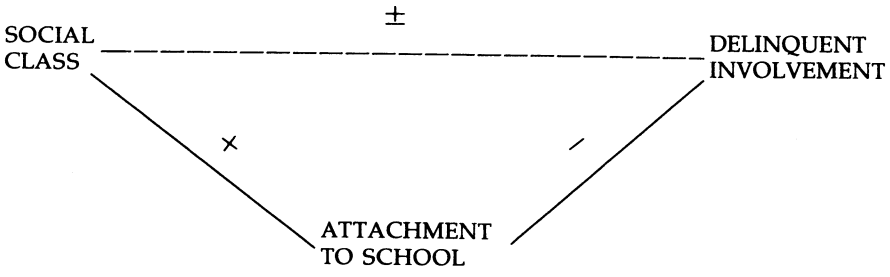
The following discussion presents such an approach for sorting out these (and other) relationships by considering the possibility that educational experiences serve to suppress the relationship between class and delinquency.

Statement of the Problem

If we were to accept the existing evidence that there is no relationship between social class and delinquency, then the purpose of specifying attachment to school as a link between the two would be to explicate a simple flow of causation from social class to attachment to delinquency.[2] The resulting

Figure 1. Representation of the Causal Relationships Between Social Class, Attachment to School, and Delinquent Involvement, as Suggested by the Literature

Figure 2. Theoretical Possibility for the Function of Attachment to School as a Suppressor in a Model of Delinquent Involvement*



*Broken path indicates the effect is due to the inclusion of attachment to school in the model.

model would show that low delinquency is a function of the negative influence of attachment to school and that attachment to school is caused by the positive effect of social class. Since we would assume that class and delinquency are unrelated, however, the model which specifies attachment to school as an independent, endogenous variable would not specify a direct path from social class to delinquent involvement (see Figure 1).

In testing for suppressor effects, on the other hand, we begin with the assumption that there *is* some association between social class and delinquency, but that it is being obscured (i.e., "suppressed") by some other factor(s). Thus, unlike the model which denotes attachment to school as simply intervening between class and delinquency, our model would specify a direct causal path from social class to delinquent involvement, as shown in Figure 2.

Attachment as a suppressor. To conceptualize attachment to school as a suppressor in the relationship between social class and delinquent behavior is to assume that class differences in levels of delinquency are revealed when variations in degrees of attachment to school are held constant, i.e., the zero-order relationship is equal to zero but the first-order partial is positive. Across levels of attachment, we would expect lower-class youths to be significantly more likely to engage in delinquent acts than their higher class counterparts (or vice-versa). Controlling for attachment to school would serve to partial out those youths who do not fit into the pattern of expectations (e.g., lower-class

youths with higher levels of attachment and low rates of delinquency), permitting a class effect to emerge.

In summary, a theoretical argument could be made that the relationship of attachment to school to both social class and delinquency may be obscuring the association between class origins and delinquent behavior. Empirical support for this argument would be obtained if a bivariate correlation between class and delinquency was significantly improved upon when controlling for attachment to school.

TESTING FOR SUPPRESSION

Data

Data for this example are taken from the Youth in Transition (YIT) project, a five-wave panel study which follows youths from age 16 through young adulthood (see Bachman et al., 1978). This example utilizes data from the first two waves of the study, incorporating a national sample of 1,592 white male adolescents.[3] Information regarding the respondents' socio-demographic and behavioral and attitudinal characteristics with respect to education is available in the initial wave (Fall, 1966) of the YIT data. In order better to approximate the causal mechanism, the delinquency measure (which provides retrospective self-reports of delinquent behavior) is taken from the second wave (Spring, 1968) of the study.

Operationalization of the Variables

The dependent variable for this study, self-reported delinquency (DELINQUENCY), is a composite of eight items measuring involvement in acts of theft and vandalism. Social class (SEL) is operationalized through a composite index of six components measuring socioeconomic level. Attachment to school is conceptualized as an unmeasured construct and is operationalized by utilizing individual items and indexes which have been suggested as indicators of attachment to school in previous research efforts (see *The school and delinquency* above). These are: academic performance (GPA), perceived teachers' interest in the respondent (TEACHERS), the respondents' perception of the value and importance of education (IMPORTED), the respondents' commitment to finishing school (COMMITTED) and actual time spent doing homework (HOMEWORK). Further specifications of the variables utilized are provided in the Appendix and in Bachman et al. (1978).

Table 1. Bivariate Correlation Matrix

Variables	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1. SEL	—						
2. GPA	.248	—					
3. TEACHERS	.106	.134	—				
4. IMPORTED	.147	.278	.175	—			
5. COMMITTED	.159	.126	.107	.230	—		
6. HOMEWORK	.139	.216	.153	.174	.105	—	
7. DELINQUENCY	.016*	−.100	−.061	−.141	−.075	−.161	—

*Nonsignificant at $p<.05$

Results

Bivariate relationships. Correlations among the elements are presented in Table 1. In keeping with the literature, these results indicate that the attachment to school variables (i.e., GPA, TEACHERS, IMPORTED, COMMITTED, and HOMEWORK) are negatively related to DELINQUENCY. Further, and of particular importance, the correlation between SEL and DELINQUENCY (.016) is nonsignificant and effectively zero.

Taken at face value, the zero-order correlation indicates that there is no relationship between social class and delinquent involvement. The logical (and rather typical) reaction to this finding would be to exclude social class from subsequent analyses. As the following discussion will suggest, however, such a decision would be premature.

Multivariate expectations. If attachment to school serves as a suppressor in the relationship between class and delinquency, then controlling for the attachment variables should reveal a stronger class effect in the multivariate equation than had been apparent from the bivariate correlation. Specifically, the inclusion of the attachment variables should cause an *increase* in the coefficient for SEL so that it moves outside the range of zero and .016.

Multivariate results. Working from within a path analytic framework, multivariate regression analyses were conducted to determine whether the attachment variables were suppressing the class effect on delinquency. The results of these analyses are presented in Table 2.

A comparison of the social class coefficients in the reduced form and structural equations reveals a positive suppressor effect in the data. Whereas the bivariate regression coefficient for SEL is extremely weak and nonsignificant, the effect of SEL increases after controlling for the attachment variables such that a one unit increment in SEL leads to an increase of almost five points in the delinquency measure. While the class effect which emerges is still rather weak, it does attain statistical significance and is far stronger than what was seen in the bivariate situation.[4]

Table 2. Reduced Form and Structural Equations for Delinquent Involvement

Independent Variables	*Equations* (1)	(2)
SEL	.016[1]	.071
	(1.094)[2]	(4.877)
GPA		–.053*
		(–.383)
TEACHERS		–.019
		(–.391)
IMPORTED		–.101*
		(–1.788)
COMMITTED		–.040
		(–.205)
HOMEWORK		–.135*
		(6.216)
R^2	.000	.047*

[1]Standardized coefficients
[2]Unstandardized coefficients
*$p<.05$

Alternative explanations. Several attempts were made to derive an alternative explanation for the results of this analysis. First, since the suppressor effect which emerged was positive, there was a question as to whether there may have been some interactions in the data; that is, whether at different levels of SEL, attachment served as a greater constraining influence for some adolescents than for others. Hence, interaction terms were created to test for significant increments to the R^2 for DELINQUENCY (Cohen and Cohen 1975:291-310). None of the interaction terms were significant, with the increments to R^2 being minimal.

Second, although the intercorrelations among the attachment variables were of insufficient magnitude to warrant strong concerns regarding multicolinearity (see Table 1), first-order partials were run in order to assess the role of each individual attachment variable as a suppressor in the relationship between SEL and DELINQUENCY. A suppressor effect for SEL was detected net of each attachment variable; i.e., each first-order partial revealed a higher coefficient for SEL than was noted in the bivariate situation. The combined set of variables yielded only a slightly higher SEL effect on DELINQUENCY than did any one (or different combination) of the attachment variables.

Third, since the YIT study offers two additional delinquency indexes (Interpersonal Aggression and Overall Frequency of Delinquent Acts), similar analyses were conducted in an attempt to replicate the suppressor effect noted

for Theft and Vandalism.[5] The results for the Overall Frequency Index (FREQUENCY) paralleled those found for Theft and Vandalism; i.e., the zero-order correlation between SEL and FREQUENCY was equal to zero, with a significant, positive class effect emerging in the multivariate equation. Contrary to the results for the other two delinquency indexes, however, SEL was significantly related to the Interpersonal Aggression Index (AGGRESS) in the bivariate situation and the association was shown to be entirely mediated by the attachment variables. Thus, there was no suppressor effect noted for SEL when predicting AGGRESS.

Although the discrepancy in results noted for AGGRESS is not surprising in light of the literature regarding differential processes with respect to type of offense, an attempt was made to determine whether the results of these analyses were simply statistical artifacts. Split-half analyses were conducted in which the sample was randomly divided into two subsamples of equal Ns. Duplicate analyses were conducted for each sample, with the results confirming the reliability of the findings. In short, the results for the aggregate sample were the same as those which emerged for the two randomly divided subsamples. Thus, the finding that attachment to school serves as a suppressor in the relationship between social class and involvement in delinquency appears to be supported.[6]

SUMMARY AND CONCLUSIONS

The state of the art in delinquency research is characterized by a tradition of models which afford little in the way of explanatory power. While the low R^2s which characterize the discipline may be attributed, in part, to methodological and theoretical issues, part of the explanation for the insufficiency of our models may lie in our failure to consider the statistical issue of suppression.

This report has discussed the necessity of searching for suppressor effects in studies of delinquency by focussing on the *processes* which lead to delinquent involvement. An empirical example was offered to demonstrate how such an approach might be applied in delinquency research. Beginning with a bivariate correlation between social class and delinquent involvement that was effectively zero, multivariate analyses were conducted in order to determine whether attachment to school served as a suppressor in this relationship. The results revealed that social class was related to delinquent involvement, but that the relationship was a function of class variations in the measures of attachment to school. Controlling for these variables permitted a significant class effect to emerge, despite the fact that no relationship between class and delinquency had been observed in the bivariate situation.

Although the class effect which emerged in this study was rather weak, it did emerge as a factor worthy of consideration. Researchers are urged to pay close attention to the etiologies of other predictor variables and to explore the

possibility of suppressor effects before choosing to omit what may turn out to be key predictors of delinquent involvement.

APPENDIX: INSTRUMENTS USED IN THE EXAMPLE ANALYSIS

Instruments for the example were derived from the initial interview schedule and questionnaire of the YIT project. The Time 1 information is extracted from the interviews and questionnaires which were administered in the Fall of 1966, and the Time 2 measure of delinquency, retrospective in nature, is taken from the data which were collected in the Spring of 1968.

Self-reported Delinquency (DELINQUENCY): The self-reported delinquency measure used in the YIT study is comprised of questions regarding the respondent's participation in nine offenses involving theft and vandalism over an eighteen-month period (the time between the first and second interviews). Specifically, the respondents were asked: "In the past eighteen months, how often have you done this?"

1. Taken something not belonging to you worth under $50.
2. Went onto someone's land or into some house or building when you weren't supposed to be there.
3. Set fire to someone else's property on purpose.
4. Damaged school property on purpose.
5. Taken something from a store without paying for it.
6. Taken a car that didn't belong to someone in your family without permission of the owner.
7. Taken an expensive part of a car without permission of the owner.
8. Taken something not belonging to you worth over $50.
9. Taken an inexpensive part of a car without permission of the owner.

Due to problems encountered during the data collection process, individual items are not available for analysis (Bachman, personal communication). Thus, delinquency is measured by an index and the respondent's score is the mean for a range of frequencies from "never" to "five or more times" (100 = Low, 600 = High). The scale was adapted from Gold (1966), but avoids the pitfalls of many self-report instruments by clearly specifying criminal (as opposed to status) offenses and by providing for considerable discrimination in the frequency of participation (Hindelang et al., 1979).

Socioeconomic level (SEL): A summary index of six equally weighted components was constructed for the YIT project, and is used in this example as the measure of an adolescent's social class. In addition to a measure of the respondent's socioeconomic status (as determined by father's occupation), this index offers an indicator of what may be considered the *quality* of the youth's home environment as measured by parents' educational attainments, number of rooms per person in the home, number of books in the home, and a

checklist of other possessions in the home. The respondent's score on this index is the mean of his scores on these items, each of which was ranked on a scale from one (low) to eight (high).

Academic performance (GPA): This measure is a single indicator of the respondent's reported overall average of his grades in the preceding year (10 = Low, 58 = High).

Perceived teachers' interest (TEACHERS): The degree to which the adolescent believes that his teachers are interested in him is measured by a scale of three questions. The first item specifically asks the respondent to assess whether his teachers are interested in him, while the remaining two questions ask him to report how often he and his teachers talk privately (3 = Low, 15 = High).

Importance of Education (IMPORTED): The degree to which an adolescent believes that getting an education is an end in itself is measured by a scale of six items which ask him to assess the value of education and the importance of being in school (6 = Low, 24 = High).

Commitment to education (COMMITTED): The YIT study offers an index of the adolescent's determination or "commitment" to finish high school in which the respondent is asked to decide whether any of the 25 possible situations might make him quit school. The assumption is that the more an adolescent is committed to getting a high school diploma, the more likely he will be to stay in school, regardless of experiences which might tempt him to do otherwise (25 = Low, 75 = High).

Time spent doing homework (HOMEWORK): The actual amount of time that the adolescent devotes to studying is measured by the respondent's report of the number of hours per week he spends doing homework (1 = Low, 8 = High).

FOOTNOTES

1. An exception involves some research which has dealt with specific types of delinquency, most notably drug use among adolescents (e.g., Akers et al., 1979).

2. This has been the approach of several researchers, where a relationship among these three factors was hypothesized without testing for causative effects (e.g., Stinchcombe, 1964; Hirschi, 1969; Elliott and Voss, 1974).

3. Williams (1968) presents a strong argument for the exclusion of non-whites when testing for social-class effects. Hence, in order to avoid the possibly confounding effects of race, this study utilizes an all-white sample.

4. Although the upper limit of the range of frequencies for DELINQUENCY is "5 or more," there is the possibility that the class effect has been further underestimated due to outliers (Elliott and Ageton, 1980).

5. The inaccessibility of the individual delinquency items precluded the possibility of creating a composite delinquency measure.

6. There is another possibility for the results of this study. The distributions of the variables in the YIT data were extremely skewed, making the possibility for distinguishing class variations in attachment and delinquency extremely difficult. Further, since the YIT data did not provide individual items for either the SEL or DELINQUENCY measures (providing instead summary measures), it was impossible to test alternative models for either attachment or delinquency. Consequently, the findings may be sample-specific, and researchers are urged to replicate these analyses in data sets which have previously failed to show class effects.

REFERENCES

AKERS, R. D., M. KROHN, L. LANZA-KADUCE, and M. RADOSEVICH
1979 "Social Learning and Deviant Behavior: A Specific Test of a General Theory." *American Sociological Review,* 44:636-55.

ALEXANDER, K. L., M. COOK, and E. L. MCDILL
1978 "Curriculum Tracking and Educational Stratification: Some Further Evidence." *American Sociological Review,* 43:47-66.

ALWIN, D. F., and R. M. HAUSER
1975 "The Decomposition of Effects in Path Analyses." *American Sociological Review,* 40:37-47.

AULTMAN, M. G., and C. F. WELLFORD
1979 "Towards an Integrated Model of Delinquency Causation: an Empirical Analysis." *Sociology and Social Research,* 63:316-27.

BACHMAN, J. G., P. M. O'MALLEY, and J. JOHNSTON
1978 *Youth in Transition: Volume VI, Adolescence to Adulthood—Change and Stability in the Lives of Young Men.* Ann Arbor, Michigan: Institute for Social Research.

BECKER, H. S.
1952 "Social Class Variations in the Teacher-Pupil Relationship." *Journal of Educational Sociology,* 25:451-65.

BRIATHWAITE, J.
1981 "The Myth of Social Class and Criminality Reconsidered." *American Sociological Review,* 46:36-57.

BRIAR, S., and I. PILIAVIN
1965 "Delinquency, Situational Inducements, and Commitment to Conformity." *Social Problems,* 13:35-45.

CERNKOVICH, S. A.
1978 "Evaluating Two Models of Delinquency Causation: Structural Theory and Control Theory." *Criminology,* 16:335-52.

CLAUSEN, J. A. (ed.)
1968 *Socialization and Society.* Boston: Little, Brown.

COHEN, J., and P. COHEN
1975 *Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences.* New York: Wiley.

CONGER, R. D.
1976 "Social Control and Social Learning Models of Delinquent Behavior: a Synthesis." *Criminology,* 14:17-40.

ELDER, G. H.
1974 *Children of the Great Depression: Social Change in Life Experience.* Chicago: University of Chicago Press.

ELLIOTT, D. S., and S. A. AGETON
1980 "Reconciling Race and Class Differences in Self-Reported and Official Estimates of Delinquency." *American Sociological Review,* 45:95-110.

ELLIOTT, D. S., and H. L. VOSS
1974 *Delinquency and Dropout.* Lexington, Massachusetts: Lexington Books.

FREASE, D. E.
1972 "Delinquency, Social Class and the Schools." *Sociology and Social Research,* 57:443-59.

GIBBONS, D. C.
1976 *Delinquency Behavior.* Englewood Cliffs, New Jersey: Prentice-Hall.

GOLD, M.
1966 "Undetected Delinquent Behavior." *Journal of Research in Crime and Delinquency,* 3:27-46.
1970 *Delinquent Behavior in an American City.* Belmont, California: Brooks/Cole.

HEPBURN, J. R.
1976 "Testing Alternative Models of Delinquency Causation." *Journal of Criminal Law and Criminology,* 67:450-60.

HINDELANG, M. J., T. HIRSCHI, and J. G. WEIS
1979 "Correlates of Delinquency: The Illusion of Discrepancy Between Self-Reports and Official Measures." *American Sociological Review,* 44:995-1014.

HIRSCHI, T.
1969 *Causes of Delinquency.* Berkeley, California: University of California Press.

HIRSCHI, T., and C. SELVIN
1967 *Delinquency Research: An Appraisal of Analytic Methods.* New York: Free Press.

JENSEN, G. F.
1972 "Parents, Peers, and Delinquent Action: a Test of the Differential Association Perspective." *American Journal of Sociology,* 78:562-75..

JOHNSON, R. E.
1979 *Juvenile Delinquency and Its Origins: An Integrated Theoretical Approach.* London: Cambridge University Press.

KERCKHOFF, A. C.
1972 *Socialization and Social Class.* Englewood Cliffs, New Jersey: Prentice-Hall.

KROHN, M. D., and J. MASSEY
1980 "Social Control and Delinquent Behavior: An Examination of the Elements of the Social Bond." *Sociological Quarterly,* 21:529-43.

LINDEN, R.
1978 "Myths of Middle-Class Delinquency: a Test of the Generalizability of Social Control Theory." *Youth and Society,* 9:407-32.

LOTZ, T.
1979 "Sociologists Benign Neglect of Juvenile Delinquency as a Social Problem." *Western Sociological Review,* 10:10-27.

MATZA, D.
1964 *Delinquency and Drift.* New York: Wiley.

MINOR, W. W.
1977 "A Deterrence-Control Theory of Crime." Pp. 117-37 in R. F. Meier (ed.), *Theory in Criminology: Contemporary Views.* Beverly Hills, California: Sage.

PHILLIPS, J. C., and D. H. KELLEY
1979 "School Failure and Delinquency: Which Comes First?" *Criminology,* 17:194-207.

POLK, K., and D. S. HALFERTY
1966 "Adolescence, Commitment and Delinquency." *Journal of Research in Crime and Delinquency,* 3:82-96.

ROSENBERG, M.
1968 *The Logic of Survey Analysis.* New York: Basic Books.

SCHUESSLER, K.
1971 "Continuities in Social Prediction." Pp. 302-39 in H. L. Costner (ed.), *Sociological Methodology 1971.* San Francisco: Jossey-Bass.

SEWELL, W. H., A. O. HALLER, and A. PORTES
1969 "The Educational and Early Occupational Attainment Process." *American Sociological Review,* 34:82-92.

SHOVER, N., S. NORLAND, J. JAMES, and W. E. THORNTON
1979 "Gender Roles and Delinquency." *Social Forces,* 58:162-75.

STINCHCOMBE, A. L.
1964 *Rebellion in a High School.* Chicago: Quadrangle Books.

TITTLE, C. R., W. VILLEMEZ, and D. SMITH
1978 "The Myth of Social Class and Criminality: an Empirical Assessment of the Empirical Evidence." *American Sociological Review,* 43:643-56.
1979 "Reply to Stark." *American Sociological Review,* 44:669-70.

WIATROWSKI, M. D., D. B. GRISWOLD, and M. K. ROBERTS
1981 "Social Control Theory and Delinquency." *American Sociological Review,* 46:525-41.

WILLIAMS, J. R.
1968 *Social Stratification and the Negro American: An Exploration of Some Problems in Social Class Measurement.* Duke University: Unpublished Dissertation.

VIII

DISSEMINATING AND UTILIZING EVALUATION DATA

In the traditional view of the conduct of either research in general or evaluation research in particular, dissemination and utilization are the final phases of the process. Indeed, some writers seemed to see a distinct separation between the planning-implementation-analysis phases and the dissemination-utilization phases. Part of the reason for this view was the lingering belief, based on the traditional laboratory model of conducting research, that the integrity of the research process could be maintained only if the conduct of the research was separated from the use of the research findings. More recent experience, however, has demonstrated that dissemination and utilization are an integral part of the entire process of evaluation and that one key to greater use of evaluation results is to start planning for that use very early in the process.

Alarmed by the apparent nonuse of so many evaluation studies, some evaluation researchers in the mid to late 1970s began to study the utilization process in detail. Among other things, they learned that many factors other than the scientific integrity of the findings were important in the use of research data. It also became apparent that use was a complex phenomenon, involving many different aspects. For example, use could be of an instrumental, conceptual, or symbolic type and could occur at just about any point in the research process. As evaluation researchers have continued to learn more about the dissemination and utilization process, it has become clear not only that there is more use of evaluation results than we might at first think but also that we need to spend a good deal more time thinking about how we can facilitate the use of evaluation.

The six papers in this section provide a variety of perspectives on the issues of dissemination and utilization. The Leviton and Boruch article that begins the set demonstrates that, as stated above, there is often more use than meets the eye. They focus on the contributions of 21 large-scale educational evaluation studies and conclude that data from these studies, particularly about implementation, contributed in many ways to changes in the conduct and management of the programs under study as well as to relevant policy changes.

The paper by Shapiro represents the kind of thoughtful analysis of the factors related to use and nonuse that has increased our understanding of the complexity of utilization. Shapiro rejects as too simple the "two cultures" (i.e., scientists versus policymakers) explanation for the use or nonuse of evaluation research results by policymakers and decision makers. Instead, he describes

four models of organizational decision making (rational choice, bureaucratic politics, organizational processes, and cognitive processing), all of which can relate to different aspects of the decision-making process, and discusses the implications of each of these for the conduct of evaluation. Shapiro's paper is noteworthy in challenging the assumption that evaluators frequently make that evaluation data of high quality are more likely to be used. Shapiro's four decision-making perspectives demonstrate that many other factors, most out of the control of the evaluator, are more important.

Deloria and Brookins provide a detailed look at a common and important part of dissemination and utilization efforts: the evaluation report. Following a brief description of the policymaker's environment, Deloria and Brookins analyze three evaluation reports that are policy-oriented rather than methods-oriented, as is the case with the traditional evaluation report. They conclude with 10 useful suggestions for the organization and content of policy-oriented evaluation reports.

One reason that evaluators sometimes do not write policy-oriented reports is that they are not sure their data are strong enough to guide policy. Grobstein shares his thoughts on this issue, which he faced as chair of the National Research Council's Committee on Diet, Nutrition, and Cancer. The committee produced a number of guidelines, one of which was that the amount of fat in the average American diet should be reduced from 40 percent to 30 percent of total caloric intake. Critics charged that the data were not strong enough to justify such a recommendation. While the level of data certainty required in the lab should remain high, Grobstein argues that the appropriate criterion in applied science is the "best available scientific information." This issue is one that frequently confronts evaluation researchers who understand that, with or without the data, policymakers must and will make decisions. Grobstein's criterion would suggest that we take a bolder role.

William Ruckelshaus, the administrator of the U.S. Environmental Protection Agency, confronts almost daily just the kind of situation that Grobstein describes. In the face of incomplete knowledge, he must make decisions about limiting or banning the use of toxic substances in our air, land, and water. In his paper in this section, Ruckelshaus provides a decision maker's perspective on how science can be helpful. He makes a useful distinction between the scientific side of the issue (i.e., risk assessment) and the policymaking side (i.e., risk management). Ruckelshaus believes that scientists can be most useful in assessing the risks of pollutants as rigorously as possible, then leaving the risk management to the policymakers. The organizational model that Ruckelshaus is proposing for the conduct of policy-relevant research would suggest that evaluators pay closest attention to the "science" of their enterprise, at least for evaluations of the products and processes similar to those addressed by the Environmental Protection Agency.

The final paper reports on the effects of a particular organizational scheme intended to produce useful evaluation data. This scheme was instituted in

Canada by the Office of the Comptroller General and involved establishing evaluation units in all Canadian federal agencies and departments. The article published here is the synopsis from a study of this innovative organizational plan conducted by the Canadian Auditor General's Office (similar to the U.S. General Accounting Office). The Canadian Government's response to the report also is included. The study involved a survey of 19 (out of a total of 56) Canadian federal departments and agencies working with the Comptroller General's Office. In its survey, the Auditor General's Office focused on the extent to which agencies and departments had instituted the organizational arrangements and evaluation practices as set out by the Office of the Comptroller General. The study has important implications for the conduct and subsequent use of evaluation in similar contexts, that is, large organizations that have the option of selecting centralized or decentralized organizational arrangements and oversight of evaluation activities.

As the variety of papers in this section demonstrates, dissemination and utilization are complex and multifaceted. While dissemination and utilization sometimes happen automatically, more often their occurrence requires careful thought and planning. As with many other aspects of the evaluation process, the sooner and more thoroughly these components are planned, the more likely it is that they will occur and that the evaluator will be sensitized to new and different opportunities for use.

33

Contributions of Evaluation to Education Programs and Policy

Laura C. Leviton and Robert F. Boruch

A truism of evaluation is that studies seldom contribute directly to identifiable decisions (Patton et al., 1977). The claim appears in several texts advocating reforms in evaluation. In Guba and Lincoln (1981) for example, it is averred that the failure to use evaluation findings has "almost assumed the proportions of a national scandal" (p.ix). In the "Ninety-Five Theses" of Cronbach et al., (1980), we find the assertion that "everyone agrees that evaluation is not rendering the service it should" (p. 3). Dunn et al., (1981) base their suggestions for reform on the presumption that evaluations are peripheral because they are seldom utilized.

The truism deserves a second look because the research on which it is based is almost a decade old, was scanty to begin with, and faced serious

AUTHORS' NOTE: *This research was supported by contract 300-79-0467 from the Department of Education to Northwestern University. We wish to thank John Evans, Janice Anderson, and the many people who patiently responded to our interviews. The work stems from a report to Congress and the Department of Education on the evaluation of federally supported education programs at the national, state, and local levels of government (Boruch and Cordray, 1980).*

From Laura C. Leviton and Robert F. Boruch, "Contributions of Evaluation to Education Programs and Policy," *Evaluation Review*, 1983, 7(5), 563–598.

obstacles to accurately assessing use. While evaluators accept that the linkage to program change may not be straightforward, they also tend to believe that this linkage is mostly absent. Just as serious, Congressional aides may share the belief (Saxe and Koretz, 1982) and cut budgets accordingly.

Reform of evaluation may indeed be warranted on a variety of grounds, including the failure to use findings. However, reformers must do more than reassert the claim of nonuse. They must show that, for a given audience and setting, the level of influence is in fact low. They must also base their prescriptions on documented examples of evaluations that were used in important ways. Some examples exist, but more are needed to understand the variety of possible contributions.

The present study addresses empirically the issue of use and nonuse. It stems from a Congressionally mandated appraisal of educational evaluation at the national, state, and local levels of government (Boruch and Cordray, 1980). Our findings challenge the truism of nonuse for the federal level in education. Evaluations at this level were found to make frequent and important contributions to decisions. They were not the only input to those decisions, of course.

Some distinctions are important to avoid the confusion of earlier studies. We distinguish between "use" and "impact," defining the former as serious consideration of findings, which may or may not relate to decisions; and the latter as actual changes in programs resulting from use (Leviton and Hughes, 1981). The evidence that is brought to bear then concerns specific links between evaluation and serious consideration, decisions, and impact. Both oral testimony and paper trails were used as evidence and, moreover, we demanded corroboration of these trails from more than one source. It is possible that trails may end with serious consideration that enlightens decisionmakers and contributes indirectly (Weiss, 1977). Trails may end with decisions but no impact, because impact takes time, and change in programs involves decisions by many people (Pressman and Wildavsky, 1973). In other cases, trails may lead to amendments, regulations, and management changes.

METHOD

SAMPLE OF STUDIES

We drew two samples in this study. The first was from the evaluation unit of the Office of Education (OE). We selected fourteen evaluations

from a total of 31 completed studies described in the "Highlights" section of OE's Annual Report on evaluation for 1978 and 1979. Excluded from the definition of evaluative studies were reports on finance, management of technical assistance, and economic projections. Although we examined slightly less than half the evaluations, we cannot claim our sample is representative. We may have selected reports that were utilized more than the ones we did not study. However, even these were often used, according to descriptions (not verified by us) in the OE *Annual Reports.*

The second sample consisted of seven studies sponsored by a variety of federal agencies dealing with education. These studies were visible and widely discussed, according to those we interviewed. They were readily accessible to us and seemed a good place to find contributions. Both samples were purposive and, we believe, biased in favor of discovering contributions. Table 1 lists the studies examined, types of information covered, their general area of education policy, their sponsor, and the firm conducting the study.

Several kinds of information from these evaluations could contribute to programs and policy. Implementation findings indicate how local and state agencies delivered education programs. Outcome information consists of causal inferences about effects of programs on students and schools. Information about federal administration reveals problem areas of management, or addresses questions of entitlement and resource allocation. The cost and cost-effectiveness of programs constitutes another category. Frequently, several kinds of findings were involved in a single evaluation.[1] Sixteen of the studies, or 65%, presented implementation findings. Eleven of the studies, or 52%, presented program outcomes. Four studies, or 19%, presented information about federal administration, while almost all made some reference to costs.

DOCUMENTATION OF CONTRIBUTIONS

For each case study of a report, we developed a pool of informants and relevant documents by using a snowball sampling technique described by Rich (1979). The pool usually started with telephone interviews of the agency project officer and of the contractor who carried out the research. Sometimes, other respondents and relevant documents, such as Congressional reports, were available at the beginning of the case study. All respondents then named others who might be knowledgeable about use, and identified supporting documents.

TABLE 1
Evaluations Examined for Their Contributions

OE Sample	*Topic and Type of Information Provided*	*Policy Area*	*Contractor*[a]
1979	Services to Neglected and Delinquent Youth (implementation)	ESEA, Title I (Compensatory education)	Systems Development Corporation
	Sustaining Effects of Compensatory Education (implementation)	ESEA, Title I (Compensatory education)	Systems Development Corporation
	Magnet Schools (implementation, outcome)	ESAA (desegregation)	Abt Associates
	Survey of Educational TV Viewership (implementation)	ESAA (desegregation)	Applied Management Sciences
	Sex Equity in Programs (implementation)	Vocational Education	American Institutes for Research
	Campus-Based Aid and Basic Grant Programs (implementation)	Higher Education	Applied Management Sciences
	Project Implementation Packages (implementation, outcome)	ESEA, Title IV-C (innovations)	American Institutes for Research
1978	Aid to Non-Profit Organizations (implementation, outcome)	ESAA (desegregation)	RAND Corporation
	Facilities Survey (implementation)	Vocational Education	Westat
	Exemplary Programs (outcome)	Career Education	American Institutes for Research
	Upward Bound (outcome, implementation, Federal administration)	Higher Education	Research Triangle Institute
	Follow Through Planned Variations Study (outcome)	Economic Opportunity Act (Compensatory education)	Abt Associates
	Bilingual Education (outcome, implementation)	ESEA, Title VII	American Institutes for Research
	Survey of Programs (implementation)	Indian Education	Communications Technology Corp.

TABLE 1 (Continued)

Sector-Wide Sample Sponsor[b]	*Topic and Type of Information Provided*	*Policy Area*	*Contractor*[a]
ASPE	Fund for the Improvement of Postsecondary Education (outcome)	Higher Education	NTS Research Corporation
ASPE	Follow Through Exploratory Evaluation (Federal administration)	Economic Opportunity Act (Compensatory education)	in-house
ASPE	Impact Aid (Federal administration)	ESEA, Aid to Schools in Federally Impacted Areas	in-house
OE, 1977	Federal Programs Supporting Educational Change (implementation, outcome)	ESEA, Several Titles	RAND Corporation
NIE	Compensatory Education Study (implementation, Federal administration, outcome)	ESEA, Title I (compensatory education)	in house with many sub-contractors[a]
NIE	Achievement Testing (implementation)	ESEA, Title I (compensatory education)	SRI International
ACYF	National Day Care Study (implementation, outcome)	HEW Day Care (interagency)	Abt Associates

a. References can be found in Leviton and Boruch, 1980.

b. Abbreviations indicate the following organizations in the Department of Health, Education and Welfare:

ASPE — Assistant Secretary for Planning and Evaluation
OE — Office of Education
NIE — National Institute of Education
ACYF — Administration for Children, Youth and Families

To the extent possible, all information was verified from additional sources independent of the snowball sample. For example, Congressional staffers were asked about uses of research information in legislation, and independent observers supplied background information on the research reports.

For the purposes of our study, contributions had to be verified by more than one source. These sources could be two or more independent respondents, or documents supporting the contentions of a single respondent. However, statements by the project officer and the contractor or grantee were not counted as corroborating each other, because they had too similar an interest in showing the study's contributions. If, on the other hand, the contractor or project officer said that a third individual, such as a program manager, had used the information, the third individual's description of the use of the information was viewed as corroboration.

Unconfirmed Instances

Some contributions could not be confirmed easily from more than one source. Conceptual uses provide an example. If a single OE staff person said that findings helped him or her to outline legislative issues for the administration, this private use might not be confirmed by others, and the relevant memos might have been long gone. Rather than eliminate unconfirmed instances, we present them separately, in order to preserve information that has, admittedly a lower standard of evidence. These instances constituted 13% of the total number of contributions discovered.

CONTENT ANALYSIS OF CASE STUDIES

The contributions of these evaluations are described at length in Leviton and Boruch (1980). In order to provide a numerical summary, we characterized the types of use and impact involved, as well as the kinds of findings that contributed in each instance.

Types of Contributions

We classified contributions into those having impact on programs or policy, those that influenced decisions without achieving impact (at the

time of our investigation), and those that involved serious consideration but no decision that we could verify. Serious consideration might involve either the use of findings to persuade others to a point of view, as when lobbyists present their agenda and support it with findings; or conceptual use, as when an official tells us that evaluation revealed new problems or new solutions in a program. Evaluations contributing to impact have also contributed to decisions; evaluations contributing to decisions have also been seriously considered. The paper trail or oral history had to verify the linkages between these levels in order for the instance to be classified at the highest level.

Decision Contexts

The contributions we located occurred in specific decision contexts. Thus, in legislation, discussions of the implications of findings by Congressional staff, lobbyists, or education officials would be termed a use of information, but one that does not go beyond serious consideration. If people said that the findings were used in drafting a bill, or if the rationale for the bill cited findings, the instance was classified as use in decision-making. An actual change in legislation, in which independent parties claimed findings had been used, was classified as contribution to impact. Generally, a finding might partially motivate decisions or impact, or it might influence the form of decisions or impact.

Similarly, evaluations could contribute to consideration of the budget, to budget proposals (decisions), or to actual appropriations and authorizations (impact). Findings could contribute to discussions of regulations, to drafting them (decisions), or to final regulations (impact). Findings could be used in thinking about management, in decisions and orders, or in observable changes in management (impact). In similar ways, findings could contribute to states' and localities' decisionmaking. Reseach and development (R&D) activities were found to involve government analysis of data for other purposes, as well as for funding of new research projects. This category was classified as "decisionmaking" in that no impact on a program was evident in the short run. However, people could also "seriously consider" new R&D and how it should be carried out, as in books on evaluation methods.

One kind of serious consideration did not fit nicely into these decision contexts: The use of findings in editorials in the media. Findings were seriously considered by influential opinionmakers, but did not relate to specific decision contexts. These were the only such cases in the sample.

Unit of Analysis

We chose the individual contribution of a finding as the unit of analysis. This could be a qualitatively distinct impact, or a distinct decision. "Distinctiveness" was judged by the content of the contribution. Thus two decisions are counted in the following example from the RAND Study, "Federal Programs Supporting Educational Change":

> [One Administration Staff member] mentioned that her office was disturbed at the RAND finding that innovative projects were isolated from the rest of the school. They therefore inserted language in the Administration's proposal that local districts must make a commitment to spend some money on the project themselves, over time, while the federal share would decrease. [This office] also inserted language in the proposal that projects should be integrated into the rest of the school [Leviton and Boruch, 1980].

The decision on funding is different from the decision on integration. Alternative units could legitimately have been used; however most of them lose information.[2]

Because we chose qualitatively distinct contributions for analysis, and because agreement was needed about what constituted a contribution, inter-rater reliability was at issue. After training in classifying three case studies, two raters reached relatively good agreement on the remainder: a kappa of .84 in rating contributions, and .75 for types of information.

RESULTS

We first present our content analysis, to reveal the extent of contributions by types of information in different decision contexts. However, this numerical summary is indeed shallow unless the reader understands the nature of these contributions in their policy contexts. Thus, we highlight some of these contirbutions and suggest reason for them. Space does not permit a full presentation, which may be found in Leviton and Boruch (1980).

CONTENT ANALYSIS

We located 156 distinct contributions of evaluation. When unconfirmed instances are included, the total rises to 180. Our verification trail

led to impact 68 times (76 verified and unverified), to decisions only, 61 times (67 verified and unverified) and to "serious consideration" only, 27 times (37 verified and unverified). Because one can never be sure one has located all contributrions in a retrospective study, all numbers presented here should only suggest an emphasis given to types of information and types of contribution. The purposive sample may be biased in favor of detecting contributions; the retrospective method is biased against detection. We cannot even guess the direction of possible bias in favor of one kind of contribution over another.

Table 2 reveals these emphases. Looking at the types of contributions, it is clear that impacts are relatively more frequent than we might expect, given the difficulty of obtaining impact, and "consideration only" is surprisingly less frequent, given the evidence of past studies (Patton, et al., 1977; Weiss and Bucuvalas, 1977). In all likelihood, this retrospective method was biased toward reporting impacts and decisions, which people remember, and which are more frequently public acts. Nevertheless, the evidence is extremely revealing.

Implementation information contributed most frequently in this set of studies. Federal administration information contributed fairly often, especially when one remembers the small number of reports containing this information. Outcome information did contribute, and at high levels, but not as much as one might expect given its emphasis in the reports we studied. Cost and cost-effectiveness information did not contribute often in this set of studies (it was not a major focus for most), but contributed to impact in all three cases.

Table 3 presents contributions by the kind of information contributing, and the decision context of the contribution. We were able to locate primarily federal-level contributions of these federal-level reports. This may represent a bias in the method of sampling respondents, or it may be that findings were not used by state and local audiences. Our federal-level informants did not name many contacts at these levels.

Findings contributed most often to the form of legislation, both proposed and passed into law. Moreover, the implementation information, and to a lesser extent, federal administration information, contributed most often to legislation activities. Findings rarely contributed to budget allocations—but decisions, in the form of budget proposals to Congress, did incorporate findings relatively often, primarily to argue for increases or decreases. Again, implementation information contributed most often.

Contributions relating to the drafting of regulations may, with time, have become impacts as final regulations were approved.[3] Twelve of

TABLE 2
Contributions of Evaluation by Type of Information and Level of Contribution

Contributions[a]	*Type of Information*						
	Implementation	Outcomes	Implementation and Outcomes	Federal Administration	Federal Administration and Implementation	Cost and Cost-Effectiveness	Total
Total Contributions	67 (84)	22 (24)	25 (29)	29	10 (11)	3	156 (180)
Total Impact	28 (33)	9 (11)	15 (16)	10	3	3	68 (76)
Total Decisions only	29 (34)	4	5	16	7 (8)	0	61 (67)
Total "Consideration" only	10 (17)	9	5 (8)	3	0	0	27 (37)

Type of Information	*Proportion of Studies Presenting Information*[b]	*Proportion of Contributions by Information*[b]
Implementation	.76	.65 (.69)
Outcome	.52	.30 (.29)
Federal administration	.19	.25 (.22)
Cost effectiveness[c]	.10	.02 (.02)

a. Verified plus unverified instances appear in parentheses.
b. Proportions do not add to 1.0 because of overlap in content of studies and contributions.
c. Costs were mentioned in almost all reports.

these decisions were Congressional orders to the effect that federal education administrators draft regulations. These orders involved implementation and federal administration information. The officials writing the regulations tended to use implementation and the combination of implementation and outcome information.

TABLE 3
Contributions of Evaluation by Type of Information and Decision Context

Decision Context	*Type of Information*						
	Implementation	Outcomes	Implementation and Outcomes	Federal Administration	Federal Administration and Implementation	Cost and Cost-Effectiveness	Total
Legislative[a]							
Impact (Law)	21 (25)	1	4	8	3	1	38 (42)
Decisions (proposals)	8 (9)	0	0	6	1 (2)	0	15 (17)
"Consideration"	9 (10)	2	1	1	0	0	13 (14)
Budget							
Impact (authorization, appropriations)	1	2 (4)	1	0	0	0	4 (6)
Decisions (proposals)	5 (8)	0	0	0	2	0	7 (10)
"Consideration"	0	0	0	1	0	0	1
Regulations							
Impact (passage)	3	0	5	0	0	2	10
Decisions (drafting)	4 (5)	0	1	4	4	0	13 (14)
Management							
Impact	0 (1)	6	2	2	0	0	10 (11)
Decisions	5	4	1	2	0	0	12
"Consideration"	1 (2)	0	0 (2)	1	0	0	2 (5)
State							
Impact	0	0	3 (4)	0	0	0	3 (4)
Decisions	0	0	1	0	0	0	1
"Consideration"	0 (5)	0	0 (1)	0	0	0	0 (6)
Local							
Impact	3	0	0	0	0	0	3
Research and Development[b]							
Decisions	7	0	2	4	0	0	13
"Consideration"	0	3	0	0	0	0	3
Other "Consideration"							
Media Editorials	0	4	4	0	0	0	8

a. Verified plus unverified instances appear in parentheses.
b. Two mandated studies were included under legislative impact.

In the instances involving management, pure outcome information was used in ten exceptional cases, involving the Joint Dissemination Review Panel (JDRP). This panel approves funding for adoption and dissemination of educational innovations, based on evidence of effectiveness. Otherwise, managers did not use outcome information alone. This is especially surprising in the case of decisions to fund further research (R&D). The three instances of "consideration" of pure outcome data involve knowledge gained about how to run an evaluation. In most other instances, R&D was funded to solve problems revealed by implementation and federal administration information.

The National Institute of Education (NIE) Compensatory Education Study accounted for 43 of the contributions we discovered. The study was an "outlier" in terms of contributions, and unusual in other respects as well. To examine whether it contributed differently from other evaluations, we contrast it with others in Table 4. The NIE study contributed in 26 of the 38 instances of legislation, and in the 12 instances in which Congress ordered new regulations to be written. Congress made heavy use of NIE's information about implementation and federal administration, especially in legislation. The legislative contributions of other studies involved implementation most often as well, but less than half of these achieved impact. Although federal administration information from other studies was used, it did not achieve impact in legislation or regulation.

In Table 5, we present the contributions made by each of the studies. The median number of impacts was two, the median for decisions only was one, and the median for serious consideration was one (whether one looks at verified instances only, or verified plus unverified instances). Table 5 also reveals that our sector-wide sample of seven studies accounts for more than half of the total contributions discovered. The NIE Study is largely responsible for this; however, the study, "Federal Programs Supporting Educational Change," ranks second in the number of contributions.

HIGHLIGHTS OF CONTRIBUTIONS

Title I: NIE Compensatory Education Study

This study was a synthesis of 35 separate research projects, involving alternative eligibility criteria, services and their effects on students, and

TABLE 4
Contrast of NIE Compensatory Education Study with All Others

Contributions[a]	*Information*			
	Implementation	*Federal Administration*	*Implementation & Federal Admin.*	*Other*
Legislative:				
NIE	14	8	3	1
All Others	7 (11)	0	0	5
Legislative Bills:				
NIE	0	1	0	0
All others	8 (9)	5	1 (2)	0
Regulation:				
NIE	0	0	0	0
All Others	3	0	0	7
Regulation—Writing:				
NIE	4	4	4	0
All Others	0 (1)	0	0	1
Other Contributions:				
NIE	1	1	0	1
All Others	30 (41)	10	2	35 (41)

a. Verified plus unverified instances appear in parentheses.

management of the program at federal, state, and local levels. Readers interested in the details of the study should consult Hill (1980). Leviton and Boruch (in press) describe the various contributions of the study. Hearings on Title I took their form directly from the six topics of the reports. In addition, every respondent felt the contributions were important, and the Senate and House reports for the legislation specifically thank the staff and commend their highly useful work, which it "consequently relied on . . . in formulating Amendments to Title I " (House Report, p. 5.).

The information on implementation largely concerned troublesome variations in the state and local management of Title I programs. NIE

TABLE 5
Contributions Made by Each Study

OE Sample	*Contributions*[a]			
Topic	*Consideration, Decisions, and Impact*	*Consideration and Decisions Only*	*Consideration Only*	*Total*
Services to Neglected and Delinquent Youth	4	1	1 (2)	6 (7)
Sustaining Effects of Compensatory Education	0 (1)	5 (6)	3	8 (10)
Magnet Schools	3	2	2 (4)	7 (9)
Survey of ESAA TV Viewership	1	0 (1)	0	1 (2)
Sex Equity in Vocational Education Programs	0	0	0 (1)	0 (1)
Campus Based Aid and Basic Programs	0	6	0	6
Project Implementation Packages	3	1	1	5
Aid to Non-Profit Organizations	5	0	0	5
Vocational Education Facilities Survey	0	4 (6)	5 (10)	9 (16)
Exemplary Programs in Career Education	6	4	0	10
Upward Bound	0 (2)	1 (2)	0	1 (4)
Follow Through Planned Variations Study	0	0	7	7
Bilingual Education	5 (9)	4	1	10 (14)
Survey of Indian Education Programs	0	1 (2)	0	1 (2)
Total OE Contributions	27 (34)	29 (35)	20 (29)	76 (98)

TABLE 5 (Continued)

OE Sample	*Contributions*[a]			
Topic	*Consideration, Decisions, and Impact*	*Consideration and Decisions Only*	*Consideration Only*	*Total*
Sector-wide Sample				
Fund for the Improvement of Postsecondary Education	2	0	0	2
Follow Through Exploratory Evaluation	2	6	0	8
Impact Aid	0	5	1	6
Federal Programs Supporting Educational Change	4 (5)	5	3 (4)	12 (14)
NIE Compensatory Education Study	27	14	2	43
Achievement Testing	2	1	0	3
National Day Care Study	4	1	1	6
Total Sector-side Sample Contributions	41 (42)	32	7 (8)	80 (82)
Total Contributions	68 (76)	61 (67)	27 (37)	156 (180)

a. Verified plus unverified instances appear in parentheses.

attributed many of these problems to lack of clarity in law, regulation, and enforcement by OE—federal administration information. Congress clarified the law, borrowing from model legislation generated by NIE, and directed OE to clarify regulations. Examples include the methods for allocating federal and local compensatory education funds, legal models for programs in elementary and high school, and clarification of the states' role in monitoring compliance. In addition, outcome information was used both in changing Congress' thinking toward a more positive view of what the program could accomplish, and in arguments increasing the budget authorization.

The Study possessed many of the features claimed to be important in a usable evaluation: communication with users, asking relevant questions through careful planning, consultation with many stakeholders, and timeliness of results. In addition, most of the contributions related to changes in program administration. As a major source of federal education money, Title I was too politically impacted for Congress to consider drastic changes in the pre-Reagan era.

Title I: The Sustaining Effects Study

A preliminary report from this longitudinal study examined eligibility of students. The report and similar findings from the NIE study fueled a debate in Congress over the criteria of eligibility for Title I services. Because the Sustaining Effects Study showed only 39% of poor children were being served, the committee chairman could argue that funds should continue to be targeted to the poor. Because only 40% of low-achieving students were served, the ranking minority member could argue that funds should be targeted to this group regardless of income. The debate was not resolved, and neither side was happy with a study that could provide ammunition for both. Our view, of course, is that it did inform the debate—although it went no further than "serious consideration" in support of opposing policy positions.

The study was used, however, in federal management of the program; for example, in renewed efforts to better target services to appropriate children. The study took a form more useful to managers, who must rely on numbers served in program operations, than to Congress.

Title I: Achievement Testing

This study recommended testing students in the fall, to encourage uniformity of reporting. Scores tend to decline over the summer months, so that spring testing can inflate estimates of the effect of Title I programs. Congress amended the law, requiring that evaluation of Title I programs be based on achievement testing "over at least a twelve-month period in order to determine whether school year programs have sustaining effects over the summer." In addition, the law was changed to allow funding of summer bridge programs. The origins of this change lie

in an earlier bill that also cited this study. In this example we see how contributions of evaluation may take more than one legislative cycle.

Title I: Services to Neglected and Delinquent Youth

A major finding of this evaluation was that Title I served only half of the eligible students in state institutions for neglected and delinquent children. Moreover, Title I students received less time in instruction than non-Title I students. Congress used these findings in amending Title I, to insure that students would get more instructional time. The program manager gave this study his full cooperation, because he wanted policy makers to become aware of the problems and resolve them. The problems were beginning to be addressed as a result of the study. An exploratory evaluation was contracted to further describe the problems of state noncompliance. State and local educators took the study back to their institutions to argue for improvements. We were able to trace one Director of Education at a school for delinquent boys who used the study to convince the school to change the curriculum.

Desegregation: Aid to Non-Profit Organizations

These organizations received federal funds to facilitate school desegregation, but the evaluation found that they did not do this well. Congress used the study in converting the funding for these organizations from a state apportionment to competitive grants, and cut the budget for the program by two-thirds (it was later restored). The study also found that nonprofit organizations frequently provided compensatory education services. Congress amended the law to focus permitted activities on desegregation, and subsequent regulations prohibited use of funds for compensatory education.

The study revealed that effective organizations used citizen action strategies. Revised regulations for the program reflected this finding in requiring that such organizations have experience in working effectively with other community organizations. Thus, the study contributed to impact by identifying an important problem and by pointing to effective methods of assisting desegregation. Identification of successful practices is frequently valued, and more examples appear below.

Desegregation: Magnet Schools

These schools have special resources designed to attract students to the school and to facilitate voluntary desegregation. The study revealed that people were attracted to the schools. However, magnet schools were found to be most effective as part of an overall desegregation plan, and OE was found to be awarding funds to school districts with poor desegregation records. Generally, the program was found to be growing out of proportion to actual desegregation activities.

Congress essentially ignored these negative findings and doubled the 1980 appropriation for the program. The Appropriations Committee noted, "A recent evaluation shows that magnet schools are an effective tool in helping to improve community attitudes toward schools," (Senate Appropriations Committee, 1979, p. 107). The program was popular with Congress, perhaps because it gave the appearance of handling desegregation in a positive way. The Administration, on the other hand, asked for a recision of the 1980 appropriation, and for reduced funding in 1981.

Some might term the budget impact a misuse of the findings by the Congress. However, it should be understood in light of the debate over magnet schools' acceptability. The findings contributed to decisions and impacts because they strengthened arguments in favor of the schools. It is up to decisionmakers to draw implications from the findings, even when evaluators would argue for a different set. As Weiss (1973) notes, "politicians have a different model of rationality in mind."

Desegregation: Survey of ESAA TV Viewership

ESAA funded public television series to facilitate desegregation. The shows were aimed at minority children of school age. A survey of viewership found, however, that less than one third of the children interviewed recalled seeing even one show. The managers of the program claimed they were already aware of the viewership problem before the survey was completed, and they had already taken actions to improve viewership. Because management change was under way, the survey rapidly became obsolete. This may have happened because of poor communication and planning, bad luck, or because a survey made management sensitive to the problem and eager to solve it.

Educational Innovation: Federal Programs Supporting Educational Change

This study surveyed school districts implementing innovations supported by OE. Although some districts started innovative projects to meet needs, others did so merely to obtain federal money. These projects fell apart, as did serious innovations, if they did not have proper planning and the commitment of participants. The study also identified factors associated with the continuation of projects after federal funding ended. The House report noted these findings in amending the law for Title IV-C, which funds adoption of innovations in school districts. Projects would be funded for a maximum of five years, with a declining federal share.

This evaluation was one of the first to cast light on the problems of implementing changes in a situation of local control. The revision in our thinking about local control is probably the most profound effect of this study and has earned it much citation. The management uses of the study, at both federal and state levels, were probably most important as government officials learned they must invite, not mandate, change in schools. In fact, California passed two laws to encourage local development of needed reforms.

Educational Innovation: Project Implementation Packages

This study revealed that, even when explicit instructions are given for replicating a successful project, modifications will be necessary when project models are adopted by new school districts. According to interviewees, the study challenged the assumptions of federal managers who had believed complete replication would happen. The federal management of Title IV-C thereafter funded project developers to provide individual technical assistance to schools adopting their projects. Together with the study of Federal Programs Supporting Educational Change, this study revealed to federal funders the nature of implementation of local projects and the extent to which ideas could feasibly be transferred.

Educational Innovation: Exemplary Career Education Projects

Effective ways of teaching career education were identified in this study. The Joint Dissemination Review Panel reviewed the ten evaluated projects and decided to approve funding of seven for dissemination and adoption. Six of the projects accepted this funding, and there have been many adoptions of these innovations throughout the country. "Impact", in the final analysis, consists of these adoptions but dissemination of the project models is also impact, as we defined it.

Vocational Education: Facilities Survey

This survey, the first comprehensive information on vocational education facilities, provided a data base used by federal researchers in the areas of sex equity, youth unemployment, and education finance. Federal managers used the data to plan. It served its intended audience: managers and researchers.

Vocational Education: Sex Equity

This study concluded that students continued to be concentrated in classes stereotyped as "appropriate" for their sex. An immense amount of public interest was generated by this conclusion, but we were able to locate only one, unconfirmed use of this information at the time of our study. Moreover, the evaluation was seriously misused, in that people drew negative implications for state sex equity coordinators, when their positions had only been authorized one year before the study concluded. An additional problem was that clearance by HHS took almost two years. The study had only been available for six months when we investigated its use. It may well have been used later—such events take time for people to digest the information and put proposals for change on the agenda.

Higher Education: The Fund for Postsecondary Education

An evaluation of this fund concluded that, by many measures, it was highly successful. The evaluation was used to justify authorizing an

increased level of funding, and the fund was transferred from the Education Provision Act to the Higher Education Act, "to give the legislative visibility deemed appropriate by the committee" (House of Representatives, 1979, p. 56). We cannot say that the evaluation, strong praise though it may be, was the sole motivation for the increased budget. It nevertheless contributed by strengthening arguments in favor of an increase.

Higher Education: Campus-Based Aid

Because this survey involved four student aid programs and provided information on student characteristics, it was more useful than existing data that examined only one program at a time. The American Council on Education produced statistical analyses of the data to use in its own policy positions on student aid. The data showed a need for a change in the BEOG program. Together with other groups, ACE proposed a change in the program that "was substantially incorporated into" a House bill (House of Representatives, 1979, p. 18). At the time of this study, the bill had not yet become law. This example illustrates how an agenda, set in part by evaluation, can emerge from any of a number of sources—provided the data are made available. In this case, management information that was shared with a lobbying group supported a perceived need for change.

Higher Education: Upward Bound

This evaluation which was a follow-up of an earlier study followed the progress of poor and minority students through high school and college. It concluded that Upward Bound was successful in getting more poor and minority students to attend college. Several federal agency staff believed that the positive evaluations of this program contributed to increases in the budget. This is likely because each year Congress appropriated more than the Administration requested. Moreover, the budget, which had been level for four years, began to increase steadily after the release of the first report. However, we were unable to find a direct linkage to budget impact, beyond these beliefs.

At the time of our study, lobbying groups in postsecondary education were pressing for a change in the Upward Bound program,

and according to one respondent it was highly likely that their proposal, which used the Upward Bound data, would change the poverty criterion for the program. Also, regulations were being written that employed criteria for awards based on outcomes that evaluations had shown to be measurable.

Follow Through: Planned Variations Study

After expenditures of over $20 million in evaluation over a decade, this outcome evaluation concluded Follow Through was not superior to regular classroom instruction, and that it was questionable whether there was any clearly superior Follow Through model. We could locate at best "serious consideration" of the findings in the press, for the "back to basic" movement. When the study was in process, it also contributed to the state of the art of evaluation itself (Rivlin and Timpane, 1975; General Accounting Office, 1975; Cronbach et al., 1980). The information was probably not the kind that policy and program audiences could readily use. The program had a vocal constituency that marshalled support against any threat. For reasons other than the evaluation, the Department of Education had been seeking budget cuts. Outcome information could not easily be used in this politically impacted setting.

Follow Through: Exploratory Evaluation

This evaluation revealed conflicting views of the mission of the program that needed resolution, and disagreement about management objectives as well. The central Follow Through Office was found to have no means of producing effective services or evaluating their effectiveness. The study recommended several management changes that were implemented. For example, the Follow Through Office was reorganized, and the sponsors of project models were given a new role in local service and knowledge production. This study capitalized on people's unhappiness with the program and skirted the effectiveness issue to focus on the central management problems that were feasible to address. The existence of the program was a given; improvement of existing arrangements was the focus, and for this reason, impact was obtained.

Bilingual Education

The evaluation of this program revealed, among other things, that Hispanic students in bilingual classes were outperformed in English proficiency by Hispanic students in regular classes. In addition, less than one-third of students in bilingual classes had limited English-speaking ability and did not leave the program when they did become proficient. These findings stirred a great deal of political opposition and criticism, but Congress nevertheless used them in justifying several amendments. The law was strengthened to indicate that measures must be developed to determine when children no longer needed assistance. The definition of the target population was expanded to include those limited in reading and understanding, the measures used in the study. In addition, Congress limited the number of English-speaking children in bilingual classrooms. Finally, a five year limit was placed on federal support for any one bilingual project.

Some believe that this study contributed little to policy on bilingual education. On the contrary, one can argue that policy began to change, though not in a bald, public way. After methodological critiques and political opposition by bilingual education advocates, Congress sidestepped a redefinition of program goals. However, it took steps to make sure that English proficiency would be emphasized. In directing the bureaucracy to develop criteria for children to enter and leave the program, Congress passed on the hot potato. OE took steps to develop these criteria, but they are necessarily controversial.

National Day Care Study

This study provided persuasive evidence that increasing the number of children per staff member in a day care center would reduce cost without affecting the quality of care. Regulations for day care programs used the study's recommended age categories and adopted one of its three policy options for specifying group size and staff to child ratios. The study found that specialized training was linked to better care, though education and experience were not. Requirements for staff reflect these findings. One of four recent studies on day care, the National Day Care Study helped to resolve a 10 year debate in Congress

over the stringency of regulations, which some states could not meet. One can usually expect a finding to be used that maintains cost can be reduced without affecting quality of service.

Impact Aid

Three options were developed for the future of a program to recompense school districts for the presence of nontaxable Federal property in the community. The HEW Administration selected one set, consisting of five recommendations, as part of the Administration's legislative proposal for ESEA. Although staff said the study was discussed, Congress did not adopt the Administration's proposal. HEW had been trying to cut back on the Impact Aid program for years, because some districts receiving funds were not greatly affected by federal property. Cut-backs, however, are in obvious conflict with members' needs to keep constituents happy.

Indian Education

This study described projects funded by the Indian Education Act. Several officials mentioned that it had not been particularly helpful, merely "listing what was out there." They would have preferred a study of effectiveness, or one that identified the kinds of projects that were useful to Indian students. However, the study was cited in justifying an Administration budget proposal for the program, and possibly, a legislative proposal to increase per-pupil expenditures where low numbers of Indian children affect the success of programs.

DISCUSSION

Evaluations do contribute to policy and program decisions, and to actual changes. We were able to trace numerous contributions to law, regulation, management, and budgets. This is not to say that evaluations alone caused either the decisions or the changes. It is to say that evaluations can be seriously considered, and they can influence the content of decisions. They can sometimes even motivate a decision.

After a decision is made, findings can support arguments in its favor and thus contribute to impact. But many other considerations go into decisions, and many other forces determine program changes. In highlighting some of these contributions, we tried to indicate the forces at work in each situation.

Because our findings contradict the truism that evaluations are not used, we must reassess our methods and state why we should be believed. First, although our sample was purposive, we examined almost half of the eligible studies completed in two years by the Office of Education's evaluation unit. Even if the rest of the studies contributed nothing to decisionmaking (and we know some of them did), the extent to which OE's evaluations contributed would still be a striking disproof of current beliefs. We do not maintain our findings generalize beyond two years of OE studies. The study that samples evaluations across many policy sectors and many years remains to be done. Yet it is now reasonable to believe that such evaluations might contribute something to decisionmaking.

Second, it is always possible our respondents embroidered on the truth, and decisions were not influenced by the evaluations. This may be a serious problem, since we performed the study for Congress. However, we verified the claims of each respondent as extensively as possible, through independent statements and documentation. A conspiracy would have to stretch from Congressional staffers themselves, to independent researchers, to bureaucrats who had varying perspectives and no particular love for evaluation. We discovered only two instances in which respondents flat-out disagreed with each other, and in both cases the likely explanation is a simple mistake. In summary, our methods, though flawed, are adequate to support the contention that these studies contributed to decisions and to impact on policies and programs. We now compare our findings to the literature on knowledge use in policymaking.

HOW ARE EVALUATIONS USED?

Our findings conflict with current beliefs about the contributions of evaluation in important ways. The beliefs are:

(1) Evaluations are rarely used in decisions.

(2) When they are used for decisions they are a minor input in most cases.

(3) They are used largely as window-dressing to legitimate decisions that were already made on other grounds.

(4) When they are used in decisions, the decisions are usually unimportant.

(5) Evaluations' major use is to enlighten policymakers about the nature of policy and programs.

Extent of Use in Decisionmaking

The belief that evaluations are rarely used in decisions comes primarily from the study of Patton and his colleagues (1977). They interviewed the decisionmaker, the evaluator, and the project officer responsible for each of twenty evaluations of health programs. Fourteen of eighteen responding decisionmakers and thirteen of fourteen evaluators said their evaluation was used. However, in no instance did respondents relate specific findings to specific decisions. Evaluations were used as "additional pieces of information in the difficult puzzle of program action" (p. 145). They were used in a context of ongoing decisionmaking, rich with other information.

Like Patton, we found evaluations served as one of many inputs. And the extent of use we discovered was relatively similar—all studies except two were used in decisionmaking. However, we were able to relate specific findings to specific decisions—129 instances and of these, 68 instances in which the decisions were implemented in changing programs and policies. By any measure, contributions to decisions were the rule, not the exception.

Several reasons are evident for the differences between the Patton study and our own. First, Patton et al. were careful to draw a random sample of evaluations, while our sample was purposive. Conceivably, our findings are an overestimate of the extent to which typical evaluations contribute. As we noted, however, the contributions from OE's evaluation unit are still surprisingly frequent. A second difference in the studies involves the sample of respondents. Patton et al., interviewed "the person identified by the project officer as being either the decisionmaker for the program or the person most knowledgeable about the study's impact" (p. 143). In our study, contributions were scattered all over the federal education sector—typically, there was no single decisionmaker who knew every contribution. Moreover, "the decision maker for the program" in Patton's study may have been so far above the day-to-day administrative responsibility for the program that

he or she did not know the whole story on use. This is especially serious, since Rich (1979) found that subordinates do not always cite findings that are the basis for a decision. We interviewed middle-level bureaucrats who had themselves used findings. Had we interviewedsay, the head of the ESAA programs insteead of those who drafted program regulations, we would not have discovered evaluations's contribution to these regulation.

A final reason for the difference in findings is that Patton examined the health sector in 1975 and we examined education in 1980.[4] Evaluation had five more years to establish itself. Also, it has been a prevalent federal activity for a longer time in education than in health services delivery (excepting clinical trials). Many of evaluation's mistakes and subsequent lessons were made in education programs. In addition, many features of OE facilitated use. A stability was evident in the staffing of education offices that many federal health agencies have not achieved. The director of Title I services had been in office ten years, and the evaluation officer, twelve years at the time of our study. Most administrators, whether they liked evaluation or not, knew what to expect and to whom they should talk. Bureaucratic problems of rapid reorganization and failed communication hinder use (Leviton and Hughes, 1981), and are everywhere in federal health programs. Also, many respondents agreed that the Congressional staff who made use of evaluations in education were much more sophisticated on the subject than is usually the case when Congress deals with data (Zweig, 1979). Staff had dealt with the same policy issues through several Congressional cycles and were prepared to integrate new information into their understanding.

Importance of Evaluations as Inputs to Decisions

The Patton et al. study as well as some case studies (Bauman, 1976; Menges, 1978; Millsap, 1978) indicate that when evaluations are used, they are likely to be one of many inputs to thinking about policy and programs. One may therefore ask: how important an input is evaluation relative to other sources of data and other considerations? The studies of evaluation (Florio, Berman, and Goltz, 1979; Fox, 1977) and of social science generally (Caplan et al., 1975; Weiss and Bucuvalas, 1977; Weiss and Weiss, 1981) indicate that federal decisionmakers can value research information highly in some circumstances. But does it carry weight in decisions?

We can only speculate about the ultimate importance of evaluations as inputs. However, the decision contexts give us clues about importance, as do the opinions of respondents. When, for example, the study "Federal Programs Supporting Educational Change" offered new information about implementing change in schools, respondent opinions and citations of the findings in laws to foster such change allow us to make an inference that evaluation was important. In contrast, our respondents said that evaluations played only a minor (though real) part in some administration budget proposals. Importance varied greatly.

Quantifying importance is meaningless in these cases—qualitative understanding is everything. One must ask, "Important in relation to what, the need for program survival or for political visibility?" No. "In relation to gossip, newspaper headlines, and opinions?" Perhaps. Information from several sources is used in combination (Caplan et al. 1975; Rich, 1977), so can we really partition the variance due to evaluation?

Contribution to Decisions Versus Legitimation of Decisions

Knorr (1977) concluded that, although social science is used in making decisions, it is also used to legitimate decisions that are made on other grounds. Weiss (1978) notes that such use is entirely appropriate in gaining political support for a position. We agree and say such use contributes to impact. Yet our task here is to distinguish legitimation from use in decisions. Again, we can only infer the extent to which legitimation was involved. We have several bases for such an inference. Respondents tell us. Also, respondents and documents may indicate that data played a role before the appearance of a formal bill, regulation, or management order. Legitimation becomes much more an issue afterward, as the proposal is challenged in public debates (Bauman, 1976; Mitchell, 1980). Third, selective use of data provides an indication, as in the case of Magnet Schools. Yet even here, there is evidence that positive findings helped resolve a political debate that in turn led to a decision to change the budget. The findings were not just window-dressing.

Thirteen of the twenty-seven instances of "serious consideration" were attempts to gain support for a political position. Such attempts were often involved when evaluations contributed to decisions as well. However, the decisions also took their form from the evaluations, and in no case could we clearly label the instances involving decisions as mere

legitimation. For example, lobbyists for higher education made use of the Campus-Based Aid survey to argue for a change, and the data helped convince Congress about the merits of the case. As Pelz (1978) noted, persuasion and other uses are not neatly compartmentalized. They should all be expected to occur in political debates.

Importance of the Decision

Weiss (1981) suggests that when decisions are made on the basis of evaluations, they are likely to be less important ones that do not threaten political constituencies. Yet one cannot maintain that the decisions to which evaluations contributed here were unimportant. Clearly, no program died because of evaluation; clearly, other considerations went into budget decisions. But the changes in the law and regulation for Title I, bilingual education, and higher education were extremely important in their potential impact on students. Even though a small start at improving education of neglected and delinquent children was not earthshaking, improvement was the all-important and correct goal. Importance is not a matter of budgets and program survival alone. The reader should judge the importance of contributions in terms of incremental political change (Lindblom, 1968).

Direct Versus Indirect Use in Policymaking

Weiss (1977) concluded from past studies of knowledge use that an important function of social science is to provide enlightenment about policies. Findings would "percolate through the consciousness" of policymakers and eventually influence programs in important ways. Given the work of Caplan et al., (1977), and Weiss and Bucuvalas (1977), we expected many instances in which people would tell us that findings confirmed or revised their thinking about programs—serious consideration, but nothing more.

Perhaps our most surprising finding is the number of decision-related uses compared to those involving serious consideration only. Even in the Rich (1977) study, where use in decisions was common, conceptual use was more common. We can say that in our study, serious consideration was involved in all decisions and all impacts. Enlightenment was essential to many of these, as when "Federal Programs Supporting Educational Change" revealed to federal officials the nature of local implementation of innovations. Yet the linkage to action was

much more direct than we were led to expect. For example, in the National Day Care Study, recommendations translated directly into regulations. In the case of bilingual education, the actions of Congress were hardly direct, as they took a step toward redefinition of goals, and then left the hard work to OE. But the implications of the study were relatively straightforward. There was no evidence of a long "percolation through consciousness" in these instances. It is possible that the decisions in retrospect appear more straightforward then they were at the time. Or, that the "percolation" may have occurred in the long years of evaluation prior to these studies. The debates over Day Care and Magnet Schools had been well-articulated for some time, and questions about Title I performance were frequently very specific. In contrast, debates over the Follow Through outcome studies may take years to result in any decisions, and arguments on sex equity were yet to be marshalled at the time of our study.

Summary

We do not maintain that current beliefs about evaluation are completely false. Several examples in our work illustrate attempted legitimation, enlightenment, or unimportant contributions. However, the generalizations may well be false, and impede a careful study of contributions. Evaluations are certainly performed for many reasons (Suchman, 1967), but to the extent that they are "decision driven," we can reasonably expect contributions to decisions.

WHAT KINDS OF EVALUATIONS CONTRIBUTE?

With many examples available to us, we are in a better position to generalize about the kinds of evaluations that can be expected to contribute, and to speculate about the reasons for this. A most striking finding is the prevalence of contributions by implementation information, in all decision contexts. It can best be understood in light of the variation in local responses to federal policies. Our decentralized system has always allowed local governments great freedom in interpreting federal policies, and the extent of local control leads to problems of implementation (Danielson et al., 1977; Pressman and Wildavsky, 1973). Moreover, federal actors are largely insulated from the reality of local programs. This is more true of the powerful decisionmakers than the program managers, because information is selectively passed

upwards in bureaucracies (Downs, 1967). Nathan (1979) makes the point that federal decisionmakers simply do not know what is going on in local projects. It should not therefore be surprising that they would favor the revealing bits of information about local behavior in the face of their attempts at control.

Use of implementation findings by program management is easily understood, given the need to monitor program operation and to plan for the future. Yet contributions to legislation are even more prevalent in our sample. Why should legislators be interested? Part of the answer lies in Congress' need to make programs accountable to the public. Programs are more easily held accountable for operations than for effects. Also, decisions about implementation directly affect constituents. For example, when local districts exclude private schools from Title I services, constituents will complain.

Another reason that implementation findings contribute to legislation is that the implicit theories about program effectiveness assume correct implementation (Pressman and Wildavsky, 1973). When implementation is poor, as in the case of Title I services to neglected and delinquent children, Congress can act. It can also act when insights are obtained about how to improve effectiveness, as in the case of day care requirements. But when a program is found to be ineffective, with little information about improvements, Congress' choices are constrained to elimination, budget changes, or further study. Only the last is easy in the face of constituent pressures.

Outcome information contributed more often than we expected, given this reasoning. Outcomes were cited in two (possibly four) budget changes and an increase in authorization. However, we can most legitimately suspect "window-dressing" in the budget context, since changes might have occurred in any case. Program managers made no use of pure outcome information, except in the ten career education projects. OE had learned that once the lid is off such innovations, they will expand with or without evidence of effectiveness. Selecting effective projects through the JDRP before they gain a constituency is easier than defunding ineffective projects with a large constituency.

Where implementation is linked to outcomes and Congress has clues to program improvement, action is based in part on the findings. Similarly, the bureaucracy can write regulations based on such clues to program effectiveness. The case of nonprofit organizations in desegregation offers an example. The combination of implementation and outcome information is a powerful tool for policymakers just as it is for evaluators.

When used by Congress, federal administration information indicated management problems, many of which contributed to the variation in local practices. It was the NIE Compensatory Education Study that revealed these, and it was able to contribute because it was reported directly to Congress. Congress would not likely have heard about problems from a study supervised by HEW, nor would it have accepted assurances as readily that the program was well-run. When management problems were revealed within HEW, in the case of the Follow Through Exploratory Evaluation, the information was used in organizational development.

In the NIE Compensatory Education Study, the National Day Care Study, and the Impact Aid Study, assessment of federal administration borders on policy analysis. In each case, an assessment of current program operation (evaluation) was married to projections about the future, and to concrete policy alternatives. In these instances, policies were adopted directly from the alternatives set forth, and were developed in concert with the policymakers themselves. Inclusion of policy analysis appears to promote usefulness.

MODELS FOR REFORM?

Some of these studies may serve as models for evaluation reform if we accept that greater use in decisions is a goal of reform. Correct use is one goal, but we should keep the danger of misuse in mind, as well as strive for fairness, quality, and good planning. Fortunately, some of the highly utilized studies satisfy these other desirable goals. For example, the NIE Compensatory Education Study began with a planning period in which all stakeholders were consulted about important evaluation questions. This served the goals of equity as well as greater relevance.

These findings fuel the fires of certain reforms and not others. None of the reports were purely case studies, and are therefore irrelevant to the merits of using such methods. On the other hand, reliance on implementation findings reinforces our appreciation for the importance of understanding the program prior to evaluating outcomes. Some may argue from our data that outcome findings are not used; but this is not the case. Outcomes in isolation from implementation are to be avoided, perhaps, as they contributed largely in the special environment

of the JDRP. However, the power of combining implementation and outcome data for program improvement is evident. It reinforces Cronbach et al.'s (1980) claim that when evaluation is used, it is used in a formative mode.

Finally, our case studies should encourage evaluators by showing that their work does have potential utility. We hope these examples will enlighten debates about evaluation, and ground them in more realism about program and policy settings.

NOTES

1. A characterization of service delivery always accompanied information about recipients in these studies. Therefore we did not distinguish the two types of information. Some critics may also quarrel with our designating the adoption of innovations as outcome information. However, it was the goal of Title IV-C to get innovations adopted.

2. We could, for example, have used the proportion of individual findings from studies that contributed. However, decisionmakers use finding in clusters (Caplan et al., 1975; Rich, 1977), so that this unit may be misleading. Another alternative unit would be the formal actions of policy and programs: sections of law and regulations, management orders, memos, and reports, RFP's, etc. Thus in the example above, a single contribution would have been counted: "an administrative proposal." Since more than one change could occur in a single formal action, however, this unit loses information.

3. Our study ended in late 1980. We have not followed developments since.

4. Patton (personal communication) believes that greater use is to be expected in the 1980's because evaluation is a more familiar tool for the manager and policymaker.

REFERENCES

BAUMAN, P. (1976) "The formulation and evolution of the health maintenance organization policy, 1970-1973." Social Sci. and Medicine 10 (March/April): 129-142.

BORUCH, R. F. and D. S. CORDRAY [eds.] (1980) An Appraisal Educational Program Evaluations: Federal, State and Local Agencies. Washington, DC: U.S. Department of Education. (ED 192 446).

CAPLAN, N., A. MORRISON, and R. STAMBAUGH (1975) The Use of Social Science Knowledge in Policy Decisions at the National Level. Ann Arbor, MI: Institute for Social Research.

COHEN, D. K. and M. S. GARET (1975) "Reforming educational policy with applied social research." Harvard Educ. Rev. 45 (February): 17-41.

COOK, T. D. and W. E. POLLARD (1977) "Guidelines: How to recognize and avoid some common problems of mis-utilization of evaluation findings." Evaluation 4: 161-164.

CRONBACH, L., S. R. AMBRON, S. M. DORNBUSCH, R. D. HESS, R. C. HORNIK, D. C. PHILIPS, D. F. WALKER, and S. S. WEINER (1980) Toward Reform of Program Evaluation. San Francisco: Jossey-Bass.

DANIELSON, M. N., HERSHEY, and BANES (1977) One Nation, So Many Governments: A Ford Foundation Report. Lexington, MA: Lexington Books.

DOWNS, A. (1967) Inside Bureaucracy. Boston: Little, Brown.

DUNN, W. N., I. I. MITROFF and S. J. DEUTSCH (1981) "The obsolescence of evaluation research." Evaluation and Program Planning 4, 3/4: 207-218.

FLORIO, D.,M. M. BEHRMAN, and D. L. GOLTZ (1979) "What do policymakers think of educational evaluation?—Or do they?" Educational Evaluation and Policy Analysis 1: 61-88.

FOX. H. W. (1977) "Evaluation research and Congressional oversight: A case study of a general phenomenon." Presented at the annual meeting of the American Association for the Advancement of Science, Denver, CO, February.

GUBA, E. G. and Y. S. LINCOLN (1981) Effective Evaluation. San Francisco, CA: Jossey-Bass.

HILL, P. (1980) "Evaluating education programs for federal policy makers: Lessons from the NIE Compensatory Education Study," in J. Pincus et al., Educational Evaluation in the Public Policy Setting. Santa Monica, CA: RAND Corp. (R-2502-RC).

KNORR, K. D. (1977) "Policymakers' use of social science knowledge: Symbolic or instrumental?" in C. H. Weiss (ed.) Using Social Science Research in Public Policy Making. Lexington, MA: Lexington Books.

LEVITON, L. C. and R. F. BORUCH (1982) "Usefulness of evaluations: The NIE Compensatory Education Study." In press, J. of Policy Analysis and Management.

LEVITON, L. C. and R. F. BORUCH (1980) "Illustrative case studies," in R. F. Boruch and D. S. Cordray (eds.) An Appraisal of Educational Program Evaluations: Federal, State, and Local Levels. Washington, DC: U.S. Department of Education. (ED 192 446).

LEVITON, L. C. and E.F.X. HUGHES (1981) "A review and synthesis of research on the utilization of evaluations." Evaluation Review 5 (August): 525-548.

LINDBLOM, C. E. (1968) The Policy-Making Process. Englewood Cliffs, NJ: Prentice-Hall.

MENGES, C. C. (1978) Knowledge and Action: The Use of Social Science Evaluation in Decisions on Equal Educational Opportunity, 1970-1973. Washington, DC: National Institute of Education.

MILLSAP, M. A. (1978) "The uses of evaluation in federal regulation writing: A case study." Presented at Evaluation Research Society Convention, Washington, DC.

MITCHELL, D. E. (1980) "Social science impact on legislative decisionmaking: Process and substance." Presented at the American Education Research Association Annual Convention, Boston, April.

NATHAN, R. P. (1979) "Federal grants-in-aid: How are they working in 1978? Presented for Conference on "The city in stress." Sponsored by the Center for Urban Policy Research, Rutgers Univ. March 9.

PATTON, M. Q., P. S. GRIMES, K. M. GUTHRIE, N. J. BRENNAN, B. D. FRENCH, and D. A. BLYTH (1977) "In search of impact: An analysis of the utilization of federal health evaluation research," in C. H. Weiss (ed.) Using Social Research in Public Policy Making. Lexington, MA: Lexington Books.

PELZ, D. C. (1978) "Some expanded perspectives on use of social science in public policy," in J. M. Yinger and S. J. Cutler (eds.) Major Social Issues: A Multidisciplinary View. New York: Macmillan.

PRESSMAN, J. L. and A. WILDAVSKY (1973) Implementation. Berkeley, CA: Univ. of California Press.

RICH, R. F. (1979) "Problem-solving and evaluation research: Unemployment Insurance policy," in R. F. Rich (ed.) Translating Evaluation into Policy. Beverly Hills, CA: Sage.

RICH, R. F. (1977) "Uses of social science information by federal bureaucrats: Knowledge for action versus knowledge for understanding," in C. H. Weiss (ed.) Using Social Research in Public Policy Making. Lexington, MA: Lexington Books.

RIVLIN, A. M. and P. M. TIMPANE [eds.] (1975) Planned Variation in Education. Washington, DC: Brookings Institution.

SAXE, L. and D. KORETZ (1982) "Editors' notes." Making Evaluation Research Useful to Congress. New Directions for Program Evaluation 1982 (June): 1-8.

SUCHMAN, E. A. (1967) Evaluative Research. New York: Russell Sage.

U.S. General Accounting Office. (1975) Follow Through: Lessons Learned from Its Evaluation and the Need to Improve Its Administration. (MWD-75-34). Washington, DC: U.S. General Accounting Office, October 7.

U.S. House of Representatives, Committee on Education and Labor (1979) Education Amendments of 1980 (96th Congress, 1st Session, Report No. 96-520). Washington, DC: U.S. Government Printing Office.

U.S. House of Representatives, Committee on Education and Labor (1978) Report: The Education Amendments of 1978, H.R. 15. (95th Congress, Second Session, Report No. 95-1137). Washington, DC: U.S. Government Printing Office, May 11.

U.S. Office of Education, Office of Evaluation and Dissemination (1980) Annual Evaluation Report, Fiscal Year 1979.

U.S. Office of Education, Office of Evaluation and Dissemination (1979) Annual Evaluation Report, Fiscal Year 1978.

U.S. Senate, Committee on Appropriations (1979) Report on the Departments of Labor and Health, Education and Welfare, and Related Agencies Appropriation Bill, 1980. Washington, DC: U.S. Government Printing Office.

WEISS, C. H. (1981) "Measuring the use of evaluation," in J. A. Ciarlo (ed.) Utilizing Evaluation: Concepts and Measurement Techniques. Beverly Hills, CA: Sage.

WEISS, C. H. (1978) "Improving the linkage between social research and public policy," in L. E. Lynn (ed.) Knowledge and Policy: The Uncertain Connection. Washington, DC: National Research Council.

WEISS, C. H. (1977) "Research for policy's sake: The enlightenment function of social research." Policy Analysis 3 (Fall): 531-545.

WEISS, C. H. (1973) "Where politics and evaluation meet." Evaluation 1, 3: 37-45.

WEISS, C. H. and M. J. BUCUVALAS (1977) "The challenge of social research to decision-making," in C. H. Weiss (ed.) Using Social Science Research in Public Policy Making. Lexington, MA: Lexington Books.

WEISS, J. A. and C. H. WEISS (1981) "Social scientists and decision makers look at the usefulness of mental health research." American Psychologist 36 (August): 837-847.

ZWEIG, F. M. (1979) "The evaluation worldview of Congressional staff," in F. M. Zweig (ed.) Evaluation in Legislation. Beverly Hills, CA: Sage.

34

Conceptualizing Evaluation Use

Implications of Alternative Models of Organizational Decision Making

Jonathan Z. Shapiro

In an unusual instance of consensus among researchers examining a common problem, it is generally acknowledged by those investigating the phenomenon of evaluation utilization that there is precious little of it. Recent reviews, such as those by Thompson and King (1981), Weiss (1982), and Duggan (1983) tend to focus on the nonuse rather than the use of evaluation information by decision makers, since it is a more common situation in evaluation. The increasing salience of what Holley (1979) calls the "tragedy of nonuse" has led to an important shift in the focus of the evaluation literature from the production of evaluation information to its utilization.

The utilization literature in educational evaluation is based upon a "two cultures" perspective on the problem of the use of social science information by policymakers (Rich, 1981). The two cultures notion, developed by C. P. Snow (1962) and modified to apply to policy analysis utilization by Caplan (1979) as the two communities theory, asserts that social science researchers and policymakers live in separate professional communities, between which exists a communication gap. As Rich (1981, p. 6) describes it,

> Among social scientists, the prevailing belief is that empirically grounded knowledge is seriously underutilized in important policy decisions: Social science still accumulates in libraries and impractical retrieval systems rather than policy and government practices. ... Policy makers, however, feel that they can not understand the reports they receive, that the reports do not deal with the immediate problems on their agenda, and that the reports are not sensitive to political and bureaucratic pressure.

One implication of the two communities assumption (that nonuse is due to a communication gap) is that use is optimized by attending to the communica-

From Jonathan Z. Shapiro, "Conceptualizing Evaluation Use: Implications of Alternative Models of Organizational Decision Making," original manuscript.

Author's Note: This paper was presented at the joint meeting of the Evaluation Network and the Evaluation Research Society, Chicago, October 1983.

tion transactions between researcher and decision maker. Rich (1981) argues that the two communities hypothesis is at the core of most studies of utilization. Duggan's (1983) identification of communication factors, nature of the decision maker, nature of the evaluator, nature and technical quality of the results and organizational structure as the primary factors related to utilization supports Rich's contention since all but the last refer to the knowledge transfer aspect of the evaluation process. This perspective is echoed by Thompson and King's (1981, p. 35) observation that

> The literature makes clear that the most critical determinant of evaluation utilization is what Patton (1978) has termed "the personal factor." As Cronbach et al. (1980, p. 6) summarize, "nothing makes a larger difference in the use of evaluations than the personal factor—the interest of officials in learning from the evaluation and the desire of the evaluator to get attention for what he knows."

Thomspon and King (pp. 35–36) go on to cite supporting assertions in the literature from Holley (1979, p. 5), that utilization is usually the result of the relationship between the evaluator and the user more than anything else; from Guskin (1980, p. 45), that use is based more on interpersonal, organizational, and psychological factors than on the actual information itself; and from Gurel (1975, pp. 27–28), that the major barriers to use are the structural constraints and requirements and the interpersonal relationships that characterize the evaluation endeavor.

The position adopted in this paper is that the two communities theory, while important, does not constitute a sufficient explanation of the organizational response to evaluation information; therefore exclusive attention to knowledge transfer factors is not sufficient to explain utilization. The two communities theory attributes nonuse to a communication gap, which implies that closing the gap will lead to utilization. Thus it is presumed that high quality, timely, comprehensible social science information from a nonthreatening evaluator will necessarily be incorporated into a policy decision. The problem with such a presumption is that it ignores the process by which information is processed, interpreted, and acted upon once it enters the organization.

Rich (1981) argues that levels of utilization may best be explained by examining routine bureaucratic and organizational roles and procedures. The set of rules, practices, and traditions may be expressed in terms of formal or informal policy that dictates how officials will produce, process, and apply information. In short, utilization may well be a function of how organizations make decisions, independent of the manner in which an evaluator produced and delivered that information to the organization. While knowledge transfer is important, it most likely constitutes a necessary but not sufficient condition for explaining utilization.

As a necessary but not sufficient condition, the two communities hypothesis can account for nonuse but not for any nonzero level of utilization. Since

nonuse appears to be the modal response to evaluation data, the two communities explanation is often all that is necessary. However, the explanation for any nonzero level of use must include the organizational decision-making process as a relevant factor. As suggested by DeYoung and Conner (1982), knowledge of the decision-making process within the client organization may help an evaluator to design an evaluation such that use is optimized. To explore this notion, four models of organizational decision making are examined with respect to their implications for understanding and maximizing the probability of evaluation utilization. The models are (1) rational choice, (2) bureaucratic politics, (3) organizational processes (Allison, 1971), and (4) cognitive processing (Steinbrunner, 1974).

ALTERNATIVE MODELS OF ORGANIZATIONAL DECISION MAKING

Rational Choice Model

The rational choice model of decision making is at the same time the most well known and least understood of the decision-making models presented in this paper. Confusion arises from the fact that "rationality" has a common language meaning different from its technical definition in the model. Steinbrunner (1974) observes that the common sense notions of rational decisions as best or worthy of approval are not consistent with the positive definition of rationality found in explications of the assumptions underlying the model (see, for example, Riker, 1962; Rikr & Ordeshook, 1973; De Swaan, 1973). Steinbrunner states (p. 27) that

> a decision process meeting the assumptions of the paradigm may or may not produce outcomes more beneficial or more worthy of approval than those achieved by other means, and the question as to whether it does or not is more a matter of investigation than of deductive assertion.

Similarly, Downs (1957, p. 5) notes that, in these types of models,

> the term rational is never applied to an agent's ends but only to his means ... whenever economists refer to a "rational man" they are not designating a man whose thought processes consist exclusively of logical propositions, or a man without predjudices, or a man whose emotions are inoperative. In normal usage all of these could be considered rational men. But the economic definition refers solely to a man who moves toward his goals in a way which, to the best of his knowledge, uses the least possible input of scarce resources per unit of valued output.

The technical definition of the rational decision maker refers to the consistency of behavior or decisions with the individual's preference order rather than with some social definition of what is good, right, or useful. The source of an individual's preference for outcomes is irrelevant; the focus of the model is on the process by which decisions consistent with individual values are gen-

erated. If there is a common sense definition of rationality, it is more closely associated with the notion of self-interest rather than collective interest or social worth.

A second misconception related to the rational choice model concerns the level at which rational decisions are made. The rational choice model is a model of individual decision making. Thus, rationality cannot be conferred upon groups, agencies, or organizations. Aggregate choice can be explained in terms of the constellation of individual preferences (Black, 1971), but the notion of aggregate rational choice as a response to an aggregate utility function does not pertain. In fact, one of the important theorems in the rational choice literature, Arrow's general impossibility theorem (Arrow, 1963), demononstrates that under reasonable democratic conditions, the outcome of aggregating individual rational preferences can be an irrational social choice. In accounting for the behavior of nations in conflict, Allison (1971, p. 252) points out that "the analogy between nations in international politics and a coordinated, intelligent human being is so powerful that we rarely remember that we are reasoning by analogy."

The rational choice model is therefore a depiction of how individuals come to their own decisions. The model presumes that decisions are based on an analysis of the net utility assigned to different outcomes under different states of nature, and decisions are consistent with maximum net utility assignment. Rational individuals assign preferences to outcomes, order their preferences, and act in the direction of the most preferred outcome.

Organizational decisions are therefore a function of individual preference orders and utility functions. Although decisions may maximize the organizational preference order (the set of ordered organizational goals), the decisions themselves are in response to individual preferences. Individually generated, organization maximizing decisions can occur for several reasons. If individual rational decision makers are in agreement—as in Downs's (1957, p. 25) definition of a political party as a coalition whose members agree perfectly on all goals such that a single, consistent preference order can be identified for the coalition—organizational choice will be rational (utility maximizing) choice. Clearly, the larger the organization, the less likely it is to obtain perfect agreement of individual preference orders.

A second situation occurs when the organizational values and choices are both defined by a single, rational individual. In this instance, the organizational preference order and decisions are individual choices elevated to an aggregate level. Peterson (1976, p. 131) observes that rational models of American foreign policymaking are based on the assumption that the president, supported by the Constitution and modern practice, plays a preeminent role.

Peterson (pp. 133–134) suggests a third situation in which the decision would be organization maximizing. He argues that under certain conditions organizations can offer incentives to or impose constraints on individual deci-

sion makers to maximize organizational utility. In each of these instances, however, the organization maximizing decision is coincident with the intention of rational decision makers to maximize their respective, individual preference orders.

Bureaucratic Politics Model

The model of organizational decision making that Allison (1971, p. 144) labels the bureaucratic paradigm portrays organizational decisions as the outcome of internal political bargaining and negotiation. The model rejects the notion of rational analysis by representing decisions as the partisan mutual adjustment of internal conflict (Lindblom, 1965; Braybrooke & Lindblom, 1970). It is assumed in this model that humans are incapable of the full-scale analysis of preferences, outcomes, and decisions described in the rational choice model and that the analysis of choice is limited to achieving mutually acceptable decisions. Such decisions are usually incremental since, the greater the change, the larger the set of alternative values, outcomes, and strategies that must be analyzed. The goal is not to arrive at an individually focused utility maximizing decision but a mutually negotiated one.

Lindblom (1965, p. 3) argues that, in the absence of a coordinated, analytical approach to decision making, "people can coordinate with each other without anyone's coordinating them, without a dominant common purpose and without rules that fully prescribe their relations to each other." An example occurs when two masses of pedestrians cross an intersection against each other because (p. 3), "they will slip through each other, each pedestrian making such threatening, adaptive or deferential moves as will permit him to cross, despite the number of bodies apparently in his way."

In many decision settings, the players have unequal power of resources. However, a major assumption of the process of "muddling through" is that decision makers recognize the efficacy of dyadic negotiation over the individual calculation of utility. In contrast with rational decisions, muddling decisions are neither the simple choice of a unified group nor a formal summary of leaders' preferences. Rather, the context of shared power but separate judgments about important choices means that bargaining and negotiation are the mechanisms of choice. Each player pulls and hauls with the power of his or her discretion for outcomes that will advance his or her interests (Allison, 1971, p. 171).

Organizational Processes Model

A third model of organizational decision making can be distinguished from the first two by the removal of the influence of individual preferences on the decision process. Allison's (1971, p. 67) organizational process model depicts decisions not as the outcome of individual conflict within the organization but rather as being due to the consensus among individuals concerning the need

for stability, consistency, and the continued existence of the organization. The implication is that, either through choice or coercion, individuals act within prescribed organizational roles, ignoring their personal preferences and simply working for "the good of the organization." This should not be confused with the rational choice notion of maximizing organizational utility, because it is assumed in the organizational process model that rational calculations are not possible and the organizational goal is simple survival.

As developed by March and Simon (1958), Cyert and March (1963), and Wildavsky (1964), decisions are viewed within the model as products of organizational routine; they are consistent with the organizational roles and rules designed to ensure the continued existence of the organization. Decisions are made when the organization is required to do so, mainly when the stability and existence of the organization is threatened. All organizational behavior, including decision making, is prescribed by a narrow repertoire of standard operating procedures. The only goal of decision making is the continued existence of the organization, which is done by creating an organizational structure insulated against the hostile environment within which the organization resides.

Cognitive Processing Model

The final model of decision making is Steinbrunner's (1974) notion of cognitive processing. Steinbrunner argues that other models of decision making represent ways in which complex decisions are handled by decision makers. He observes that decision makers are required to exert some form of control over complexity by analyzing it, selectively ignoring most of it, or by insulating against it.

In the face of overwhelming complexity, not amenable to rational calculation or bargaining, the mind functions as a mechanism for resolving ambiguity. In citing Chomsky's model of language acquisition as a process in an environment not sufficiently structured to convey the rules of grammar, Steinbrunner (1974, p. 90) asserts that the inferential capacity of the mind to construct grammar rules reveals its ability to impose structure on highly ambiguous data. Steinbrunner argues that the known principles of cognitive operations suggest a very different response both to uncertainty and utility calculations from that projected by other decision paradigms.

From the cognitive or cybernetic perspective, it is presumed that the process of decision making is dominated by the mechanisms of

(1) grooved thinking, in which all information is categorized in fixed ways,
(2) uncommitted thinking, in which each successive piece of information is taken to be authoritative, and
(3) theoretical thinking, in which information is fit into prexisting abstract and extensive patterns of belief in noncontradictory manner.

It is argued (Steinbrunner, 1974, p. 136) that these inference mechanisms of the mind impose structure in ambiguous (highly complex) situations in systematic ways under given organizational conditions, and that the cognitive decision process operates within the fixed structure thus established. Cognitively processed information, appearing as reality, may be far different from the actual situation. Because cognitive processes are unconscious processes, this would not be perceived by the decision maker.

ON THE NEED FOR ALTERNATIVE MODELS OF DECISION MAKING

Given the intention in this paper to propose the design of evaluations consistent with the implications of decision making, a question may be raised concerning the need for multiple, seemingly mutually exclusive, models of decision making. It is argued in this paper that the four models are not descriptions of alternative realities, but alternative descriptions of a single reality. Based upon Theil's (1971) dictum that models are to be used but not believed, it is asserted that the models can be employed to emphasize different aspects of the process of decision making, but elements of all the models are present to some degree in any decision. This position is in agreement with Allison's (1971) observation that alternative models function as perceptual screens through which different aspects of the decision process can be magnified.

The primary assumption underlying the perspective on decision making employed in this paper is that decision makers evaluate decisions by attempting to calculate the consequences of decisions. The four alternative models identify the levels at which decision consequences can be calculated. The rational choice model represents a method for analyzing consequences for the decision maker alone. The bureaucratic politics model represents the analysis of decision consequences for sets of individuals within an organization. The organizational process model represents the analysis of decision consequences for the organization as a holistic entity. Finally, the cognitive processing model represents the method of decision making when the complexity of the decision setting does not permit the analysis of consequences at any level.

While it is accepted that most decisions contain elements of all the models, it is argued that under different conditions, one of the models will most closely approximate the reality of the decision setting. This condition, as suggested by Steinbrunner, is the degree of complexity in the decision setting. Therefore, the models can be understood to represent approximate behavior under different degrees of complexity in the decision-making setting.

The rational choice model may represent decision making in the face of relatively minimal complexity. When alternatives are few, consequences of action are clear, and preferences for alternative outcomes are distinct, individuals may analyze the decision according to the rational calculus (see, for example, Stratman, 1974). When decisions become more complex, individuals

may choose to analyze only a subset of outcomes and actions and settle for consensually acceptable, rather than utility maximizing, outcomes.

When the decisions become too complex to analyze over any subset with respect to individual costs and benefits, the decision may be arrived at by considering only costs and benefits to the organization qua organization. This simpliflies the decision since organizational preferences are assessed against the relatively simple organizational values of stability, consistency, and continued existence. Finally, when the decision setting is too complex even to be analyzed through organizational routine, decisions will be made through cognitive processing in which complexity will be distorted to fit preexisting belief structures. To assert that the models represent aspects of any decision process is to imply that decision makers attempt to analyze all levels of consequence. To suggest that one of the models is particularly appropriate for a given decision is to identify a particular level of consequence as particularly salient in that decision.

IMPLICATIONS OF THE ALTERNATIVE MODELS FOR EVALUATION USE

The basic premise of this paper is that the alternative models of decision making have implications for study of evaluation use and the design of useful evaluation. For each model, the nature of the decision and characteristics of relevant information will be considered in terms of expectations and strategies available to an evaluator.

The rational choice model assumes individually focused, utility maximizing decision makers. When a single individual is (or a group of like-minded individuals are) responsible for a decision, the decision will be a function of an individual preference order. When a set of dissimilar individuals are responsible for the decision, some aggregation of preference orders is required.

Historically, evaluation researchers have maintained a rational choice notion of the nature of decision making. As Weiss (1972, p. 2) has observed,

> Evaluation research is viewed by its partisans as a way to increase the rationality of decision making. With objective information on the outcomes of programs, wise decisions can be made on budget allocations and program planning. Programs that yield good results will be expanded; those that make poor showings will be abandoned or drastically modified.

Similarly, DeYoung and Conner (1982) argue that most evaluators mistakenly presume that decisions are rational choices. It is too commonly assumed, they state, that decision makers (p. 434)

> have clarified and ordered the goals of the organization and pursue logical strategies for attaining these goals. The rational evaluator believes that the decision maker operates according to an overall plan of action and that he needs relevant information to select sound means to achieve desired ends.

This notion of rationality implies that when a program is measured against its goals, the organizational decision will reflect the evaluation data which, if valid and reliable, will indicate a direction for maximizing those goals. The problem with this definition of rationality is that it is based upon the misconceptions due to common language meaning discussed above. To assert, as Weiss does, that rational decisions are wise decisions is to misunderstand the technical definition of rationality explicated by Steinbrunner and Downs. To state, as DeYoung and Conner do, that rational decision makers order the goals of the organization and act to attain them, is to mistake the level at which rational decisions are made. In short, to assume that rational decisions are wise decisions, and that they reflect an organizational preference order, is to misrepresent the rational choice model.

The problem with this misrepresentation of the rational choice model is that it sets up rationality as a "straw person." That is, if rationality is defined as wise, aggregate focused decision making, and experience reveals that organizational decisions are neither wise nor aggregate utility maximizing, the implication is that the rational choice model is not a reasonable representation of organizational decision making. As a consequence of this straw person definition, evaluators such as Patton (1978, p. 18) have concluded that, "the visions of government based on rational decision making undergirded by scientific truth were beginning to fade."

The implications for evaluation use of correctly interpreting rationality as individual and self-interested, rather than collective and wise, decisions are significantly different from the straw arguments defeated by Patton. When a rational individual comes to a decision, the decision is based upon an individual preference order even when the decisions are coincidentally organization maximizing. In the rational choice model there is no a priori reason for expecting a decision maker to adopt the organizational preference order as his or her individual order. A rational decision maker would make organizational maximizing decisions only if organizational interests are placed above personal interests *in the individual preference order*—an unlikely situation given the nature of self-interest. To assert that a set of rational decision makers will make organization maximizing decisions is necessarily to assert that organizations are made up of rational altruists, which is likely a contradiction in terms.

The point is that rational decision makers would welcome any information that reduces uncertainty concerning the true state of a program, but the decision based upon that information will be a function of self-interest. A rational decision maker who would suffer a loss in utility if a current unsuccessful program would be replaced by some other program would be unlikely to make a decision in which that failure was acknowledged. A decision maker who would be better off if a program failed, perhaps because he or she would be given greater authority in a new program, would make a decision based on acknowledging the program failure. Therefore, the rational choice model suggests that decision makers will utilize valid and reliable evaluation information concerning their programs. However, the form of utilization, particularly

with respect to the direction of the program decision, is a function of individual preference, and organizational goals that are not also personal preferences will not influence a rational decision. Evaluators shoud strive to maximize the reliability and validity of the data they produce for rational decision makers, but utilization, subject to the preferences of the individual decision maker, is beyond the evaluator's control and may not conform to the evaluator's expectation.

The bureaucratic model advanced by Braybrooke and Lindblom (1970) asserts that decision makers analyze only a subset of relevant values and outcomes related to a decision, and they then seek a consensual decision position. Decisions, being the result of bargaining and negotiation, are likely to be based on compromise and result in only incremental change. Since decisions will be incremental, data are likely to be perceived as relevant provided they suggest incremental rather than large-scale directions for change. It is likely, under partisan mutual adjustment, that Weiss's (1972, p. 326) observation that "utilization might be increased if the evaluation included ... analysis of the effectiveness of components of the program, or alternative approaches, rather than all or nothing, go or no-go assessment of the total program" is a correct statement.

This suggests that in the bureaucratic politics decision process the classical field research design that compares a treatment to a control group will not yield relevant information since the finding of no difference would imply a decision too large to be consensually acceptable. Instead, data pointing to incremental changes would be required; for example, an evaluator could assess alternative programs that are incrementally different. Of course, the likelihood of finding statistically significant differences between similar programs is less than the instance when a treatment and "no treatment" (or radically different treatments) are compared. Thus, as designs become more incremental, the probability decreases that an evaluation of group differences based on statistical inference will point toward change. If an evaluation were based upon a treatment/no treatment design, any decision resulting from the data may fall in the direction suggested by the data, but it is likely to be of a lesser magnitude. Finally, if group consensus lies in a direction away from the evaluation results, the results may simply be ignored.

The organizational process model also limits the range of useful evaluation findings. Since decisions are based upon a repertoire of standard operating procedures, evaluation findings can only be utilized if the action implied by the data falls within the set of routine organizational behavior. The evaluation question must be framed in terms of what decisions can be made, rather than what decisions should be made. Such a requirement is likely to restrict the set of evaluation recommendations an organization can act upon, for example, all things being equal, incremental decisions are more likely to be viewed as routine than are large-scale ones. Thus, some of the implications of the bureaucratic politics model will apply here.

However, the organizational process model has further implications for evaluation research. While the rational choice and bureaucratic politics models reveal the problems of assuming the direct utilization of evaluation findings, at least the search for information is based on a desire to clarify the consequences of program decisions. However, using a model that suggests that organizational behavior is a set of routine operations, the motivation for making the decision to seek information must be examined rather than assumed.

In a recent article, Feldman and March (1981) discuss reasons for the organizational search for information consistent with the organizational process model of behavior. It is first asserted that the decision theoretic notion of gathering information to solve problems is not applicable at the organizational level, an observation echoed by Allison's note on the use of rational choice as analogy at the organizational level. Feldman and March then suggest that information gathering has an important symbolic function in organizational behavior. Organizations gather information to take on the appearance of rationality and competence in decision making. It is therefore the gathering rather than analysis of information that organizations usually pursue. Thus, the motivation for information search is the symbolic value it conveys.

Feldman and March argue that despite this motivation, once information searching becomes part of the organizational routine, it can become instrumental. Again, consistent with the assumptions of the model, organizations can utilize information concerning the environment within which it operates. Therefore, Feldman and March suggest that organizations use information as surveillance, scanning the environment for indications of trouble. They further contend, however, that since information can be subject to strategic misrepresentation, decision makers learn routinely to discount much information (p. 177). In brief, organizations routinely gather information as a symbolic gesture and can utilize surveillance information, but will discount information that is open to misrepresentation.

Feldman and March's conclusions suggest the following implications for evaluation. First, organizations may often request information they have no intention of using. What appears to be evaluation use may actually be evaluator use; that is, it is the process rather than the product that the organization values.

Moreover, the observations that organizations can utilize surveillance information but will discount information open to strategic misrepresentation suggest that the most relevant evaluation information will be nonjudgmental information. Obviously, the type of information most susceptible to strategic misrepresentation is information that contains interpretation, such as the evaluator's judgment concerning the success of a program. Consequently, when decision makers are operating out of the organizational process mode, the traditional evaluation inference comparing programs to program goals is least likely to be utilized. Under the organizational process model, evaluators should try to produce surveillance information, that is, the program equiva-

lent of social indicator data. Such relatively nonjudgmental information can then be scanned by the decision maker, who will act upon it when homeostasis is threatened.

The final model of decision making is Steinbrunner's cognitive processing specification. The model asserts that under conditions of complexity too great to be controlled at any level by the decision maker, cognitive processing functions such as grooved thinking, uncommitted thinking, and theoretical thinking drive the decision. The most important implication for evaluation use is that decision makers may unconsciously distort the evaluation findings to make them fit preconceived decision orientations. When this is the case, utilization is clearly beyond the control of the evaluator. However, it should also be noted that if cognitive processing is more likely to occur in the face of greater complexity, then evaluators should strive to produce information in as simple a format as possible in order to minimize the subsequent distortion.

CONCLUSIONS

This paper began with the premise that the two communities theory of utilization is not a sufficient explanation for the utilization of evaluation data. It was asserted that utilization and the design of useful evaluation is also a function of organizational decision-making procedures. It has been further argued that under different conditions of decision setting complexity, the salient aspect of the decision process varies. In order to maximize use, evaluators will have to identify the process and structure the evaluation according to the demands of information relevant to that process.

One important effect of acknowledging the influences of the decision-making process on utilization is to undermine a favored evaluation assumption—that information of high technical quality concerning program performance can be useful. In analyzing the characteristics of useful information under each decision-making model, it becomes clear that usefulness is based on the individual needs of decision makers, not the aggregate needs of the organization. In none of the models, except for a highly unlikely rational choice situation (where decision makers are rational altruists), is there a demand for valid and reliable information concerning the performance of a program against the organization's goals. Thus, the evaluator's goals of producing high-quality information on program performance and producing useful information cannot be simultaneously maximized. It appears that evaluators will have to choose between (a) producing highly useful information by functioning as the personal consultants to individuals within an organization and (b) producing high-quality information about a program (in effect becoming the advocate of the organization rather than individuals within it) and accepting a lower probability that the information will be utilized.

REFERENCES

Allison, G. T. (1971). *Essence of decision.* Boston: Little, Brown.

Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New York: Wiley.

Black, D. (1971). *The theory of committees and elections.* London: Cambridge University Press.

Braybrooke, D., & Lindblom, C. E. (1970). *A strategy of decision.* New York: Free Press.

Caplan, N. S. (1979). The two community theory and knowledge utilization. *American Behavioral Scientist,* **22,** 459–470.

Cronbach, L. J., et al. (1980). *Toward a reform of program evaluation.* San Francisco: Jossey-Bass.

Cyert, R.M., & March, J. G. (1963). *A behavioral theory of the firm.* Englewood Cliffs, NJ: Prentice-Hall.

De Swaan, A. (1973). *Coalition theories and cabinet formations.* San Francisco: Jossey-Bass.

DeYoung, D. J., & Conner, R. F. (1982). Evaluator preconceptions about organizational decision making. *Evaluation Review,* **3,** 431–440.

Downs, A. (1957). Economic theory of democracy. New York: Harper & Bros.

Duggan, J. D. (1983). *Client use of evaluation findings: An examination of salient variables.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April.

Feldman, M. S., & March, J. G. (1981). Information in organizations as symbol and signal. *Administrative Science Quarterly,* **26**(1), 171–186.

Gurel, L. (1975). The human side of evaluating human service programs: Problems and prospects. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills, CA: Sage.

Guskin, A. E. (1980). Knowledge utilization and power in university decision making. In L. A. Braskamp & R. D. Brown (Eds.), *Utilization of evaluation information.* San Francisco: Jossey-Bass.

Holley, F. (1979). *Catch a falling star: Promoting the utilization of research and evaluation findings.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April.

Lindblom, C. E. (1965). *The intelligence of democracy.* New York: Free Press.

March, J. G., & Simon, H. A. (1958). *Organizations.* New York: Wiley.

Patton, M. Q. (1978). *Utilization focused evaluation.* Beverly Hills, CA: Sage.

Peterson, P. E. (1976). *School politics Chicago style.* Chicago: University of Chicago Press.

Rich, R. F. (1981). *Social science information and public policy making.* San Francisco: Jossey-Bass.

Riker, W. H. (1962). *The theory of political coalitions.* Forge Valley, MA: Murray Printing Co.

Riker, C. H., & Ordeshook, P. (1973). *An introduction to positive political theory.* Englewood Cliffs, NJ: Prentice-Hall.

Snow, C. P. (1962). *Science and government.* London: Oxford University Press.

Steinbrunner, J. D. (1974). *The cybernetic theory of decision.* Princeton, NJ: Princeton University Press.

Stratman, W. C. (1974). The calculus of rational choice. *Public Choice,* **18,** 93–105.

Theil, H. (1974). *Principles of econometrics.* New York: Wiley.

Thompson, B., & King, J. A. (1981). *Evaluation utilization: A literature review and research agenda.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Weiss, C. H. (1972). *Evaluation research.* Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C. H. (1982). Measuring the use of evaluation. In E. R. House & Associates (Eds.), *Evaluation studies review annual* (Vol. 7). Beverly Hills, CA: Sage.

Wildavsky, A. (1964). *The politics of the budgetary process.* Boston: Little, Brown.

35

The Evaluation Report

A Weak Link to Policy

Dennis DeLoria and Geraldine Kearse Brookins

As secretary of the U.S. Department of Health, Education, and Welfare (HEW) from 1977 to 1979, Joseph Califano personally requested many of the evaluations that were carried out by the HEW Office of the Inspector General. Among the hundreds of department priorities, issues commanding Califano's direct attention were of greater than usual importance. Following his request, the evaluation staff of the Office of the Inspector General would spend six or eight months gathering data, often traveling to many regional offices and local projects across the country. When data collection and analyses were completed, the inspector general and his staff reported the findings directly to Califano. Califano stipulated that the findings be summarized in a written report not longer than 15 pages and summarized orally in 20 minutes, followed by 40 minutes for his questions. From this brief interchange he decided what action, if any, should result from the months of evaluation.

Some dearly held evaluation practices are called into question when the secretary of a major department permits but 15 pages and 20 minutes for reporting important findings, when evaluation reports about federal programs and policies often are 100 to 300 pages in length. Given this discrepancy, it seems necessary to reexamine their contents and organization. By doing so we may find ways to refocus them to better meet the needs of policy makers such as Califano.

Here we first discuss the work of policy makers and some reasons why evaluation reports tend to be long. We then examine three policy reports to determine their similarities in meeting the needs of policy makers.

From Dennis DeLoria and Geraldine Kearse Brookins, "The Evaluation Report: A Weak Link to Policy," pp. 254–271 in *Learning from Experience: Evaluating Early Childhood Demonstration Programs,"* edited by J. R. Travers and R. J. Light.

Finally, we summarize 10 features that appear to make evaluation reports more useful.

POLICY MAKERS: PEOPLE IN A RUSH

Managers' activities are generally characterized by brevity, variety, and fragmentation, claimed Mintzberg (1973) in a broad review of studies examining the nature of managerial work. He pointed out that managers' jobs are remarkably alike, including senior and middle managers in business, U.S. presidents, government administrators, production supervisors, foremen, and chief executives. He found the brevity of managers' activities surprising: telephone calls averaged 6 minutes, unscheduled meetings averaged 12 minutes, and work sessions averaged 15 minutes. Brevity was also reflected in the treatment of mail. Executives expressed dislike for long memos and skimmed most long reports and periodicals quickly. Most surprising, significant activity was interspersed with the trivial in no particular order. Managers must be prepared to shift moods quickly and frequently.

Mintzberg found strong indications that managers preferred the more active elements of their work: activities that are current, specific, and well defined. Among written communications, they seemed to prefer those dealing with active, concrete, live situations. The managers typically received about 20 periodicals and many reports per week. "Most were skimmed (often at the rate of 2 per minute), and an average of only 1 in 25 elicited a reaction," stated Mintzberg (1973:39). From this it would appear that to be effective, or to be even thoughtfully considered, evaluation reports written for policy makers must make some carefully thought-out concessions to such a frenzy of executive activity.

EVALUATORS: PEOPLE CONCERNED WITH METHODS

Evaluators are typically social scientists, with extensive training in the scientific method. Central to that training is the notion that any statement of evaluation or research findings must be accompanied by a careful description of the precise methods used, so other scientists can replicate them to verify the findings. By training and scientific necessity, evaluators devote a substantial part of most reports to detailed descriptions

of the methods used. Such reports typically follow the classical "dissertation" style, having chapters on background, purpose, hypotheses, subjects, design, measures, data collection, statistical analysis, findings, and discussion. The many variations of this style share one essential characteristic: Their fundamental organization emerges from the scientific method. Practically, this dictates that the overall report format be organized around the methods used, and findings are embedded as a subsection within.

The dissertation-style report may contain facts needed by policy makers, but they are usually fragmented because of the need to respect the conventions of science. For example, the details needed to answer a single policy question may be scattered across several chapters--some in the chapter describing the subjects, some in the discussion of child measure outcomes, some in the discussion of parent measure outcomes, some in the discussion of staff interview outcomes, and some in the chapter presenting overall findings. The burden falls on the policy maker to locate the fragments and piece them together to answer complex questions.

TWO REPORTS ARE NEEDED: ONE SCIENTIFIC, ONE POLICY

The methods-oriented evaluation report is necessary to uphold the conventions of science, but a policy-oriented report seems necessary to reach policy makers. Coleman (1972) elegantly described the relationship. He said that the original policy questions must be translated into questions that can be addressed by the methods of science; at the conclusion of the scientific process the findings must be translated into the world of policy. Viewed in this way, most evaluations stop short of completion if the final report is a conventional, methods-oriented one. Only a rare policy maker would spend the time and effort needed to extract policy information from a methods-oriented report while being bombarded by the dizzying activity described by Mintzberg.

An alternative would be a brief, policy-oriented report that describes concrete action items in language understandable to policy makers. Passages detailing methods used to conduct the evaluation would be removed so the policy maker would not have to sift through them to locate passages with findings of interest. Policy questions and their answers would form the major organiz-

ing theme of the report. The jargon of evaluation would be avoided. Policy makers might well consult such a report in making important decisions--at present a too-rare occurrence.

Three Sample Policy Reports

To explore our hunches we examine three policy reports that embody many of the features needed by policy makers. All three were written to directly inform or influence policy, and they advocate specific policy actions. The authors appear familiar with matters of policy and policy reporting. They are situated differently in relation to the policy makers they attempt to inform: Some work in a federal agency responsible for administering programs, some in a private research consulting firm, and some in a child advocacy group.

The reports are different in important ways. One report presents original data only, another presents findings from other studies only, and one presents some of each. One looks only at the process of implementing a major piece of legislation, another at the effects on children of existing school enrollment practices, and another looks at both program process and effects on children. One project had a budget of more than $7 million, another less than 5 percent of that, and one used existing staff in a federal agency. One was requested by Congress, another by a program administration agency, and one was undertaken solely through private initiative. This diversity makes their similarities even more significant.

Although the three reports have certain exemplary features, they are also not without faults, some of which may be serious. Whatever faults they possess, however, do not detract significantly from the policy-oriented characteristics we are interested in. This paper examines and emphasizes the strengths of these reports, rather than their faults, in the belief that this strategy can more directly contribute to future improvements.

This paper does not attempt to assess the actual policy impacts that these reports have already had, nor does it lay out a sequence of events to increase policy impact. Past experience suggests that policy reports, no matter how well written, will not have much influence without deliberately organized support of one kind or another. Such a topic lies outside the intent of this paper.

Our examination is based on simple inspection rather than quantitative analysis. It should be considered a search for hypotheses to be confirmed, rather than a confirmation itself. To the extent our conclusions appeal to common sense, we consider them sufficient. To orient our examination we looked to the reports for answers to four questions:

1. What policy perspective did the authors adopt?
2. What policy questions did they address?
3. What methods did they use to answer the questions?
4. What format of presentation did they use?

There are many smaller questions buried in each of these; the answers are implicit in the narrative. From this examination has evolved some guidelines that may be of use to others preparing policy reports.

Report 1: *Progress Toward a Free Appropriate Education*

Policy Perspective This report (U.S. Office of Education, 1979) is the first of a series of annual reports to Congress on progress in the implementation of P.L. 94-142, the Education for All Handicapped Children Act of 1975. The act requires reports to be delivered to Congress each January.

The Bureau of Education for the Handicapped (BEH, now located in the U.S. Department of Education), which prepared the report, is the agency responsible for carrying out provisions of the act. This, of course, gives the authors a vested interest in the findings, since their purpose is to report BEH's success or lack of success in implementing the act. Despite the potential for a conflict of interest, the report maintains an objective tone throughout; problems as well as successes in implementation are highlighted. The report does not stress future policy actions, but its discussions of problems often include descriptions of corrective actions initiated by BEH or references to the need for additional money or work.

Although BEH wrote the report mainly for Congress, the authors explicitly kept in mind many others who might use the findings, such as federal administrators in HEW, the Office of Education, and BEH; state directors of special education and state evaluators; leaders of professional

associations and advocacy groups; and members of the academic community (U.S. Office of Education, 1979:77).

The report addresses issues of importance to federal policy by virtue of the source of its mandate, the position of its authors, and its stated audiences. Depending on the nature and seriousness of its findings, the report could influence many kinds of decisions: federal legislative authorizations and appropriations, federal regulations and guidelines, federal program implementation practices, training and technical assistance, and similar state (and local, where appropriate) decisions. Moreover, massive funds are involved for implementing the act. For fiscal 1979 the federal appropriation was $408 million, and the states projected outlays up to 30 times as great, for a possible total of $24 billion nationwide (U.S. Office of Education, 1979:113). The act affects every state and every local school district, involving thousands of educators and millions of children.

Policy Questions Six policy questions are addresssed in the report:

- Are the intended beneficiaries being served?
- In what settings are the beneficaries being served?
- What services are being provided?
- What administrative mechanisms are in place?
- What are the consequences of implementing the act?
- To what extent is the intent of the act being met?

All six are closely tied to the concerns of Congress and the requirements of the act. Their final wording was arrived at by a task force, which invited consultation and review from all persons directly concerned with administration of the act. None of the questions explicitly inquires about the changes in children resulting from implementation of the act; instead, they explore the process of providing required services and whether the intended children are being served.

Each of these questions implies a host of subordinate questions, which are discussed either directly or indirectly in the narrative. For example, under the question "Are the intended beneficiaries being served?" the main issue appears to be "How many eligible children are not being served?" Another subordinate question examines inconsistencies among states in the percentages

of children served and the reasons for the differences. Another asks if only eligible children are being served.

None of the major questions directly mentions costs, although costs are prominently discussed in many of the subordinate questions.

Methodology This report summarizes data from other sources rather than presenting original data. Sixteen sources are cited, although the body of the report says little about the studies or their methods. Readers wishing more information are referred to notes, appendixes, or to the studies themselves; references to them are made mainly through the use of footnotes or credits under tables and figures. By thus removing most discussion of the supporting sources, the full emphasis of the report is place on substantive issues, producing a high ratio of substantive findings to supporting explanation.

The policy questions are stated in general terms, but each section of the report begins by clarifying the intent of its question. The clarifications are taken directly from language in the act or related committee print, and the authors provide additional interpretation when needed. They cite findings from previous studies or court rulings when specific problem areas need to be emphasized. This results in a thorough contextual description for readers, setting clear expectations for the kinds of findings needed to answer the questions. The authors present and discuss data from the appropriate sources. The report often points out discrepancies or conflicting findings and isolates these areas for examination in future studies.

Throughout the report the methodology is subordinated to policy considerations. For example, historical narrative and case examples are interwoven with statistical tabulations for answering a single question. This is an improvement on the frequent practice of grouping statistical results in one part of the report, historical background in another, and case examples in a third; such fragmentation forces the reader into several disconnected sections of the report for partial answers to a single question. The BEH report avoids this problem.

Format The BEH report addresses six policy questions; the questions are used as chapter headings to organize the entire report. This permits the reader to go directly to the questions of interest and find all the needed information in one place.

An executive summary, which can be read in about 15 minutes, provides an overview of the report. A reader wishing to follow up one of the statements in the executive summary can find the corresponding sections of the report fairly easily. Two improvements would have made it even easier to locate them: page references following statements in the summary and a more complete table of contents. Policy-related subheadings are used throughout the report and could easily have been listed in the table of contents.

Most topics in the report are presented in self-contained, well-labeled sections that are readable in 15 minutes or less. This permits rapid access to the authors' conclusions in any area of the report, eliminating the need to sequentially read the report from cover to cover for answers to specific subordinate questions. This vastly improves accessibility of information compared with more traditional evaluation reports and saves much time and work for the reader.

The readability of the report is lower than anticipated, measuring near the "very difficult" score of Flesch's (1949) readability formula. A close look at the language in the report shows that there is just as much jargon as in the typical evaluation report, but with one important difference: The jargon is that of policy makers, not of evaluators. Much of the language derives from the act itself and from related legislative processes; some originates in the discipline of special education; the rest originates in the federal and state processes for implementing the act. Most of this jargon, unlike evaluation jargon, is likely to be familiar to the policy makers who will read the report or its summary. The report could nonetheless benefit from more deliberate use of plain English.

Statistical presentations were kept simple throughout, and graphic displays were used frequently. No special training is required of the reader to interpret the statistical data. Only the most elementary statistics were presented: counts, percentages, ranks, and costs.

Any backup materials that did not directly assist in answering the policy questions were relegated to appendixes or referenced in other sources. Throughout the report, however, sufficient information was included to eliminate almost all need for reference to the appendixes or sources in order to understand the report.

Report 2: *Children at the Center*

Policy Perspective *Children at the Center* (Abt Associates, Inc., 1979) is the final report of the National Day Care Study (NDCS), a large-scale study of the costs and effects of day care. NDCS was initiated in 1974 by the Office of Child Development, now the Administration for Children, Youth, and Families (ACYF). This large-scale research project was designed to "investigate the costs and effects associated with variations of regulatable characteristics of center day care--especially care giver/child ratio, group size, and care givers qualifications" (Abt Associates, Inc., 1979:xxv). These three characteristics are generally considered to be central determinants of quality in center day care and are key factors in state and federal regulations.

One of the central issues of federal policy in subsidized day care is the relationship of day care costs to its effects on children. Undergirding this issue are a number of assumptions regarding the characteristics of center care, the quality of care, and the developmental well-being of children in day care settings. ACYF was particularly committed to the assumption that ". . . developmental well-being and growth of children (could) be fostered in a day care setting" (Abt Associates, Inc., 1979:xxvi). Hence it seems the NDCS was implemented to determine whether federal regulations could be developed to incorporate ACYF's commitment to quality without nullifying the indirect economic benefits that have motivated day care legislation.

Although ACYF was the primary source that influenced the structure of the study, there were also other sources and issues. The Federal Interagency Day Care Requirements lacked empirical evidence to support the assumptions upon which the requirements were based, and this lack to a large degree motivated the structure of the NDCS. There were few data available on a large-scale basis regarding characteristics, such as group size, staff/child ratio, and care giver qualifications, their effects on children, and the relationship of costs to effects--all of which are policy issues. The NDCS combined some of the concerns of ACYF and the needs of the Federal Interagency Day Care Requirements into one study by examining the effectiveness of varying center day care arrangements while taking into consideration such demographic variables as regions, states, socioeconomic groups, etc. At least with respect to center care, it was thought that the results of such a

study could provide essential information for policy reformation regarding standards and regulations.

The report speaks to several policy audiences. It is explicitly addressed to administrators within ACYF and to those preparing the Federal Interagency Day Care Requirements. It is also addressed implicitly to state and local governments that regulate day care licensing, monitoring, and standards. In addition, the report can be viewed as being addressed to Congress, which approves the appropriations for federally funded day care.

Policy Questions In this report, three major policy questions were addressed (Abt Associates, Inc., 1979:13):

- How is the development of preschool children in federally subsidized day care centers affected by variations in staff/child ratio, care-giver qualifications, group size, and other regulatable center characteristics?
- How is the per child cost of federally subsidized, center-based day care affected by variations in staff/child ratio, care-giver qualifications, group size, and other regulatable center characteristics?
- How does the cost-effectiveness of federally subsidized, center-based day care change when adjustments are made in staff/child ratio, care-giver qualifications, group size, and other regulatable center characteristics?

The answers to these questions were intended to play a major role in decisions about current regulations and practices that affect day care centers serving federally subsidized preschool children. Adequate answers require that the policy variables have a direct relationship to the major policy issues and questions. Staff/child ratio and care-giver qualifications were assumed to affect children's cognitive and social development. These two characteristics of day care were also known to have a significant impact on the cost per child of day care. Group size was specified in the Federal Interagency Day Care Requirements and therefore was of interest. Given the variety of issues regarding day care, federal involvement, and regulation, an attempt to deal with more than three major policy questions would have merely diluted the report's policy effectiveness. The policy issues are clearly identified and, notably, so are issues that are not a focus of the study. The authors' disclaimers are significant because they further delimit the research

being considered and restrict the readers' attention in the proper context. By calling attention to issues that are not a focus, the authors demonstrate a recognition that there are other important questions that could be addressed.

Methodology One of the major challenges of a study with national policy significance is the selection of a sample. To this end the evaluators carefully and deliberately selected a sample with appropriate classroom composition, care-giver qualifications, and racial composition. Fifty-seven centers with such diversity were selected within three sites.

Selection of sites was based on four general criteria. These criteria required that the sites have a sufficient number of eligible centers, represent different geographic regions of the country, show different demographic and socioeconomic characteristics, and exhibit regulatory diversity. The actual selection of sites resulted from an analysis that grouped urbanized areas according to measures of socioeconomic status. The analysis yielded six prototypical cities within three regions--South, North, and West. On the basis of feasibility of study implementation, the final choice of sites was Atlanta, Detroit, and Seattle.

In one phase of the study, a quasi experiment was executed to compare three groups of centers: treated high-ratio centers, matched low-ratio centers, and unmatched high-ratio centers. The authors point out that the staff/child ratio was selected for manipulation because of its critical policy relevance. The quasi experiment included only 49 of the centers within the total sample.

Given the policy questions involved, it was important to employ measures of classroom composition and staff qualifications that were reliable and valid. Classroom composition was defined in terms of number of care givers per classroom, group size, and staff/child ratio. These particular variables were measured by both direct observation and schedule-based measures. However, only measures based on direct observation were used in the effects analyses. Information regarding care-giver qualifications was gathered through interviews with care givers. Measures based on direct observation were also used to determine teacher behavior and child behavior. In addition, standardized tests were used to measure the impact of center characteristics on aspects of school

readiness. Parent interviews were also conducted to obtain information on parental involvement and family use of center services. These measures were used primarily to assess quality of care at the centers--the outcomes.

The data were subjected to multivariate statistical analyses, but the findings that link classroom characteristics to measures of quality and measures of costs are correlational. The statistical strengths of the reported relationships are sufficient to be used as significant indicators of both quality and costs. The researchers in the NDCS used methodological procedures that were sophisticated and appropriate to the study's goals and mandate.

Format The authors present the policy-relevant findings at the beginning of the volume, allowing the reader to become aware of the major findings immediately. Policy recommendations, which stem directly from the findings, are concretely stated and provide a contextual framework that encourages the policy maker to consider actual policy decisions. The recommendations are grouped by area, providing the reader with a logical progression. For example, the authors present first the findings for preschool children, then the findings for infants and toddlers. After the findings, the authors recommend regulations and guidelines for both groups. The summary gives suggestions for fiscal policy.

Unlike the authors of many research and evaluation reports, the authors of *Children at the Center* do not assume that all readers are familiar with key terms used in the study and therefore provide a glossary at the beginning of the volume. This feature guards against misinterpretation of terms and results and, hence, of implications on the part of the reader. Since the glossary precedes the executive summary, the reader does not have to turn to a specific section of the volume to determine how the variables were defined in order to place the findings and recommendations within the proper context; thus, time is saved for the policy-making reader.

All information is presented in discrete chunks, each of which represents a whole in itself. Specifically, a reader can glean from the executive summary the major findings regarding day care and federal policy. Or, to gain some insight into the manner in which regulatory language should be constructed, the reader could turn to that section and obtain information in a few minutes.

Just as written information is presented in discrete chunks, most of the data are presented in bivariate tables that are concrete presentations of statistical relationships. This kind of uncomplicated presentation seems more likely to be retained by the reader than are complex multivariate tabular presentations.

Report 3: *Children Out of School in America*

Policy Perspective *Children Out of School in America* (Children's Defense Fund, 1974) is a national comprehensive study of the nonenrollment of school-age children, conducted in 1973 and 1974 by the Children's Defense Fund, a child advocacy organization. Inspired by a similar one conducted by the Massachusetts Task Force on Children Out of School, the study was initiated by the Children's Defense Fund, rather than by any particular federal or state agency. It was principally addressed to HEW's Office for Civil Rights but has wide applicability to other federal agencies, state and local governments, school districts, and parent advocacy groups. The findings are presented in three categories: barriers to attendance, children with special needs and misclassification, and school discipline. Specific recommendations are set forth for the federal government, state and local governments, and parents and children. Inherent in the recommendations is a strong advocacy position. The authors advocate that specific actions take place within the federal government, state and local governments, and among parents and children regarding the exclusion of children from school.

Policy Questions The major issue in this report is the denial of a basic education to any child by schools, by either overt or covert practices and procedures. While the policy questions are not explicit in the report, one can identify at least one major policy question and three subsidiary ones:

- How do exclusionary practices (overt and covert) of schools and school systems affect the education of a significant proportion of school-aged children?
- How does the lack of specific procedures for individual assessment and placement affect the education of all children?
- What is the relationship between school attendance and various school charges for essential educational services and material?

• How are suspensions and other disciplinary actions of school mediated by the race, ethnicity, and socioeconomic status of school-aged children?

The exploration of these questions provided a rich data base for policy makers at the federal, state, and local levels. Indeed, such exploration fostered more specific questions to be answered by a number of agencies within these levels of government. The study also provided a basis for active advocacy for children being excluded from school.

Methodology This report uses both 1970 census data on school nonenrollment and survey data obtained via a questionnaire developed by the Children's Defense Fund. The survey instrument was used to augment the census data as well as to address issues of special policy concern to the researchers. More than 6,500 households were represented in the study. The data were collected in 30 areas of the country within various geographic regions that encompassed 8 states and the District of Columbia. In addition, school principals and superintendents were interviewed about nonenrollment, classification procedures, suspensions, and other disciplinary actions.

The data analyses include frequency counts and percentages, with comparisons being drawn between census data and the Children's Defense Fund data. These comparisons are presented in single, straightforward tables. Descriptions of specific methodological procedures appear in an appendix.

Format The major findings of this study are reported at the beginning of the volume. This allows the reader to immediately become aware of the major issues and the scope of the work that is required to remedy the problems at issue. Most of the information is organized in short chapters that can be read quickly. In the case of longer chapters, the subordinate sections can be read within a short time, facilitating access to particular issues. For example, to understand the ways in which children are misclassified for special programs, the reader could turn to that section in the chapter on exclusion of children with special needs and thereby quickly become familiar with the subject.

The document is written in simple, nontechnical language and is basically organized around the three main issues: barriers to school attendance, exclusion of

children with special needs, and school discipline and its exclusionary impact on students. The role of statistics in minimal; the technical information is placed in appendixes. The interspersal of case history and anectodotal data with survey and census data is a particularly effective mechanism for holding the reader's attention and focusing it on specific issues.

MEETING POLICY MAKERS' NEEDS

These three reports share a few features that set them apart from methods-oriented reports. The similarities are not fully consistent across reports, but for purposes of discussion there appear to be about 10 from which we can learn.

1. The questions addressed are clearly linked to real policy decisions. In each report the principal questions arose from a policy context: debates about day care regulations, progress toward implementation of new legislation, or inequities keeping children out of school. Policy makers and people affected by these issues were directly involved in formulating the questions in each case. They participated in meetings to explore and define the questions, and the questions determined the evaluation methods used.

2. At least some questions in each report consider the costs affecting policy. Nearly all policy decisions involve cost (or other resource) trade-offs, either directly or indirectly. When appropriate cost data are presented in a policy report, its possible influence is greatly increased. The cost data can be obtained in different ways: In the National Day Care Study, cost data were collected concurrently with the process and outcome data; in the BEH report to Congress, cost data were estimated from several outside sources.

3. Policy questions form the central organizing theme of the report. The overall organization of these reports contrasts markedly with methods-oriented reports. A glance at the three tables of contents makes the policy orientation immediately apparent. They list the policy questions examined in a reasonably direct fashion, immediately immersing the reader in the substantive issues. This reflects the fact that each chapter typically discusses a single policy question or a small related subset of questions.

4. The reports describe enough of the policy context to permit informed interpretation without outside sources. All three reports went to great lengths to present readers with broad policy perspectives surrounding specific questions. This permits ready interpretation of the findings by readers who are not already familiar with the policy or decision-making context.

5. Evaluation methodology is played down. Evaluation methods used to answer the questions are scarcely mentioned in the three reports. This is not to say that the studies were not built on solidly crafted methods, for by and large they were; rather, the authors chose not to present details of methodology in these reports, which were intended for policy makers. Quite likely the omission is insignificant, considering the purposes of the three reports, since few policy makers possess the training to interpret technical methods. Moreover, the reports provide adequate references to other sources (often appendixes or other volumes accompanying the report) that detail the methods, so readers who wish to can learn more.

6. Reports begin with a brief summary of essential findings. Usually called an executive summary, it permits readers to quickly learn essential conclusions from the report and to decide which other parts of the report they want to read. It seems important for the summary to be brief (10 pages or less). Brickell et al. (1974) interviewed top-level officials from several government agencies and found they preferred 1- to 10-page reports to longer ones. They commonly requested a short report for themselves and a longer one for their subordinate staff; their subordinate staff in turn requested short reports for themselves and longer reports for their subordinates, and so on down the hierarchy.

7. Backup narrative for the executive summary is "chunked" into easily locatable brief segments throughout the body of the report. The reports are generally organized such that a reader who wants to learn more about something in the executive summary can find the backup narrative easily and read it quickly. Throughout most of the reports, information is organized into self-contained, short chunks. This lets a reader quickly follow up on one or two findings of particular interest, without requiring cover-to-cover reading. Authors can usually assume that none of the policy makers will read their report from cover to cover; rather, they will be selective, reading the executive summary and little else

unless it is of high interest, easy to find, and quick to read. Every incremental improvement in accessibility and readability increases the amount of the report likely to be read by the policy maker and, hence, increases the likelihood of policy impact.

8. Only simple statistics are presented. For the most part, statistical presentations in the four reports included only counts, percentages, ranks, averages, ranges, costs, and bivariate tables or graphs. If complex statistical findings cannot be reduced to these simpler forms, they probably will have little meaning to policy makers. Few of them are trained in advanced statistics, and the elegance of advanced techniques may escape them. Moreover, liberal use of statistics will often obscure other information in the report because of the demands it places on the reader.

9. Where jargon is used, it is the jargon of policy makers, not of evaluators. We thought the three reports would minimize jargon to achieve maximum clarity in presenting findings, but to our surprise they did not--they were cluttered with jargon throughout. In contrast to methods-oriented evaluation reports, however, their jargon was taken from policy makers' language, not evaluators' language. Policy makers are likely to comprehend it easily. The use of policy jargon may even enhance the credibility of these reports for many policy makers, by implying that the evaluators understand issues well enough to become familiar with the appropriate language.

10. Concrete recommendations for action are based on specific findings. The reports encourage policy action by presenting specific recommendations. These recommendations tend to be down to earth and specific, avoiding abstract platitudes. This translation from findings to recommendations not only relieves the reader of the burden of interpretation, but it also helps ensure that the authors' intended interpretation will not be misunderstood. The concreteness of the recommendations coincides with the preferences Mintzberg observed among executives for activities that were specific and well defined.

Our 10 observations are little more than hypotheses at this time, but they begin to provide a framework for distinguishing policy-oriented reports from the methods-oriented reports that underlie them. To the extent they are incorporated in future policy-oriented reports, we feel the policy impact of evaluations will increase, even without the further improvements in methodology that we feel are also needed.

REFERENCES

Abt Associates, Inc.
(1979) Children at the Center: Volume 1, Summary Findings and Their Implications. Cambridge, Mass.: Abt Associates, Inc.

Brickell, H. M., Aslanian, C. B., and Spak, L. J.
(1974) Data for Decisions: An Analysis of Evaluation Data Needed by Decision Makers in Educational Programs. New York: Educational Research Council of America.

Children's Defense Fund
(1974) Children Out of School in America. Washington, D.C.: Children's Defense Fund.

Coleman, J. S.
(1972) Policy Research in the Social Sciences. Morristown, N.J.: General Learning Press.

Flesch, R.
(1949) The Art of Readable Writing. New York: Collier Books.

Mintzberg, H.
(1973) The Nature of Managerial Work. New York: Harper & Row, Publishers.

U.S. Office of Education
(1979) Progress Toward a Free Appropriate Public Education. DHEW Publication No. (E) 79-05003. Bureau of Education for the Handicapped, Office of Education, U.S. Department of Health, Education, and Welfare.

36

Should Imperfect Data Be Used to Guide Public Policy?

Clifford Grobstein

When a scientist offers data that bear on some question of public policy—the health hazard of toxic wastes, for example, or pinpointing a cause of acid rain—how reliable should the data be? Should such data and their interpretation be called upon only when they achieve the level of certainty demanded within science itself? Or are lower levels of certainty significant when the issue is one of protecting public health?

The question arose recently in reaction to the first report of the National Research Council's Committee on Diet, Nutrition and Cancer, which I chaired. Commissioned by the National Cancer Institute, the report summarized a large body of literature at various levels of scientific certainty and suggested guidelines for public action. Among these was one to which a number of critics took exception. It called for efforts to reduce the amount of fat in the average American diet from 40 percent of the total caloric intake to 30 percent. This was offered as a step that might lower the incidence of certain common kinds of cancer. The critics said that the data supporting this guideline were inadequate for any scientific conclusion.

Because of the importance of the specific case and the general issue, it is useful to examine the nature of the data in question. The information came from epidemiological and laboratory studies—the first giving the more direct information on human populations, the second the more scientifically convincing data but on another species.

A number of worldwide epidemiological studies show that the incidence of certain cancers differs among countries with different dietary habits. Correlations between diet and cancer type have also been found within countries where subgroups follow different dietary regimens. The studies used methodologies of varying cogency, from simple statistical correlations between incidences and food consumption to more rigorous procedures such as case control comparisons. The reports are not fully consistent, but there is general agreement that high-fat diets are associated with cancer of the breast, colon and prostate.

From Clifford Grobstein, "Should Imperfect Data Be Used to Guide Public Policy?" *Science 83*, 1983, December, p. 18.

Partly because of these findings, the same relationship has been tested in laboratory animals, especially ones that develop tumors similar to the human tumors thought to be sensitive to fat consumption. Typically such studies involve giving low doses of a known carcinogen followed by diets at several levels of fat. In most instances, the higher the fat intake, the less time before tumors develop, the larger the number of tumors, and the higher the death rate. The fat-cancer link is buttressed by a reasonable hypothesis as to the mechanisms by which fats produce their effect. The three-way convergence of epidemiological and laboratory data with the early indications of a mechanism provides adequate support for a provisional conclusion.

Few scientists operating under the strictest canons of scientific certainty would regard the case as closed. But what about a scientist asked to advise policy makers or public health educators? It is easy enough to call for further research before reaching any conclusions. But in the face of growing evidence that most cancers have a long history from initiation to diagnosis, is it responsible to withhold advice, even if based on imperfect data? What we are eating today may be determining the incidence of cancer 20 to 30 years from now. Should people *not* be told about the possible effects of their diets or other behaviors? If they *should* be told, what level of scientific certainty is required first?

In fundamental science we properly demand incontrovertible evidence, else we would be building on shifting sands. In applied science, however, aggressive application often begins before the requirements for full certainty are satisfied. When it works, this is called imaginative initiative and inspired judgment—and the rewards are high. With respect to cancer prevention, people do not ask for the level of certainty appropriate to pure science. Rather they ask what is the best available scientific information on which to make the soundest judgment. And the time for decision often is *now*. Given the level of public concern about cancer and the apparent delay between cause and effect, it would be just as unfortunate for the scientific community to be too late as too early in making people aware of trends that are developing in scientific data.

In general terms, therefore, when working in the policy mode, scientists must recognize that the declared purpose is an important determinant of the necessary level of certainty. In all cases it is essential to communicate accurately what the level of certainty is, as well as how it can be improved. But if science is to be used as constructively as it must be, the rigid criteria of fundamental science are often inappropriate. What often is needed is the best *available* advice for a complex decision arena. Soundly assessed and accurately communicated, the current state of knowledge can be a most important guide, even though not fully complete and not yet wrapped up in the golden trappings of complete certainty. We would be remiss to withhold what can be useful because it is not perfect.

37

Science, Risk and Public Policy

William D. Ruckelshaus

Summary. A climate of fear now dominates the discussion of environmental issues. The scientific community can help alleviate this fear by making a greater effort to explain to the public the uncertainties involved in estimates of risk. Current statutory mandates designed to protect public health both demand levels of protection that technology cannot achieve and are uncoordinated across government agencies. A common statutory framework for dealing with environmental risks is needed. In addition, care must be taken to separate the scientific process of assessing risk from the use of such assessments, together with economic and policy considerations, in the management of risks through regulatory action.

We are now in a troubled and emotional period for pollution control; many communities are gripped by something approaching panic and the public discussion is dominated by personalities rather than substance. It is not important to assign blame for this. I appreciate that people are worried about public health and about economic survival, and legitimately so, but we must all reject the emotionalism that surrounds the current discourse and rescue ourselves from the paralysis of honest public policy that it breeds.

I believe that part of the solution to our distress lies with the idea that disciplined minds can grapple with ignorance and sometimes win: the idea of science. We will not recover our equilibrium without a concerted effort to more effectively engage the scientific community. Frankly, we are not going to be able to emerge from our current troubles without a much improved level of public confidence. The polls show that scientists have more credibility than lawyers or businessmen or politicians, and I am all three of those. I need the help of scientists.

This is not a naive plea for science to save us from ourselves. Somehow, our democratic technological society must resolve the dissonance between science and the creation of public policy. Nowhere is this more troublesome than in the formal assessment of risk—the estimation of the association between exposure to a substance and the incidence of some disease, based on scientific data.

Science and the Law at EPA

Here is how the problem emerges at the Environmental Protection Agency. EPA is an instrument of public policy, whose mission is to protect the public health and the environment in the manner laid down by its statutes. That manner is to set standards and enforce them, and our enforcement powers are strong and pervasive. But the standards we set, whether technology- or health-related, must have a sound scientific base.

Science and the law are thus partners at EPA, but uneasy partners. The main reason for the uneasiness lies, I think, in the conflict between the way science really works and the public's thirst for certitude that is written into EPA's laws. Science thrives on uncertainty. The best young scientists flock into fields where great questions have been asked but nothing is known. The greatest triumph of a scientist is the crucial experiment that shatters the certainties of the past and opens up rich new pastures of ignorance.

But EPA's laws often assume, indeed demand, a certainty of protection greater than science can provide with the current state of knowledge. The laws do no more than reflect what the public believes and what it often hears from people with scientific credentials on the 6 o'clock news. The public thinks we know what all the bad pollutants are, precisely what adverse health or environmental effects they cause, how to measure them exactly and control them absolutely. Of course, the public and sometimes the law are wrong, but not all wrong. We do know a great deal about some pollutants and we have controlled them effectively by using the tools of the Clean Air Act and the Clean Water Act. These are the pollutants for which the scientific community can set safe levels and margins of safety for sensitive populations. If this were the case for all pollutants, we could breathe more easily (in both senses of the phrase); but it is not so.

William D. Ruckelshaus is Administrator of the U.S. Environmental Protection Agency, Washington, D.C. 20460. This article is based on a talk he gave at the National Academy of Sciences, Washington, D.C., on 22 June 1983.

From William D. Ruckelshaus, "Science, Risk, and Public Policy," *Science*, 1983, 221, 1026–1028.

More than 10 years ago, EPA had the Clean Air Act, the Clean Water Act, a solid waste law, a pesticide law, and laws to control radiation and noise. Yet to come were the myriad of laws to control toxic substances from their manufacture to their disposal—but that they would be passed was obvious even then.

When I departed EPA a decade ago, the struggle over whether the federal government was to have a major role in protecting our health, safety, and environment was ended. The American people had spoken. The laws had been passed; the regulations were being written. The only remaining question was whether the statutory framework we had created made sense or whether, over time, we would adjust it.

Scientific Realities

Ten years ago I thought I knew the answer to that question as well. I believed it would become apparent to all that we could virtually eliminate the risks we call pollution if we wanted to spend enough money. When it also became apparent that enough money for all the pollutants was a lot of money, I came to believe that we would begin examining the risks very carefully and structure a system that would force us to balance our desire to eliminate pollution against the costs of its control. This would entail some adjustment of the laws, but not all that much, and it would happen by about 1976. I was wrong.

This time around as administrator of EPA, I am determined to improve our country's ability to cope with the risk of pollutants over where I left it 10 years ago. It will not be easy, because we must now deal with a class of pollutants for which it is difficult, if not impossible, to establish a safe level. These pollutants interfere with genetic processes and are associated with the diseases we fear most: cancer and reproductive disorders, including birth defects. The scientific consensus is that any exposure, however small, to a genetically active substance embodies some risk of an effect. Since these substances are widespread in the environment, and since we can detect them down to very low levels, we must assume that life now takes place in a minefield of risks from hundreds, perhaps thousands, of substances. We can no longer tell the public that they have an adequate margin of safety.

This worries all of us, and it should. But when we examine the premises on which such estimates of risk are based, we find a confusing picture. In assessing a suspected carcinogen, for example, there are uncertainties at every point where an assumption must be made: in calculating exposure; in extrapolating from high doses where we have seen an effect to the low doses typical of environmental pollution; in what we may expect when humans are subjected to much lower doses of a substance that, when given in high doses, caused tumors in laboratory animals; and finally, in the very mechanisms by which we suppose the disease to work.

One thing we clearly need to do is ensure that our laws reflect these scientific realities. The administrator of EPA should not be forced to represent that a margin of safety exists for a specific substance at a specific level of exposure where none can be scientifically established. This is particularly true where the inability to so represent forces the cessation of all use of a substance without any further evaluation.

Functions of Regulatory Agencies

It is my strong belief that where EPA, OSHA (the Occupational Safety and Health Administration), or any other social regulatory agency is charged with protecting public health, safety, or the environment, we should be given, to the extent possible, a common statutory formula for accomplishing our tasks. This statutory formula may well weigh public health very heavily, as the American people certainly do.

The formula should be as precise as possible and should include a responsibility for assessing the risk and weighing it, not only against the benefits of continued use of the substance under examination, but against the risks associated with substitute substances and the risks associated with the transfer of the substance from one environmental medium to another through pollution control practices. I recognize that legislative change in the current climate is difficult. It is up to those of us who seek change to make the case for its advisability.

But my purpose here is not to plead for statutory change; it is to speak of risk assessment and risk management and the role of science in both. It is important to distinguish these two essential functions, and I rely here on a recent National Academy of Sciences report on the management of risk in the federal government. Scientists assess a risk to find out what the problems are. The process of deciding what to do about the problems is risk management. The second procedure involves a much broader array of disciplines and is aimed toward a decision about control.

In risk management it is assumed that we have assessed the health risks of a suspect chemical. We must then factor in its benefits, the costs of the various methods available for its control, and the statutory framework for decision. The NAS report recommends that these two functions—risk assessment and risk management—be separated as much as possible within a regulatory agency. This is what we now do at EPA and it makes sense.

Risk Assessment

We also need to strengthen our risk assessment capabilities. We need more research on the health effects of the substances we regulate. I intend to do everything in my power to make clear the importance of this scientific analysis at EPA. Given the necessity of acting in the face of enormous scientific uncertainties, it is more important than ever that our scientific analysis be rigorous and the quality of our data be high. We must take great pains not to mislead people about the risks to their health. We can help to avoid confusion be ensuring both the quality of our science and the clarity of our language in explaining hazards.

I intend to allocate some of EPA's increased resources to pursuing these ends. Our 1984 request contains significant increases for risk assessment and associated work. We have requested $31 million in supplemental appropriations for research and development, and I expect that risk assessment will be more strongly supported as a result of this increase as well.

I would also like to revitalize our long-term research program to develop a base for more adequately protecting the public health from toxic pollutants. I will be asking the outside scientific community for advice on how best to focus those research efforts.

In the future, this being an imperfect world, the rigor and thoroughness of our risk analyses will undoubtedly be affected by many factors, including the toxicity of the substances examined, the populations exposed, the pressure of the regulatory timetable, and the resources available. Despite these often conflicting pressures, risk assessment at EPA must be based only on scientific evidence and scientific consensus. Nothing will erode

public confidence faster than the suspicion that policy considerations have been allowed to influence the assessment of risk.

Risk Management

Although there is an objective way to assess risk, there is, of course, no purely objective way to manage it, nor can we ignore the subjective perception of risk in the ultimate management of a particular substance. To do so would be to place too much credence in our objective data and ignore the possibility that occasionally one's intuition is right. No amount of data is a substitute for judgment.

Further, we must search for ways to describe risk in terms that the average citizen can comprehend. Telling a family that lives close to a manufacturing facility that no further controls on the plant's emissions are needed because, according to our linear model, their risk is only 10^{-6}, is not very reassuring. We need to describe the suspect substances as clearly as possible, tell people what the known or suspected health problems are, and help them compare that risk to those with which they are more familiar.

To effectively manage the risk, we must seek new ways to involve the public in the decision-making process. Whether we believe in participatory democracy or not, it is a part of our social regulatory fabric. Rather than praise or lament it, we should seek more imaginative ways to involve the various segments of the public affected by the substance at issue. They need to become involved early, and they need to be informed if their participation is to be meaningful. We will be searching for ways to make our participatory process work better.

For this to happen, scientists must be willing to take a larger role in explaining the risks to the public—including the uncertainties inherent in any risk assessment. Shouldering this burden is the responsibility of all scientists, not just those with a particular policy end in mind. In fact, all scientists should make clear when they are speaking as scientists, ex cathedra, and when they are recommending policy they believe should flow from scientific information. What we need to hear more of from scientists is science. I am going to try to provide avenues at EPA for scientists to become more involved in the public dialog in which scientific problems are described.

Lest anyone misunderstand, I am not suggesting that all the elements of managing risk can be reduced to a neat mathematical formula. Going through a disciplined approach can help to organize our thoughts so that we include all the elements that should be weighed. We will build up a set of precedents that will be useful for later decision-making and will provide more predictable outcomes for any social regulatory programs we adopt.

In a society in which democratic principles dominate, the perceptions of the public must be weighed. Instead of objective and subjective risks, the experts sometimes refer to "real" and "imaginary" risks. There is a certain arrogance in this—an elitism that has ill served us in the past. Rather than decry the ignorance of the public and seek to ignore their concerns, our governmental processes must accommodate the will of the people and recognize its occasional wisdom. As Thomas Jefferson observed, "If we think [the people] not enlightened enough to exercise their control with a wholesome discretion, the remedy is not to take it from them, but to inform their discretion."

Interagency and International Coordination

Up to this point I have been suggesting how risks should be assessed and managed in EPA. Much needs to be done to coordinate the various EPA programs to ensure a consistent approach. I have established a task force with that charter.

I further believe we should make uniform the way in which we manage risk across the federal regulatory agencies. The public interest is not served by two federal agencies taking diametrically opposed positions on the health risks of a toxic substance and then arguing about it in the press. We should be able to coordinate our risk assessment procedures across all federal agencies. The risk management strategies that flow from that assessment may indeed differ, depending on each agency's statutory mandate or the judgment of the ultimate decision-maker.

But even at the management stage there is no reason why the approaches cannot be coordinated to achieve the goal of risk avoidance or minimization with the least societal disruption possible. I have been exploring with the White House and the Office of Management and Budget the possibility of effecting better intragovernmental coordination of the way in which we assess and manage risk.

To push this one step further, I believe it is in our nation's best interest to share our knowledge of risks and our approach to managing them with the other developed nations of the world. The environmental movement has taught us the interdependence of the world's ecosystems. In coping with the legitimate concerns raised by environmentalists, we must not forget that we cope in a world with interdependent economies. If our approach to the management of risk is not sufficiently in harmony with those of the other developed nations, we could save our health and risk our economy. I do not believe we need to abandon either, but to ensure that it does not happen, we need to work hard to share scientific data and understand how to harmonize our management techniques with those of our sister nations.

In sum, my goal is a government-wide process for assessing and managing environmental risks. Achieving this will take cooperation and goodwill within EPA, among Executive Branch agencies, and between Congress and the Administration, a state of affairs that may partake of the miraculous. Still, it is worth trying, and the effort is worth the wholehearted support of the scientific community. I believe such an effort touches on the maintenance of our current society, in which a democratic polity is grounded in a high-technology industrial civilization. Without a much more successful way of handling the risks associated with the creations of science, I fear we will have set up for ourselves a grim and unnecessary choice between the fruits of advanced technology and the blessings of democracy.

38

Synopsis from Program Evaluation: 1983 Report of the Auditor General of Canada

The Government's Response

Development of Program Evaluation

3.1 In the 1960s, the awareness that program effectiveness lies at the heart of good public sector management led increasingly to a recognition that formal procedures to measure the effectiveness of public programs were necessary. This happened for a variety of reasons, but two are noteworthy. First, the state of the art in social research had advanced to the point where useful measurement of attainment of program objectives was possible. Second, a growing concern for value for money in complex and expensive public programs led to a rising demand for proof that the expenditures were cost-effective.

3.2 The Government of Canada began to place a growing emphasis on program evaluation in the late 1960s when departments and agencies were encouraged to establish central Planning and Evaluation Divisions by having Treasury Board make resources and person-years available to them. The departments responded. As a 1974 Treasury Board study noted, they had established planning and evaluation units involving approximately 3,500 person-years. However, that study also found that little program evaluation was being done. Most of the effort was apparently being directed to policy analysis and planning.

3.3 In 1977-78, we conducted a government-wide study of program evaluation, looking at 23 programs in 18 departments and agencies. We also found few successful attempts to conduct program evaluations. During that audit, the Government issued a Policy Circular (Treasury Board, 1977-47) which required all departments and agencies to establish procedures to evaluate systematically the efficiency and effectiveness of their programs.

3.4 In 1978, the Office of the Comptroller General was established and given functional responsibility for implementing this new policy. Since then, it has developed a policy framework to guide and structure departmental evaluation functions. This has been published in two documents: *Guide on the Program Evaluation Function* (May 1981); and *Principles for the Evaluation of Programs* (September 1981). By establishing a team of liaison officers to give guidance and advice to departments, the Office of the Comptroller General has also taken an active role in encouraging and assisting departments to implement the policy framework.

Editors' Note: This paper is a summary of a recent review of the status of program evaluation activities in the Canadian government. The review was conducted by the Auditor General's Office, an independent unit similar to the U.S. General Accounting Office, responsible for government-wide oversight. The Auditor General reports directly to the Canadian House of Commons. The analysis paid special attention to the results of actions by the Office of the Comptroller General to establish evaluation units and to promote evaluation activities throughout the Canadian government. The Comptroller General's Office, established in 1978 specifically to improve management practices in the Canadian government, is part of the Treasury Board, a cabinet ministry in the executive branch equivalent to the U.S. Office of Management and Budget. Following the synopsis of the Auditor General's report, we have included the "Government's Response" which was prepared by the Comptroller General.

From *Program Evaluation: Report of the Auditor General, 1983*, pp. 1–8. Published by the government of Canada.

3.5 It is not easy to establish a new function like program evaluation on a consistent basis throughout government. In this case, the task was made more difficult because many senior managers were sceptical about the value of program evaluation, and there was a shortage of qualified program evaluators. To deal with this, the Office of the Comptroller General set as its first target getting evaluation units established in the larger departments and agencies and encouraging them initially to attempt useful program evaluations of small programs, or those of limited scope, to gain experience and to enhance credibility with management.

Audit Focus

3.6 The purpose of this audit was to report on the progress made in establishing program evaluation in the federal government over the five years since our 1978 Report. To do this, we selected a sample of 19 departments and agencies across government and audited their program evaluation functions. In each case, we looked at two areas:

> the infrastructure for program evaluation, including its policy, plans, resources and management, to assess the degree to which the organizational elements necessary to maintain a functioning and productive unit were in place;
>
> the conduct, reporting and use of program evaluations to assess the degree to which they have been carried out in accordance with government guidelines, reported in a balanced and fair way to the appropriate officials, and used by them.

3.7 Our audit criteria were developed from two sources: The Report of the Public Accounts Committee, tabled in the House in July 1980, in which it endorsed the five basic criteria set by our Office for auditing evaluations; and the Office of the Comptroller General's policy framework which was used to provide a more detailed elaboration of these criteria.

3.8 Our observations are divided into three parts. The first two deal with the findings of our audit of program evaluation in 19 departments and agencies. The last section deals with matters of a government-wide nature affecting program evaluation.

Organization and Management for Program Evaluation

3.9 In auditing the development of program evaluation functions in 19 departments and agencies, we found that the government has made significant progress in establishing a program evaluation function in its departments and agencies. All the departments and agencies we audited have established corporate program evaluation units. In 1982-83, the 19 units used 168 person-years of staff time and spent just over $3 million contracting with outside consultants for professional and special services. Only two of the departments did not have deputy-approved policies for program evaluation, although one of these has subsequently been approved.

3.10 To ensure that program evaluations are carried out systematically and on a cyclical basis, departments are required to develop a long-term plan to evaluate all their

programs. Sixteen of the 19 departments had approved long-term evaluation plans, 2 had draft plans and 1 had no plan.

Editors' Note: "Tabling" in the Canadian governmental system has an opposite effect to that of tabling in the U.S. system. In Canada, tabling is a formal action whereby a report is presented to the House of Commons (the main Canadian legislative body) and so is available for governmental and public review and action. In the U.S. system, by contrast, tabling is an action that effectively stops any further consideration of an issue.

3.11 There were a number of areas in which the management practices of the evaluation units could be improved. In particular, these concerned the control of projects, both with respect to timing and resources. Further, in many units, adequate documentation for studies was not maintained, nor was information available to some managers to enable them to account for resources used on evaluations.

Program Evaluation Initiatives

3.12 By contrast with our findings in 1978, we found that departments have made considerable progress in planning and carrying out corporate program evaluations. Seventeen of the 19 departments have evaluation initiatives under way, and 15 of these have completed at least one program evaluation study. Many have completed more. Overall, we found that 6 evaluation frameworks, 119 evaluation assessments, and 86 program evaluation studies had been completed.

3.13 In those departments with relatively more experience and in which we were able to identify a trend, we observed improvements in the quality of evaluations undertaken. However, we also identified a number of areas where further improvements are required. A significant proportion of evaluation assessments did not form an adequate basis for sound advice to the deputy for planning the evaluation study. In a number of cases, limitations in the study could be traced back to weaknesses in the assessment.

3.14 A substantial number of the evaluation studies had significant weaknesses in the methods used to carry out planned work, particularly with respect to measuring program effectiveness. We found poorly designed questionnaires; unreliable data; data that were incomplete and/or lacked objectivity; inadequately specified sample designs; and samples that were too small for the purpose intended, or biased. As a result of these problems, approximately half the studies which attempted to measure the effectiveness of programs were unable to adequately attribute outcomes to activities.

3.15 It is important to note that most study reports included discussions of qualifications of the findings, although, in some instances, these were not as complete as they should have been.

3.16 Even in those cases where difficulties were encountered in planning and carrying out assessments and studies, it was often the case that parts of these produced information that is sound and that departmental managers found useful.

3.17 The 43 studies that we audited in depth covered a wide range of topics, and most addressed at least one of the four basic program evaluation issues:

over three-quarters attempted to address program rationale, raising questions about the continuing relevance of, or need for, the program;

three-quarters attempted to measure the degree to which the program had achieved one or more of its goals and objectives;

a similar proportion tried to determine the extent to which the observed changes had occurred as a result of the program, and

half attempted to assess the degree to which there were cost-effective alternatives to the program.

Editors' Note: The deputy is the highest career government employee in a department or agency. The minister is the appointed head of one or several departments and is a member of Parliament.

3.18 The recommendations arising from these studies covered an equally wide range. They dealt with matters such as changing program support activities, program design, and the size of the program. In most instances these recommendations have been accepted by senior management and acted upon.

3.19 We found that, in almost all instances, evaluation reporting has followed procedures set out in the departmental evaluation policy. All departmental policies require reporting to the deputy, and nearly all require reporting to other responsible levels of program management. Instances of failure to report evaluations have been infrequent, and steps have been taken to ensure appropriate reporting. However, we also found that the reporting of studies in the Strategic Overviews under the Government's new Policy and Expenditure Management System could be improved.

Government-wide Issues

3.20 We used the program evaluation guidelines issued by the Office of the Comptroller General as a basis for expanding and elaborating on the audit criteria developed by our Office for our 1978 study and endorsed by the Public Accounts Committee in 1980. We believe that these guidelines provide a useful basis for departments to organize, plan and conduct program evaluation work. Also, as noted earlier, the implementation strategy followed by the Office of the Comptroller General involved an initial emphasis on establishing evaluation units in the major departments and agencies and on encouraging these departments to conduct evaluations, even of a limited nature. In our opinion, this has been a reasonable way in which to proceed. However, we noted several areas where the policy framework may need to be modified or extended.

3.21 The current policy framework for program evaluation has the department and the deputy head as the central focus of program evaluation activity. However, many important programs are delivered in a way that involves more than one department. Further, the basic structure of the Policy and Expenditure Management System stresses the interdepartmental nature of program decision making.

3.22 Although current policy and guidelines recognize the existence of interdepartmental programs, they fail to specify procedures to be followed in conducting evaluations of them. The consequence of this is that interdepartmental programs are not systematically being subjected to the same type of orderly review and evaluation as programs administered wholly within single departments and agencies.

3.23 Most of the agencies which had not developed an infrastructure for evaluation were small. As part of its implementation strategy, the Office of the Comptroller General has only recently given full attention to the way in which evaluations of these agencies' programs should be carried out.

3.24 Crown corporations are being used by the Government to achieve public policy objectives, and funds for these purposes are being provided through the Estimates. We found that, in general, public policy objectives of Crown corporations were not subject to program evaluation, nor were they being scheduled for program evaluation. As of 31 March 1983, five Crown corporations were establishing program evaluation functions in liaison with the Office of the Comptroller General.

3.25 In its review of the 1978 Auditor General's Report, the Public Accounts Committee recommended to the House, in July 1980, that all effectiveness evaluations be tabled in the House of Commons within 60 days of their completion. The Government responded, in October 1980, through the President of the Treasury Board, that:

> With enactment of the Access to Information Bill as currently proposed, and approval of a draft Treasury Board policy on the documentation of program evaluation studies, such information would be publicly available. Some procedure, perhaps tabling of evaluation reports within 60 to 90 days of their completion by the responsible Minister, will have to be established to ensure easy access by Members of Parliament.

3.26 On the basis of our audit work we found that, since 1980, only one program evaluation report has been tabled in the House.

Summary

3.27 While we were sharply critical of the situation we observed in 1978, we also stated that "the recent emphasis the Government of Canada has placed on program evaluation could, if developed and sustained, put it among the leaders in the field." Leadership does not come easily. The task which the Government has set for itself is to make program evaluation an integral part of public sector management. This requires no less than a commitment on the part of the Government to critically examine the success of its own programs and policies.

Editors' Note: Crown corporations are publicly owned corporations. The "Estimates" is the proposed federal budget of Canada.

3.28 We found that very real progress has been made. Most of the major departments and many of the agencies have the basic infrastructure for evaluation in place. Many of these are actively attempting to carry out evaluations. However, the quality of the evaluations needs to be improved.

3.29 To achieve these improvements in quality, the government's commitment must be made evident through requests for the evaluation of its programs, appropriate staffing of departmental evaluation branches, and the distribution of evaluation reports.

3.30 Program evaluation requires asking fundamental questions about a program's existence. Requests for evaluations should make explicit what the program is trying to achieve and against what it should be evaluated.

3.31 In getting the results achieved thus far, one of the major problems the Office of the Comptroller General and departments have faced has been a shortage of appropriately trained and qualified evaluators. This shortage continues, and if the progress and momentum achieved thus far are to be maintained, the development of a cadre of qualified professional program evaluators will be one of the major challenges to be met.

3.32 Finally, the quality of evaluations should improve with experience. This process requires that evaluation reports be widely distributed. Broader scrutiny will help ensure higher quality products. It will also help to ensure that lessons learned in one study will be available to build on in the next.

The Government's Response to the Auditor General's 1983 Audit Report on Program Evaluation

The Government's approach to evaluation of programs aims at ensuring that relevant and reliable information is produced on the continuing need for, performance of, and relative cost-effectiveness of programs. In its approach to the establishment of the program evaluation function, the Government has been mindful of the need to ensure procedures for the conduct and consideration of evaluations which enhance their credibility and actual use, as well as being mindful of the significant cost of conducting studies. Accordingly, the approach taken to establishing the program evaluation function is one of integrating evaluation within the improving management practices in departments and with the Policy and Expenditure Management System, while the approach to conducting evaluations is one which balances the need for objectivity with the need to link evaluation closely with line management of government programs.

This audit, coming midway through the planned initial development phase, is a timely progress report on the approach the Government has taken, the accomplishments to date and the main work yet to be done. The findings, taken as a whole, represent to the Government an endorsement of its approach and confirm the substantial progress achieved. The recommendations, in general, are helpful and supportive of the Government's efforts and, with a few exceptions, outline a program of work largely consistent with the Government's plans in this area.

The infrastructure in terms of evaluation policies, plans and organizations is in place in most of the major departments and agencies throughout the Federal Government. Furthermore, the Office of the Comptroller General is working now with smaller departments and agencies and Crown corporations to assist them in developing an appropriate approach to evaluating their programs. As of September 30, 1983 a total of 90 departments and agencies were liaising with the Office on program evaluation, 19 of which were Crown corporations.

The audit identified problems in the quality of a number of evaluation studies produced between 1980 and March 1983. It also notes significant improvements in quality over this time period, especially in departments which have had relatively more experience in conducting studies. This confirms the Government's view and, notwithstanding the noted progress, this is an area where further work is required. Among the several recommendations aimed at improving quality is the recommendation that further steps be taken to ensure the availability of qualified evaluators. This is not only a question of acquiring technical skills. Fully qualified evaluators require a combination of technical and managerial skills and a thorough knowledge of programs and program management in the Federal Government. Such knowledge and skills are fully obtained only through appropriate experience. Accordingly, while the need is recognized and is being addressed, progress will be gradual. The Office of the Comptroller General, for its part, offers an ongoing series of seminars and workshops on evaluation, is consulted on most senior staffing actions in evaluation and has recently distributed a discussion paper on human resources management in the evaluation function to departments and agencies. The paper includes a number of proposals concerning the development of evaluation personnel.

Although adequate infrastructure and appropriate quality are required for producing good quality evaluations, the Government's central concern is that the findings of the studies be used in the ongoing management of government. As this report notes, some action has been taken on the recommendations of most of the studies produced to date. This audit finding is particularly welcome and likely reflects the approach several departments have taken in their initial evaluations. In particular, approaches which involve both senior management and line managers in the evaluation and which focus on issues on which departments can act appear to be the most successful.

This audit report includes a number of recommendations relating to the availability of evaluation findings to Parliament. At this time, it appears that the most effective and appropriate way for reporting such information to Parliament would be to include the key findings of evaluations, where relevant, in the Part III of each department's Estimates. This would provide for relevant findings on the effectiveness of programs to be presented in a concise manner to Members of Parliament. Access to information legislation, of course, provides for any member of the public to request a particular evaluation study. Accordingly, it would seem that a requirement to report also in departmental annual reports would be unnecessary, as would be, perhaps, other procedures to table reports in the House. With these developments, it may be appropriate for the Public Accounts Committee to consider the need for tabling all evaluation study reports, as had been recommended prior to these developments.

Finally, the audit report recommends that there is a need for an improvement of procedures to conduct interdepartmental evaluations. Such evaluations have taken place in the past and several are now under way. Existing procedures of the Cabinet Committee system do provide for the identification of the need for such evaluations, but few have been called for, perhaps due to the need first to demonstrate the success of evaluation on less complex issues. The Office of the Comptroller General is currently preparing discussion papers on evaluation in several interdepartmental areas and will be exploring further procedures to facilitate such evaluations whenever there is a demonstrated need for them by an agency which can act on the findings.

This audit of the program evaluation function in the Federal Government has been useful and its findings provide a valuable information base for deciding the future directions the Government will be taking in developing and enhancing the evaluation capability of departments and agencies. As confirmed by this audit report, the continued development and indeed existence of this evaluation capability will require a continual monitoring by the Office of the Comptroller General. Increasing attention will be devoted toward improving the quality of evaluation products and to ensuring that the evaluation findings are used by all levels of mangement within government.

IX

PROFESSIONAL ISSUES AND FUTURE DIRECTIONS

Members of the evaluation community represent a diversity of backgrounds, job roles, and evaluation interests. That any sense of community has been maintained among such a diverse group is both impressive and informative. It indicates that, at some level, individual differences across evaluators become secondary to overarching professional issues and that there is a shared need among evaluators to address some issues collectively. Two such overarching issues are *how* evaluation is practiced and *where* evaluation is conducted. The importance of these issues is perhaps most evident when one views evaluation as a marketable product. The image and perceived utility of the product, as well as the optimal market for the product, are bottom-line issues that influence the viability of the evaluation enterprise.

The Evaluation Research Society (ERS) has dealt explicitly with the issue of how evaluation is practiced by developing the Standards for Program Evaluation, which are reprinted here. The ERS Standards were developed with several goals in mind, including the desire to improve the quality and image of evaluation, to guide program evaluation practice, and to provide funding agencies with a means of assessing the relative value of evaluation proposals and evaluation products.

Given the need to represent an extremely diverse audience of evaluators, the ERS Standards were purposely stated in very general terms. In his review of them, Cronbach praises this uncertainty in the standards because he believes that highly specific standards would have an adverse effect on evaluation. Such standards would, for example, inhibit creativity and innovation in the practice of evaluation, and they would establish a ceiling for what is considered excellence, leaving little room for improvement.

While Cronbach views the ERS Standards as satisfactory at the present time, he does discuss several areas in which they could be improved. He notes in particular that the standards' treatment of validity is deficient and that their treatment of reliability is stereotyped and ambiguous. He notes as well several of the standards' unvoiced messages, including their bias toward evaluation as a service for officials, to the near exclusion of evaluation as a service for persons who are served by a program or persons who will be affected by the outcome of the evaluation. Cronbach concludes that the ERS Standards will not function to standardize (and hence, restrain) evaluation practice; instead, he believes they primarily have a symbolic function, as the mark of the maturity of the evaluation profession and as a reminder of the ideals of the evaluation community.

Fetterman discusses professional issues that fall outside the realm of the ERS Standards. Based upon his own ethnographic work with educational programs, he discusses the ethical dilemmas and hazards that arise from urban fieldwork and contract research. Some of the issues raised by Fetterman are the following:

- How should evaluators respond when they obtain "guilty knowledge" (i.e., confidential knowledge of illegal activities)?
- How should they respond when they acquire "dirty hands" (i.e., when they are in a situation from which they cannot emerge innocent of wrongdoing)?
- Should all potentially damaging information about a program be included in the evaluation report, and with whom is the evaluator morally bound to share the findings of his or her research?

Fetterman acknowledges that resolving these ethical dilemmas may be a tortuous process. Nevertheless, they are part of the reality of the fieldwork experience, and Fetterman believes that developing moral decision-making guidelines is imperative if one is to deal effectively with these professional issues.

The second issue that is of overarching concern to the evaluation community is where evaluation is practiced. Traditionally, the primary marketplace for evaluation services has been social programs in the public sector. In recent years, political and fiscal support for social programs has declined, so much so that the continued growth and viability of the evaluation profession is threatened. While the seriousness of this threat is equally recognized by Calingo, Perloff, and Bryant and by Cook, their suggestions regarding how the evaluation community should address this threat are quite different. Calingo et al. believe that the solution lies in seeking new marketplaces for evaluation, specifically in the private sector. They propose a typology of evaluation opportunities in the private sector, and they believe that traditional evaluators can effectively compete for some, but not all, of these opportunities. The key to successful entry into the private sector, according to Calingo et al., is in attempting to enter only those areas where the skills of traditional evaluators give them an advantage over potential competitors in industry. Calingo et al. conclude that there are many evaluation opportunities in the private sector and that traditional evaluators may open up this marketplace if they so desire.

In contrast, Cook argues that very few opportunities exist for evaluators who wish to enter the private sector, primarily because evaluation has long been practiced there in what seems to be a satisfactory manner. To deal with the threat of declining support for evaluation, Cook suggests that, rather than trying to expand, evaluators should take stock of their professional experiences, consolidate what they know, and prepare for the time when a more socially conscious administration takes office. Such consolidation, Cook argues, should take place along two lines. First, evaluators should confront

the widespread beliefs that social programs do not work and that evaluations are too insensitive to detect true, but modest, effects. Based upon his understanding of the logic of evaluation, Cook explains how the difficulties evaluators face in implementing this logic have reduced the perceived utility of evaluation. To confront the undeserved aspect of this image, Cook advocates consolidating the research evidence on the programs, types of local projects, and elements of practice that seem most useful. Specifically, evaluators may conduct meta-analyses of past evaluations to detect otherwise undocumented evidence of program and evaluation utility. In addition, Cook feels that those responsible for the welfare of evaluation should begin to summarize some of the mistakes of the past that evaluators now know how to avoid or overcome.

The second way in which Cook believes the evaluation community may consolidate its position is by determining the types of evaluations that promise the greatest and least payoffs. In this regard, Cook questions the utility of mandated evaluations conducted by in-house evaluators, and he advocates that the evaluation community give more thought to deciding what to evaluate and what questions they should try to answer when they do evaluate.

The articles by Calingo et al. and Cook suggest that evaluators must move either in the direction of expansion or in the direction of consolidation. For the individual evaluator, this choice is real because one evaluator cannot pursue both of these goals simultaneously. Within the evaluation community, however, there exists the potential to pursue both of these professional goals. Such diversity of interests has been a hallmark of the evaluation community since its beginning and will continue to be a vital resource as the profession confronts the challenges of the future.

39

Evaluation Research Society Standards for Program Evaluation

ERS Standards Committee

The Evaluation Research Society was established in 1976 to serve the professional needs of the growing number of people engaged in program evaluation. More often than not, evaluators work under some other official title, such as program analyst, research associate, auditor, or program planner. The programs they examine range over a wide spectrum; for example, health, education, welfare, law enforcement, public safety, rehabilitation, urban development, defense, environmental protection, training, certification, licensing, business and personnel systems, museums, and media. Further, evaluators have diverse backgrounds and come from a variety of disciplines, including economics, psychology, sociology, public policy, operations research, engineering, systems analysis, and biometry.

This diversity is reflected in the membership of the Evaluation Research Society (ERS) and is its distinctive characteristic. Recently, there was a further enrichment of this diversity of individuals and interests, which resulted from the merger of the Council for Applied Social Research (CASR) with the ERS. There are, of course, other groups that focus on evaluation within more narrowly defined domains—for example, higher education or public policy.

From ERS Standards Committee, "Evaluation Research Society Standards for Program Evaluation," pp. 7–19 in *Standards for Evaluation Practice* (New Directions for Program Evaluation, no. 15).

Why Did the ERS Develop Standards for Program Evaluation?

The ERS believes that even though evaluators have different titles, work in different areas, and come from different backgrounds, they have common concerns and interests; and that evaluation theory and practice will benefit from interdisciplinary sharing. These beliefs have led to a search for standards to guide program evaluation practice and to focus attention on issues facing the emerging profession.

Some have asked why, in the face of a well-conceived and well-publicized national effort to develop program evaluation standards (that is, the Joint Committee on Standards for Educational Evaluation, Daniel L. Stufflebeam, chair), the ERS undertook a parallel activity. First, the Joint Committee's standards focus only on educational programs and would therefore not represent the broader interests of the ERS membership. Further, standards and guidelines were available from a number of other sources, including the U.S. General Accounting Office (GAO) and the Office of the Auditor General of Canada. The ERS acknowledges a considerable debt to the Joint Committee and these other sources for ideas we have borrowed and incorporated in the ERS standards.

Some have asked why the ERS limited its attention to program evaluation. ERS recognizes that program evaluation is only one kind of evaluation that is important for use in decisions about individuals, institutions, and society. However, program evaluation is the enterprise in which the majority of the ERS membership is involved. In the future, the ERS may need to give similar attention to personnel evaluation, product evaluation, proposal evaluation, and other evaluation applications. However, we felt that a more modest initial standards-development effort would stand a better chance of completion and application than an attempt to encompass such a diverse range of evaluation targets. As indicated below, the ERS standards are quite broad, even within the program evaluation delimitation.

What Aspects of Program Evaluation Do These Standards Encompass?

While some people tend to think of program evaluation as a one-shot effort to determine the overall worth of a program, in fact this is only one of several general categories of evaluation. These general categories can be defined both by the purpose of the evaluation effort and by the kinds of activities that are stressed. Some of the categories are associated more with some program contexts and settings than with others, and, as a consequence, the work of some evaluators is likely to fall more in one category than in another. How-

ever, many evaluators are comfortable working in several of the categories, many evaluation efforts encompass more than one of the categories, and indeed—especially in the case of resident evaluators in an agency—evaluators are often expected to be expert in a wide range of evaluation services. The general categories are as follows:

1. *Front-end analysis (preinstallation, context, feasibility analysis).* This includes evaluation activities that take place prior to the installation of a program: to confirm, ascertain, or estimate needs (needs assessments), adequacy of conception, operational feasibility, sources of financial support, and availability of other necessary kinds of support (for example, organizational). The results should provide useful guidance for refining program plans, determining the appropriate level of implementation, or deciding whether to install the program at all).

2. *Evaluability assessment.* This includes activities undertaken to assess whether other kinds of program evaluation efforts (especially impact evaluation) should be initiated. The emergence of evaluability assessment as a legitimate and distinctive enterprise represents a growing professional concern with the costs of evaluations in relation to their benefits, as well as with identifying the general characteristics of programs (significance, scope, execution, and so forth) that facilitate or hinder formal evaluation efforts. Evaluability assessment may encompass inquiries into technical feasibility (for example, Can valid performance indicators be devised?), policy matters (for example, Do program directors understand what kinds of information the proposed evaluation would produce? Is the funding agency's interest in the program likely to be short lived?), and, of course, the characteristics of the program itself (for example, Has it in fact been installed?).

3. *Formative (developmental, process) evaluation.* This includes testing or appraising the processes of an ongoing program in order to make modifications and improvements. Activities may include analysis of management strategies and of interactions among persons involved in the program, personnel appraisal, surveys of attitudes toward the program, and observation. In some cases, formative evaluation means field-testing a program on a small scale before installing it more widely. The formative evaluator is likely to work closely together with program designers or administrators and to participate directly in decisions to make program modifications.

4. *Impact (summative, outcome, effectiveness) evaluation.* This evaluation category corresponds to one of the most common definitions of evaluation—that is, finding out how well an entire program works. The results of impact evaluation—or of *program results review* or similar terms used in some governmental settings—are intended to provide information useful in major decisions about program continuation, expansion, or reduction. The challenges for the evaluator are to find or devise appropriate indicators of impact and to

be able to attribute types and amounts of impact to the program rather than to other influences. Some knowledge or estimate of conditions before the program was applied—or of conditions in the absence of the program—is usually required. Impact evaluations differ in the degree to which the search for appropriate indicators goes beyond the stated objectives or expectations of the program formulators, directors, funders, or other sponsors of the evaluation. However, there is rather substantial agreement that the more independent the evaluator is, the more credible the results of the impact evaluation will be, so long as the expectations of people who manage, oversee, or influence the program are reflected in the evaluation. Achieving a balance among potentially conflicting criteria will be a continuing challenge.

5. *Program monitoring.* This is the least acknowledged but probably most practiced category of evaluation, putting to rest the notion that the evaluator necessarily comes in, does the job, and then gets out. From the GAO to human service agencies in states and provinces to military training installations, there are substantial requirements to monitor programs that have already been installed, sometimes long ago. These programs may or may not once have been the subject of front-end analysis, process evaluation, impact evaluation, and perhaps even secondary evaluation (see 6, below). The kinds of activities involved in these evaluations vary widely, ranging from periodic checks of compliance with policy to relatively straightforward tracking of services delivered and counting of clients. Program monitoring may include purposes or results found also under other evaluation categories; for example, it may involve serious reexamination of whether the needs the program was originally designed to serve still exist, or it may suggest system modification, updating, and revitalization.

6. *Evaluation of evaluation (secondary evaluation, metaevaluation, evaluation audit; may include utilization evaluation).* These activities are applied most frequently to impact evaluations and are stimulated by various interests, such as scholarly investigation, requirements of agencies in coordination or oversight roles, unwillingness of the evaluatee to accept the original evaluation results, or interest in the after-effects of the evaluation on the program. Evaluations of evaluations may take a variety of forms, ranging from professional critiques of evaluation reports and procedures to reanalyses of original data (sometimes with different hypotheses in mind) to collection of new information. In the case of programs that generate widespread public interest (for example, Head Start and veterans' programs), secondary evaluators may examine the results of a number of different evaluations (including evaluations of program units and components) in order to estimate overall impact. Those involved in the relatively new movement to study whether and how evaluation results are used caution that, although utilization evaluation is included here as a special kind of evaluation of evaluation, failure of utilization is not necessarily or solely a failure of evaluation.

The preceding descriptions of six general classes of program evaluation make it clear that a broad range of meaning can be attached to the statement that someone is evaluating a program. As a frame of reference for this document, the classification scheme allows for the applicability of some standards to some categories and not to others. In fact, most of the standards apply to all categories, but when a standard is particularly relevant to only some categories, that case is specifically noted.

What Are the Standards Like and How Are They Organized?

The Standards are organized into six sections: (1) Formulation and Negotiation, (2) Structure and Design, (3) Data Collection and Preparation, (4) Data Analysis and Interpretation, (5) Communication and Disclosure, and (6) Utilization. These are listed roughly in order of typical occurrence, and all six of these phases are normally included in front-end analysis, evaluability assessment, formative evaluation, impact evaluation, and program monitoring. Secondary evaluations, however, may not include any new data collection or data analysis.

Frequently, there are significant implications of the standards in one section for standards in later sections. For example, if the Formulation and Negotiation standards are followed, the evaluator should be in a much better position to meet the Structure and Design standards. Or, violations of Data Collection and Preparation standards could make it very difficult to meet Data Analysis and Interpretation standards.

The Standards themselves take the form of simple admonitory statements. It has been suggested that more detail may be needed in some of the statements and that examples of acceptable practices in different contexts might enhance their meaning. However, the drafting committee concluded that the decision to make such additions should follow identification of ambiguities encountered in attempts to use the document and that examples of acceptable practices should be derived from those experiences.

The committee wishes to underscore its view that this initial formulation of standards is just that. This document is, and should continue to be, a live one, subject to periodic reexamination and revision.

In practice, judgment about the compliance of a given evaluation with the Standards will of course require that the context of the evaluation effort be considered. Moreover, the basis for judgment should be what an informed, disinterested party would consider reasonable and appropriate in the circumstances.

These Standards are specific to program evaluation and do not encompass the full body of legal requirements, governmental regulations, and accepted norms for professional and corporate conduct to which evaluators are subject.

Formulation and Negotiation

Before an evaluation project or program is undertaken, the concerned parties should strive for a clear mutual understanding of what is to be done, how it is to be done and why, and for an appreciation of possible constraints or impediments. However, the knowledge initially available will vary widely, and the parties to the evaluation should be prepared to modify early formulations as information and circumstances change.

1. The purposes and characteristics of the program or activity to be addressed in the evaluation should be specified as precisely as possible.

2. The clients, decision makers, and potential users of the evaluation results should be identified and their information needs and expectations made clear. Where appropriate, evaluators should also help identify areas of public interest in the program.

3. The type of evaluation which is most appropriate should be identified and its objectives made clear; the range of activities to be undertaken should be specified. (See categories 1-6, Introduction).

4. An estimate of the cost of the proposed evaluation and, where appropriate, of alternatives should be provided; this estimate should be prudent, ethically responsible, and based on sound accounting principles.

5. Agreement should be reached at the outset that the evaluation is likely to produce information of sufficient value, applicability, and potential use to justify its cost.

6. The feasibility of undertaking the evaluation should be estimated either informally or through formal evaluability assessment (see page 9).

(Some of the factors to consider are the clarity of the program description and objectives; prospects for needed cooperation; the plausibility of any postulated cause-effect relationships; the availability of time, money, and expertise to carry out the evaluation; and administrative, fiscal, and legal constraints.)

7. Restrictions, if any, on access to the data and results from an evaluation should be clearly established and agreed to between the evaluator and the client at the outset.

(In some cases—for example, government-sponsored studies where freedom of information statutes apply and where it is understood that the results of evaluation studies automatically go into public domain—the right-to-know question is not negotiable. The sponsor or evaluator is obligated to point this out at the beginning. In other cases—for example, confidential studies undertaken for private individuals and organizations—the client may rightfully expect the confidentiality of the findings to be maintained.)

8. Potential conflicts of interest should be identified, and steps should be taken to avoid compromising the evaluation processes and results.

9. Respect for and protection of the rights and welfare of all parties to the evaluation should be a central consideration in the negotiation process.

10. Accountability for the technical and financial management of the evaluation once it is undertaken should be clearly defined.

11. All agreements reached in the negotiation phase should be specified in writing, including schedule, obligations and involvements of all parties to the evaluation, and policies and procedures on access to the data. When plans or conditions change, these, too, should be specified.

12. Evaluators should not accept obligations that exceed their professional qualifications or the resources available to them.

Structure and Design

The design for any evaluation cannot be conceived in a vacuum. It is necessarily influenced by logistical, ethical, political, and fiscal concerns and therefore must take these as well as methodological requirements into account. This applies to each of the six types of evaluation specified in the introduction. Some of the principal concerns that extend beyond methodological requirements and influence the design itself are embodied in standards 1, 2, 3, 6, 7, 8, and 9. Designs will vary in rigor, and not all measurements are equally objective. However, even with these broad variations, the following standards generally apply. For example, the approach to a case study is as subject to specification as the design of an experimental study; the reliability of judgments is as much at issue as the reliability of objective tests.

13. For all types of evaluations, a clear approach or design should be specified and justified as appropriate to the types of conclusions and inferences to be drawn.

14. For impact studies, the central evaluation design problem of estimating the effects of nontreatment and the choice of a particular method for accomplishing this should be fully described and justified.

15. If sampling is to be used, the details of the sampling methodology (choice of unit, method of selection, time frame, and so forth) should be described and justified, based on explicit analysis of the requirements of the evaluation, including generalization.

16. The measurement methods and instruments should be specified and described, and their reliability and validity should be estimated for the population or phenomena to be measured.

17. Justification should be provided that appropriate procedures and instruments have been specified.

18. The necessary cooperation of program staff, affected institutions, and members of the community, as well as those directly involved in the evaluation, should be planned and assurances of cooperation obtained. (See standard 11.)

Data Collection and Preparation

These standards assume that data collection is carried out within the specifications of a sound design and plan of work. (See standards 1–18.) However, at the time the data collection methods are specified, reasonable changes should be made in the design in order to accommodate the realities of the situation. During the data collection process, if logistical difficulties occur or circumstances change significantly, the design and work plan should be revised accordingly.

19. A data collection preparation plan should be developed in advance of data collection.

20. Provision should be made for the detection, reconciliation, and documentation of departures from the original design.

21. Evaluation staff should be selected, trained, and supervised to ensure competence, consistency, impartiality, and ethical practice.

22. All data collection activities should be conducted so that the rights, welfare, dignity, and worth of individuals are respected and protected.

23. The estimated validity and reliablity of data collection instruments and procedures should be verified under the prevailing circumstances of their use. (See standard 16.)

24. Analysis of the sources of error should be undertaken, and adequate provisions for quality assurance and control should be established.

25. The data collection and preparation procedures should provide safeguards so that the findings and reports are not distorted by any biases of data collectors.

26. Data collection activities should be conducted with minimum disruption to the program under study and with minimum imposition on the organizations or persons from whom data are gathered.

27. Procedures that may entail adverse effects or risks should be subjected to independent review and then used only with informed consent of the parties affected.

28. Data should be handled and stored so that release to unauthorized persons is prevented and access to individually identifying data is limited to those with a need to know. (See standard 7.)

29. Documentation should be maintained of the source, method of collection, circumstances of collection, and processes of preparation for each item of data.

30. Appropriate safeguards should be employed to ensure against irrecoverable loss of data through catastrophic events.

*Where secondary data are used, the evaluator should try to ascertain whether the processes through which the data were originally produced conform to these standards.

Data Analysis and Interpretation

The choice of analytic procedures, like the choice of data collection methods, is more or less dictated by the structure and design of the evaluation. At the data analysis stage, the evaluator no longer has much freedom to change the design and is required to temper the analyses to characteristics of the data actually collected. New methods and procedures, some detailed and rigorous, are appearing in the literature, and evaluators should be aware of these innovations and give them full consideration.

31. The analytic procedures should be matched to the purposes of the evaluation, the design, and the data collection.

32. All analytic procedures, along with their underlying assumptions and limitations, should be described explicitly, and the reasons for choosing the procedures should be clearly explained.

(The level of detail required in the descriptions will vary with the familiarity of the procedure to the primary audience.)

33. Analytic procedures should be appropriate to the properties of the measures used and to the quality and quantity of the data.

34. The units of analysis should be appropriate to the way the data were collected and the types of conclusions to be drawn.

35. Justification should be provided that the appropriate analytic procedures have been applied.

36. Documentation should be adequate to make the analyses replicable.

37. When quantitative comparisons are made, indications should be provided of both statistical and practical significance.

38. Cause-and-effect interpretations should be bolstered not only by reference to the design but also by recognition and elimination of plausible rival explanations.

39. Findings should be reported in a manner that distinguishes among objective findings, opinions, judgments, and speculation.

Communication and Disclosure

Good communication is obviously essential to a well-formulated and well-executed evaluation and to use of the results. In particular, good communication is necessary to clarify the nature of the program, the expectations for the evaluation, and even the type of evaluation required (see standards, 1, 2, and 3); to anticipate restrictions on release of results and potential conflicts of interest (see standards 7 and 8); to establish accountability for the effort (see standards 10 and 11); to secure the cooperation of parties involved in the program and the evaluation (see standards 18 and 27); and to distinguish objec-

tive findings clearly from opinion and interpretation (see standard 39). In short, communication is not to be equated solely with the final report. However, most evaluation efforts do produce certain formal reports, intermediate and final, written and oral, and there are standards these reports should meet.

40. Findings should be presented clearly, completely, and fairly. (See standard 39.)

41. Findings should be organized and stated in language understandable by decision makers and other audiences, and any recommendations should be clearly related to the findings.

42. Findings and recommendations should be presented in a framework that indicates their relative importance.

43. Assumptions should be explicitly acknowledged.

44. Limitations caused by constraints on time, resources, data availability, and so forth should be stated. (See standards 5, 6, 7, 11, and 12.)

(Suggestions should be included about those issues and questions that need further study.)

45. A complete description of how findings were derived should be accessible.

46. Persons, groups, and organizations who have contributed to the evaluation should receive feedback appropriate to their needs.

47. Disclosure should follow the legal and proprietary understandings agreed upon in advance (standard 7), with the evaluator serving as a proponent for the fullest, most open disclosure appropriate.

48. Officials authorized to release the evaluation data should be specified.

49. The finished data base and associated documentation should be organized in a manner consistent with the accessibility policies and procedures. (See standards 7, 28, 29, 32, and 36.)

Use of Results

The usual reasons for conducting an evaluation are functional ones: to help those affected determine the feasibility of undertaking the program or to assess its operation and effects. The use of evaluation results cannot be guaranteed, of course, but it will be more likely if careful attention is given to the information needs of the potential users of the results throughout all phases of the evaluation (see especially standards 2, 3, 18, 40–46.) Beyond the day-to-day processes of encouraging responsiveness to the evaluation, there are some other considerations that the evaluator needs to keep in mind.

50. Evaluation results should be made available to appropriate users before relevant decisions must be made.

51. Evaluators should try to anticipate and prevent misinterpretations and misuses of evaluative information. (The evaluator, of course, cannot be

held responsible for misuses of evaluative information. Nevertheless, follow-up contacts with users, rebuttals of misinterpretation, and promotion of an open exchange of information should be a part of the evaluator's responsibility.)

52. The evaluator should bring to the attention of decision makers and other relevant audiences suspected side effects—positive or negative—of the evaluation process.

53. Evaluators should distinguish clearly between the findings of the evaluation and any policy recommendations based on them.

(If evaluators are called upon to go beyond the findings and to make policy recommendations or if they initiate such recommendations, they must make clear the difference between such recommendations and the actual findings of the evaluation.)

54. In making recommendations about corrective courses of action, evaluators should carefully consider and indicate what is known about the probable effectiveness and costs of the recommended courses of action.

55. Evaluators should maintain a clear distinction between their role as an evaluator and any advocacy role they choose to adopt.

(Evaluators should be aware of the apparent conflict between advocating certain positions and presenting evaluation results. Evaluators may wish to take advocacy stands, but when they do they should not assume that they possess any special status or competence.)

Sources

American Personnel and Guidance Association. "Responsibilities of Users of Standardized Tests." *Guidepost,* October 5, 1978.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. *Standards for Educational and Psychological Tests.* Washington, D.C.: American Psychological Association, 1974.

Anderson, S. B., and Ball, S. "Ethical Responsibilities in Program Evaluation." In S. B. Anderson and S. Ball, *The Profession and Practice of Program Evaluation.* San Francisco: Jossey-Bass, 1978.

Auditor General of Canada. "Study in Procedures in Cost Effectiveness: Chapter 4: Measuring Efficiency and Chapter 5: Evaluating Effectiveness." In 100th Annual Report to the House of Commons (*Fiscal Year Ended March 31, 1978*). Available from Printing and Publishing, Supply and Services Canada Hull, Quebec, Canada, KIA 0S9.

Baron, J. B., and Baron, R. M. "In Search of Standards." In R. Perloff and E. Perloff (Eds.), *New Directions for Program Evaluation: Values, Ethics, and Standards in Evaluation,* no. 7. San Francisco: Jossey-Bass, 1980.

Code of Federal Regulations. Title 45, CFR, part 46. Washington, D.C.: U.S. Department of Health, Education and Welfare, revised January 11, 1978.

Committee on Evaluation Research, Social Science Research Council. *Audits and Social Experiments: A Report Prepared for the U.S. General Accounting Office.* Washington, D.C.: U.S. General Accounting Office, 1978.

Comptroller General of the United States. *Assessing Social Programs Impact Evaluations: A Checklist Approach* (Exposure Draft). Washington, D.C.: U.S. General Accounting Office, 1978.

Comptroller General of the United States. *Evaluation and Analysis to Support Decision Making.* Washington, D.C.: U.S. General Accounting Office, 1976.

Comptroller General of the United States. *Federal Program Evaluation: Status and Issues.* Washington, D.C.: U.S Government Printing Office, 1978.

Comptroller General of the United States. *Finding Out How Programs Are Working: Suggestions for Congressional Oversight.* Washington, D.C.: U.S. General Accounting Office, 1977.

Comptroller General of the United States. *Guidelines for Model Evaluation* (Exposure Draft). Washington, D.C.: U.S. General Accounting Office, 1979.

Comptroller General of the United States. *Standards for Audit of Governmental Organizations, Programs, Activities, and Functions.* Washington, D.C.: U.S. Government Printing Office, 1972.

Division of Industrial-Organizational Psychology, American Psychological Association. *Principles for the Validation and Use of Personnel Selection Procedures.* Dayton, Ohio: Industrial-Organizational Psychologist, 1975.

Educational Testing Service. *Principles, Policies, and Procedural Guidelines Regarding ETS Products and Services.* Princeton, N.J.: Educational Testing Service, 1979.

Emrich, R. L. *Proposed Evaluation Guidelines and Standards.* Sacramento: California Council on Criminal Justice, 1973.

Emrich, R. L. *Standards for Metaevaluation.* (Preliminary version.) Hackensack, N.J.: NCCD Research Center, 1974.

Flaherty, D. H. "The Bellagio Conference on Privacy, Confidentiality, and the Use of Government Microdata." In R. F. Boruch (Ed.), *New Directions for Program Evaluation: Secondary Analysis,* no. 4. San Francisco: Jossey-Bass, 1978.

Gibbs, L. E. "A Code of Ethics for Evaluators? Detailed Responses to Questions 1 and 3 Appearing in *Evaluation*." Unpublished manuscript, 1977.

International Personnel Management Association Assessment Council. *Standards for Item Bank Sharing.* Washington, D.C.: International Personnel Management Association, 1978.

Joint Commission on Accreditation of Hospitals. "The Balanced Service System." In *Principles for Accreditation of Community Mental Health Programs.* Chicago: Joint Commission on Accreditation of Hospitals, 1976.

Joint Committee on Standards for Educational Evaluation. *Standards for Evaluation of Educational Programs, Projects, and Materials* (Draft). Kalamazoo: Western Michigan University Research Center, 1979.

Joint Dissemination Review Panel. *Ideabook.* Washington, D.C.: U.S. Department of Health, Education and Welfare, 1977.

Klein, S. *Ethics for R&D Management: What Is Needed?* Washington, D.C.: National Institute of Education, U.S. Department of Health, Education and Welfare, 1977.

McClintock, C. C. "Issues in Establishing and Enforcing Professional Research Ethics and Standards." *Advances in Consumer Research,* 1977, *4,* 258-261.

Molner, S. F. "Trapped Bedfellows: A Comment on Windle and Neigher." *Evaluation and Program Planning,* 1978, *1,* 109-112.

Robbin, A. "Ethical Standards and Data Archives." In R. F. Boruch (Ed.), *New Directions for Program Evaluation: Secondary Analysis,* no. 4. San Francisco: Jossey-Bass, 1978.

Sheinfeld, S. N. "The Evaluation Profession in Pursuit of Value." *Evaluation and Program Planning,* 1978, *1* (2), 113-115.

Sieber, J. E., and Sanders, N. "Ethical Problems in Program Evaluation: Roles, Not Models." *Evaluation and Program Planning,* 1978, *1* (2), 117-120.

Stockton, R. *Principles and Practice for Education R&D Management with Ethical Considerations.* (Outline for a monograph.) Washington, D.C.: American Educational Research Association, Special Interest Group on Research Management, November 1978.
Task Force on Development of Assessment Center Standards. "Standards and Ethical Considerations for Assessment Center Operations." In J. L. Moses and W. C. Byham (Eds.), *Applying the Assessment Center Method.* New York: Pergamon Press, 1977.
Windle, C., and Neigher, W. "Ethical Problems in Program Evaluation: Advice for Trapped Evaluators." *Evaluation and Program Planning,* 1978, *1* (2), 97–108.

Members of Drafting Committee

Scarvia B. Anderson (Chair to 1980)
Educational Testing Service

Larry A. Braskamp
University of Illinois

Wallace M. Cohen
U.S. General Accounting Office

John W. Evans
Educational Testing Service
(formerly with U.S. Department of Education)

Alan Gilmore
Office of Auditor General of Canada

Keith E. Marvin (Chair 1980 to present)
U.S. General Accounting Office

Virginia C. Shipman
Educational Testing Service

James J. Vanecko
Peter Merrill Associates, Boston
(formerly with U.S. Department of Education)

Ronald J. Wooldridge
New York State Office of Mental Health
(formerly with Georgia Department of Human Resources)

40

In Praise of Uncertainty

Lee J. Cronbach

Two sentiments about the ERS Standards I express at the outset. First, the document as it stands is as satisfactory as could be hoped for in this field at this time. Every principle is open to a reasonable interpretation. Second, murky depths lie beneath statements that seem like placid pools reflecting a limpid sky. Like Descartes, "I encountered nothing so dubious that I could not draw from it some conclusion that was tolerably secure, if this were no more than the inference that it contained in it nothing that was certain" (1969 [1637], p. 99).

The evaluation community will gain something if every member takes these standards as a Hippocratic oath. The community will gain a more profound education if each reader asks of each standard: Under what circumstances would it make sense to go counter to this advice? My critical remarks are intended to stimulate this instructive process, not to scorn the Standards Committee's gift.

Standards: Intentions and Side Effects

Standards have functions, intentional and unintentional. I propose to explore these functions before I consider the probable role of the evaluation standards. This is an opportunity to reflect on what has been in the back of my mind since the early 1950s, when I was one of the group appointed to formu-

From Lee J. Cronbach, "In Praise of Uncertainty," pp. 49–58 in *Standards for Evaluation Practice* (New Directions for Program Evaluation, no. 15).

late what became known as the *Standards for Educational and Psychological Tests* (American Psychological Association, 1954; 1974).

By definition, the function of a standard is to standardize. But, standardization is inimical to innovation; should an agency want to standardize practice? The American Psychological Association (APA) committee that prepared the document on testing tried to avoid a standardizing effect, and so, it seems, did the ERS committee. What can be the function of standards that do not standardize? Before I speak to that point, let me consider functions of standardizing.

Some standards institutionalize mere conventions, for what must be aesthetic reasons. Excellence at the hundred-meter dash is not more meritorious than excellence at eighty-two meters. The hundred-meter standard enhances excellence, however, enabling athletes to perfect techniques specific to a distance, as they could not in preparing for races of miscellaneous lengths. A haiku of seventeen syllables is not intrinsically preferable to a poem of fifteen syllables, but poems that satisfy that additional, arbitrary specification give prouder evidence of human determination. Just as some runners are better at one distance than another, so are some talents presumably better able to express images in fifteen syllables than seventeen. Not only can standards penalize arbitrarily selected producers; as abortifacient, the seventeen-syllable standard deprives the world of delightful fifteen-syllable poems. In general, imposition of conventions supports an academic smugness that has repeatedly brought sterility to arts.

Some standards are intended to raise the quality of information and so to increase the power of purchasers, an example being the octane numbers posted on gasoline pumps. To help consumers become masters of their fate is a fine objective, but even information standards can carry a price. The variable that is standardized tends to become the mark of excellence, while other qualities tend to be overlooked. Americans came to think of automobile engines in terms of horsepower; if a standard for engine efficiency had been advertised over the decades, it would have changed the history of the industry. The old requirement that margarine be white implied its inferiority to butter and inhibited consumer choice.

Reducing cost is one motive for restricting variety. Huge inventories of replacement parts are needed if dimensions vary with the whim of their maker. It was impractical to ship goods over long distances when reloading was required because one rail line ended and the next continued on a different gauge. Paradoxically, reducing variety can increase consumers' options and expand competition. But, standards such as building codes can also keep superior technology from coming into use. Indeed, standards are typically designed to restrain trade. Certification requirements for entry into a profession

are one conspicuous example. Such standards are almost never validated empirically, and they often reflect only the view predominant in a given time.

A standard is likely to be a political compromise, after hard-liners demand severe restriction, and heterodox objections cumulate in a call for laissez faire. Political forces can be so balanced, or understanding of the topic can be so immature, or the topic can be inherently so complex that the standards are no more than platitudes. Those who legislate inexplicit standards endow bureaucrats with power that can be abused. The stringency of guidelines and the rigor of enforcement can shift with the political balance.

The APA test standards were left deliberately open. Some zealots of 1950 wanted to discourage "invalid" tests and hoped to set some unequivocal standard—perhaps something like "The validity coefficient must reach 0.40 before a test is recommended for a given use." Alternative views prevailed; the standards did not draw a line between good tests and bad. Rather, they called for providing professionals with information that would enable each one to judge a test's adequacy for the use intended. Since the research to collect such information was costly, the standards discouraged development of innovative tests with small markets. Test development came to be concentrated in well-capitalized firms that sought returns by producing for the largest, hence most traditional, markets. In the 1970s, government regulations and court opinions on fair employment incorporated the standards by reference; their deliberate equivocality then proved to be an embarrassment, because employers could not demonstrate that they were in compliance.

Standards, even weak ones that do nothing to restrain practice, have a symbolic function. They are reminders of community ideals; loyalty to motherhood and apple pie contribute to social stability, even though some mothers produce indigestible pies. Also, pronouncing in favor of virtue advertises the virtue of the pronouncer. For an organization to issue a set of standards is to claim maturity of thought for the field and puberty for the group itself. Each subsequent edition reasserts a claim to authority over the territory staked out. The sheer existence of the APA test standards strengthened the influence of the psychological profession when regulatory agencies took aim at personnel testing.

The Placid Surface

It is neither surprising nor unsuitable that the ERS Standards reiterate pieties. Evaluators should be competent (standards 6, 12, and 21), respectful of those whom they observe and or sponsors (standards 9 and 22), attentive to their own biases (standards 8 and 15), and lucid about plans and findings (standards 13, 40, and 41).

The standards are delicately worded, bespeaking an admirable diplomacy that found language to accommodate many points of view. The document has an ecumenical concept of the profession; no hint appears that competence in statistics or in economics, say, counts for more than any other qualification. A crucial sentence discourages would-be enforcers from taking the Standards as literal and inviolable; in no respect is the evaluator's plan or action to be more diligent or more nearly perfect than is "reasonable. . . in the circumstances." One functional consequence of the Standards, then, is to foster community among evaluators by promulgating a spirit of tolerance—a spirit that is not the usual companion of high-mindedness.

Ostensibly, the ERS document is the opposite of restrictive, welcoming the greatest variety of evaluation approaches. Five broad categories of inquiry are recognized as aspects of program evaluation. By implication, the document denies that measuring effect size is primary and that all other inquiry is of lesser value, and it puts behind us the notion that there is one best design for an evaluation. A door is left ajar for impact evaluations that do not focus on goal attainment (although the document does not speak wholeheartedly on that point).

The Standards are unlikely to restrain trade directly. It would be impossible, I think, for a regulator to convert them into a scorecard for an evaluation or a set of evaluations. Almost all the document constitutes a call for fuller communication. It says that evaluators should plan deliberately and should be prepared to give their reasons for each choice. The document does not place a greater burden of proof on particular choices (for example, of a case study over a quantitative survey). This amounts to an outright rejection—no less important for being silent—of efforts to improve evaluation by imposing academic standards.

It is unfortunate that the attempt to be both high-minded and permissive leaves some sentences so general as to suggest that some issues were not faced squarely. I am inclined to think that attempts to be more definite would have come to naught, because there are defensible options at almost every step in an evaluation, and, with regard to a particular option, the balance of costs and benefits changes from one evaluation to the next. Indeed, it may never become possible to make explicit the contingencies under which, for example, information should be collected from or on uncooperative subjects (in violation of standard 18).

In sum, the Standards lack bite, and their contribution lies mostly in the symbolism of their existence. Beyond that, they tell evaluators to concern themselves with financial accountability, subjects' rights, documentation of the steps from data to conclusions, clarity of communication, and so on, not solely with research operations. These reminders can raise consciousness in

evaluators who conceive of their roles too narrowly. Also, they can strengthen the hand of evaluators in negotiating with sponsors who are insensitive to one or another of these ideals.

Unvoiced Messages

The Standards subtly encourage some evaluation approaches over others. This imbalance probably arises from the inevitable character of standards. Standards purport to convey directives for action, not mere sentiments, and it is far easier to make explicit recommendations on some topics than on others. An unintended weighting results, which throws the resulting standards out of balance. What, then, is the slant of these Standards?

The ERS document is aligned more with Wholey's (1979) views than with those of any other recent position statement on evalution. Wholey sees evaluation primarily as a service to officials and the impetus of evaluation as arising from what officials want to know. In these Standards, the agency contracting for the evaluation is referred to as "the client" (standard 17). Is not the true client a pluralistic policy-shaping community and the sponsor merely its agent in buying information (Cronbach and Associates, 1980)? The ERS Standards state that "the parties to the evaluation" are asked to come to a mutual understanding; this phrase seems not to include persons who are served by the program or persons who will be affected by decisions. House, Scriven, and some of the rest of us have stressed the evaluator's special responsibility to introduce questions that officials overlook or prefer not to have investigated. The Standards admit "the public interest" (standard 2) only in a grudging way that surely has in mind only the abstract, generalized public, not fragmented constituencies and categories of clients.

Significant content is conveyed by what the Standards do not say. The first such message is that evaluators have no reason to concern themselves with values; the word *values* does not appear in the document. Yet, evaluators could reasonably take as a primary task the identification of value issues pertinent to a program.

Although the Standards are called standards for evaluation, these are in truth standards of conduct for evaluators. What about the conduct of sponsors? A few references to clarity in negotiating the contract (standards 5 and 7) seem to place symmetric responsiblity on sponsor and evaluator. Beyond that, however, the Standards Committee did not choose to put a finger on acts of sponsors that impair evaluations. Even a few timorous platitudes about good sponsor behavior would serve as an entering wedge for things that the profession should be starting to say. The following standard—a suggestion that I do not expect to have accepted as worded here—exemplifies what I think is missing:

> Unless restriction of reports has been agreed on in advance, the sponsor should release the evaluator's report, as written, within two months of its delivery. The sponsor may appropriately provide comments on an earlier draft for the evaluator's consideration and may appropriately release its own interpretation alongside the report.

The kernel of the ERS document is the Standards proper. The prefatory remarks—a kind of outer wrapping—will receive less attention than the kernel; their less hortatory tone gives their content the weight of a footnote. This would not matter if forematter and text were in accord, but they are not. That is, the forematter expresses an awareness that is imperceptible in the Standards proper.

Whereas the Standards speak throughout of "the evaluation," reference to an evaluation "program" (found only in a preamble) is more in keeping with recent writings. Studies on utilization indicate that it is rare for an outcome evaluation to stimulate large change in the particular service studied. Evaluations are used, in the sense that officials draw thoughts about general policy from the whole emerging corpus of social research, including evaluations. It, therefore, is wise to plan not separate evaluations but a portfolio of studies with overlapping time schedules. Such planning calls for the collective wisdom of sponsoring agencies, social scientists, and concerned citizens (Cronbach and Associates, 1980). None of us is yet clear about the implications of this recent insight for investment in evaluation, but for the Standards to reduce the notion of investigative programs to a single word—printed, as it were, on the wrapper—is a shortcoming.

The preamble gives adequate emphasis to the study of process, of the interconnected events in the setting, delivery, reception, and sequelae of the social service. We do, indeed, need to ask how things work, whether or not we can agree on how well they work. Unfortunately, the word *process* is absent from the Standards proper (and the word *explanation* appears nowhere).

Concern for cause-and-effect relationships does appear (standards 6 and 38). The evaluator who chooses to make a causal statement is directed to "eliminate" plausible rival hypotheses (standard 38). This—one of the few standards that actually issues a command—asks too much. A given design will make some counterexplanations less plausible, but even a strongly controlled design cannot provide certainty regarding the meaning of any one study (Cook and Campbell, 1979, p. 83). Here, in the endorsement of evaluability assessment as a preliminary (standard 6) and in the call for a no-treatment control group or a surrogate in the outcome-oriented study (standard 14), the Standards strongly suggest that evaluators should be concerned with "treat-

ment made a difference" conclusions. Hence, the Standards themselves are less liberal than the preamble.

Standard 6 misconstrues one of Wholey's (1979) main themes. For him, evaluability assessment is not a preliminary; it comes at the end of substantial evaluation activity. One by-product of that inquiry is a recommendation for or against undertaking a formal summative study. For Wholey, the comparative study will most often be a comparison of pretested treatment alternatives. The kernel seems to be preoccupied with old-fashioned treatment–no-treatment null hypotheses at a time when many are calling for studies that trace processes and contingencies to their conclusions. To restore balance, the ERS Standards should include standards intended to invite and strengthen process studies. For openers, I suggest these two:

1. The evaluator should try to make sense of any numerical result.
2. The evaluator should try to explain why outcomes within a treatment differ from site to site.

Process standards probably have to be nebulous. That they are absent from the ERS Standards reflects the natural tendency of standard setters to speak only of the more formal, standardizable aspects of their subject.

Once an evaluation begins, information is brought back from the field—some of it unanticipated. Then, there should be reconsideration and revision of the priorities assigned to evaluation questions and the plans for inquiry. The preamble says this, but the theme receives only a backhanded, passive reference in the Standards themselves (standard 11). The Standards proper are committed to prespecification of evalution purposes, questions, and procedures. The first two sections, which have to do with the launching of the evaluation, contain eighteen standards. Ten of these ask that something be "specified" or "made clear" or "described and justified"; not one suggests that something could be left open for later consideration. Of course, evaluators should plan and specify and make clear; what I miss is a modifier—for example, "tentative" or "preliminary and incomplete."

"The approach to a case study," we are told, "is as subject to specification as the design of an experimental study." The remark is true in a whimsical sense, because an intelligent field experiment is subject to little prespecification. In a number of true experiments that started with impartial allocation of subjects to treatment groups, the inquiries paid off primarily because the investigators thought freshly about the data as they came in. Often, the experimental manipulation played the role only of increasing the range of a variable, such as food intake or income; the analyst regrouped cases on the actual extent

of service or food or income received, without regard to ostensible assignment, in order to bring out more basic phenomena. (See Cronbach, 1982, on this and many other matters on which I question the Standards.)

Construct Validation: Where Art Thou?

The Standards' remarks on validity are off target. The validity of each measurement method should be "estimated" (standards 16 and 23); and should the evaluator not consider the validity of interpretations, rather than of measurements? And, does an estimate not call for a number? Although a regression coefficient can play some part in justifying an employment practice, validating an interpretation of any kind is a qualitative, judgmental process. The interpreter can do no more than lay out an argument defending the interpretation; its plausibility is for the audience to determine (Cronbach and Associates, 1980; House, 1980; Messick, 1981).

Validity in evaluation is almost always of the construct sort. The issue is whether indicators collected by another procedure (and on another realization of the treatment concept) would—when put "in language understandable by" the audience (standard 41)—lead to the same conclusion. The writers of the Standards seem never to have heard of construct validity as the term is applied to evaluation by Cook and Campbell (1979). My sympathetic guess is that their hearts were in the right place, but they found it impossible to write crisp one-sentence regulations for subtle reasoning activities.

Evaluators should be self-critical. They should ask whether the data are biased by respondents' desire to speak well of a service lest it be discontinued or by the overloading of an educational outcome measure with tasks on which one of the competing instructional programs concentrated. Treatment events as well as instruments require this scrutiny. It is an error to regard schools funded by a certain program as treated and schools that received no such funds as untreated. Many school superintendents draw funds from a second source to get innovation into control schools. Demonstrating that this did not happen is a validity study as much as any check on an instrument is.

Construct validation consists of an attempt to falsify a proposed interpretation. The evaluator identifies plausible rival hypotheses and looks for evidence that would support them, hoping not to find it. Because time is limited and rival interpretations are inexhaustible, validation cannot be thorough. An adequate standard would stop with a request for the evaluator, before and after collecting data, to think about possible contaminants. Of course, a multimethod procedure (multiple realizations of treatments as well as multiple indicators) in commendable, but I am pleased that the Standards do not call for it.

The art of design is to decide when and how to make each such investment, as the investment inevitably diverts resources from another aspect of the inquiry.

The call for information on reliability in these same Standards (standards 16 and 23) is stereotyped, and it could be misread. The right question is whether the sample statistics, not individual measurements, are reliable—standard errors are what matter. Procedures that have unimpressive reliability coefficients can be entirely satisfactory for a conclusion based on group means. The Standards rightly do not say that all instruments should have high reliability; their silence contains wisdom that most readers will overlook.

Concluding Comments

The preamble emphasizes that the Standards will change and grow. This change may take the form of increasingly detailed advice. It is said that, as the Standards are used, "examples of acceptable practices" will accumulate as material for a later edition. The hint that some practices are unacceptable makes me uneasy. To be sure, fiscal irresponsibility, for example, is unacceptable. But, in the areas to which most of my comments have been directed, another mode of extension may work to our advantage. Could we not have examples in which different evaluators have followed (or advocated) different courses of action with regard to, for example, extent of quality control? Describing these divergent positions and the arguments for each would make explicit the extent to which evaluation decisions are contingent. That would enhance the educational power of the document and avoid holding up a practice that was right in one circumstance as a model for most circumstances.

These Standards will do little to standardize evaluations, and it will be extremely difficult to use them as a checklist in approving or disapproving an evaluation plan. We should all be grateful for the wisdom and effort that brought about such a happy result.

References

American Psychological Association. *Technical Recommendations for Psychological Tests and Diagnostic Techniques.* Washington, D.C.: American Psychological Association, 1954.

American Psychological Association. *Standards for Educational and Psychological Tests.* Washington, D.C.: American Psychological Association, 1974.

Cook, T., and Campbell, D. T. *Quasi-Experimentation.* Chicago: Rand-McNally, 1979.

Cronbach, L. J. *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass, 1982.

Cronbach, L. J., and Associates. *Toward Reform of Program Evaluation: Aims, Methods, and Institutional Arrangements.* San Francisco: Jossey-Bass, 1980.

Descartes, R. *The Philosophical Works of Descartes Rendered into English.* Vol. 1. Cambridge, England: Cambridge University Press, 1969. (Originally published 1637.)
House, E. *Evaluating with Validity.* Beverly Hills, Calif.: Sage, 1980.
Messick, S. "Evidence and Ethics in the Evaluation of Tests." *Educational Researcher,* 1981, *10* (10), 9-20.
Wholey, J. S. *Evaluation: Promise and Performance.* Washington, D.C.: Urban Institute, 1979.

41

Guilty Knowledge, Dirty Hands, and Other Ethical Dilemmas

The Hazards of Contract Research

David M. Fetterman

FIELD-WORKERS ENCOUNTER MANY PERSONAL and professional hazards in contract research. A few of the situations that can prove hazardous include entrance into the field,' role conflicts, fieldwork in the inner city, ethnographic reports, and the dissemination of findings. Job stress and "burnout" pose additional problems. Urban fieldwork, in particular, forces an ethnographer to confront the realities of guilty knowledge (Polsky 1967)—confidential knowledge of illegal activities—and dirty hands—a situation from which one cannot emerge "innocent of wrongdoing" (Klockars 1979:271). Like Klockars (1969:265), "I personally have little use for the kind of moral study which seeks to understand how angels should behave in paradise and do not intend this analysis to be a contribution to that literature." Implicit in this discussion is that good field-workers are both "competent at their vocations and decent as human beings" (Klockars 1979: 265). Moreover, as a colleague has expressed, a field-worker can only "act morally and responsibly if one knows the situation and understands the actors." Ethical dilemmas generated in the day-to-day interactions between sponsor, researcher, and informant warrant closer examination. "When only outstanding and scandalous cases are defined as matters for ethical concern, then the daily perplexities, interactions, and decisions occurring in the field may well be perceived as merely 'personal.' Ethics then becomes an academic subject, consisting primarily of abstract concepts counterposed by shocking violations" (Cassell 1980:42). A review of these issues serves to guide researchers in this growing field. Moreover, it is hoped that this discussion will be reflexive, encouraging field-workers in various fields to reevaluate their own roles in the pursuit of research.

David Fetterman is a senior member of Stanford University administration. He is conducting qualitative evaluations, appraisals, and audits of academic and business departments at Stanford University. Concurrently, he is a guest lecturer in the Anthropology Department and the School of Education, and serves as a consultant at RMC Research Corporation, a Bay Area educational evaluation firm. *The author is indebted to Michael H. Agar, Carl B. Klockars, and Ned Polsky for producing works that have greatly influenced his thoughts and actions in moral decision making in the field. The author is also indebted to G. D. Spindler, G. K. Tallmadge, and D. S. Waxman for their assistance in the preparation of this manuscript. The National Institute of Education, the State of California's Office of Planning and Educational Research, and Stanford University have provided generous support for the research projects referred to in this discussion. Appreciation is also extended to Corina and her two girls for their contribution to this fieldwork. The views expressed in this article are the author's and do not necessarily reflect the policy of any agency or institution with which the author is associated.*

Context

The hazards discussed in this review were based on my experience as an ethnographer in a Bay Area contract research corporation for the last four and a half years. This corporation is typical of many modern research corporations in that it is part of a corporate conglomerate of unrelated industries. The mechanics of daily routine are typical of most research corporations. The personnel in a research company respond to governmental requests for proposals, gather the appropriate expertise, and write proposals to compete with other firms for the same research contracts. (See Fetterman 1982b for a detailed description of research corporation life.)

Ethnographers are hired by research companies to provide a qualitative insight into proposed research and to fulfill

From David M. Fetterman, "Guilty Knowledge, Dirty Hands, and Other Ethical Dilemmas: The Hazards of Contract Research," *Human Organization*, 1983, 42(3), 214–224.

research specifications required by the sponsor. Typically, ethnographers are hired to conduct fieldwork for a portion of the study. Participant-observation may range from five years to 100 hours on-site, depending on the nature and priority of the task and on the funding. (See Fetterman 1982a regarding the current role of ethnography in educational contract research.)

The ethical dilemmas explored in this discussion were based primarily on my experiences as a contract ethnographer in four separate evaluation studies: a study of alternative high school programs for dropouts, a study of gifted and talented education, a migrant education study, and an arts education contract.

Entrance into the Field

Entering the field of educational evaluation as an evaluator represents one of the first problems encountered by the ethnographer. The experience is similar to the hazards of entering the field in a foreign culture. The ethnographer must establish him or herself in an unknown and potentially hostile environment. In contract research it is not uncommon for the ethnographer to participate in a series of harrowing interviews paralleling Clinton's description (1975). The next step in this initiation is the corporation's rites of passage. This can range from frustrating methodological discussions to a routine exchange of ritual insults regarding the difference in fields. The last stage, much like a conventional employment experience, involves "proving oneself" in this precarious role as a competent researcher and employee, for example, gathering reliable and valid data, working under pressure, and constructively working with colleagues.[2] This process is evidenced in a portion of an ethnographer's recommendation. According to the director:

> Stanley (pseudonym) was hired by ABC research corporation to work as an ethnographer. . . . He entered a somewhat hostile environment in that most other members of the staff had rather strong quantitative biases and were suspicious of qualitative approaches. Despite this inauspicious set of circumstances, he quickly established himself as a valued member not only of the project team, but of the entire office. He became well liked on a personal basis and well respected as a professional (President, ABC Research Corporation, 1982).

Anthropologists must determine from the onset if their values and temperament are suited to weathering those preliminary challenges. The consequences of being ill prepared or personally unsuited for such a role can be devastating to the profession as well as for the individual.

These trivial but personally draining difficulties are overshadowed by the problems resulting from conflicting expectations with sponsors. Sponsors have become increasingly aware of the strengths and weaknesses of ethnography in evaluation. Many sponsors, however, have been lured by ethnography's reputation for "finding out what's going on," without understanding what it is, or more to the point, what it is not. For example, a request for a proposal may specify the use of ethnography, the proposal may specify the use of ethnographic techniques, and upon award of the contract the project officer may expect a priori closed questionnaire-type interview protocols—with statistical correlations. These expectations may represent useful approaches in other studies; however, these expectations do not meet the realities of ethnographic research. A sponsor's acceptance of a proposal is a binding contract and it marks the formal entrance of the field-worker into evaluation. Ethnographers entering such an agreement must recognize that the two parties may have differing sets of responsibilities and expectations. It is both the ethnographer's and the sponsor's responsibility to resolve these conflicts in a manner that serves each party's pragmatic interests without compromising the methodological integrity of the agreement.

Role Conflicts

A major problem for the anthropologist in the field is being viewed stereotypically as an evaluator (Everhart 1975; Colfer 1976). The stereotypic concept of an evaluator as someone looking for problems or deficiencies effectively blocks many communication channels. Since the ethnographer is interested in finding out how the system works from the insider's perspective, such barriers to communication must be broken down. The problem was illustrated dramatically during a site visit when personnel wouldn't speak to the site visitors, believing them to be spies. Colfer (1976), Clinton (1976), and Thorne (1980) have reported similar experiences.

The anthropologist-evaluator is faced with more than the methodological dilemma of data collection in the field. The ethnographer must function as an intermediary between informants and sponsors, informants and the research corporation, and between informant and informant. One of the most serious ethical dilemmas that emerge from working in this setting is the development of conflicting roles and interests.

> Even in unusually benign instances the field researcher must be very sensitive in his presentation of self and management of social interactions. In most cases, though, the fieldworker encounters social complexities and problems at every turn, and successful role maintenance demands great presence of mind, flexibility, and luck (Pelto 1970:200).

Politics further compounds these role maintenance problems. The ethnographer is required to play many roles in the political context of contract research. These roles confer many responsibilities.

Conducting research in a recent national evaluation illustrated the complexity of these relationships and the diversity of roles required to function in this setting. The ethnographer conducted research in the street, the classroom, student and community members' homes, public schools, the program's local and national disseminating organizations, city governments, the research corporation, the governmental managing agency, and the sponsoring agency. Each of these levels have conflicting groups within each strata—for example, student, teacher, and principal in the school level. As Klockars (1977: 219) explained:

> The problem of conflicting role obligations in biomedical experimentation, where researcher-subject and physician-patient dilemmas arise, has been highly troublesome to attempts to develop ethics for biomedical research. However, such problems do not begin to approach the complexity of conflicts and reciprocal obligations and expectations characteristic of anthropological or life history fieldwork.

It is difficult to maintain a rapport with rival groups unless one establishes oneself as an independent entity sensitive to each party's concerns, and interested in collecting information

from all sides. Taking sides (purposely or inadvertently) early in the research erects barriers to communication with rival groups (see Berreman 1962). First and foremost, however, the anthropologist's responsibility lies with the informant at the center of the research task—in this case the student. The anthropologist must respect the informant's rights and maintain an intricate web of obligations, including confidentiality and reciprocity. The anthropologist must maintain his or her perspective within this convoluted structure and remember that the central informant's rights must take priority according to personal and professional ethical codes, if we are to continue to work with informants, as Mead said, "in an atmosphere of mutual trust and respect" (1969). In addition, this position serves to protect future anthropological endeavors.

This juggling act becomes more difficult with the addition of another party. The ethnographer is also responsible to the taxpayer. Supporting the federal or state bureaucracy (a representative of the taxpayer) is often an unpopular position. An "agency relationship with the state" is created when a researcher accepts government funds. The state assumes both legal and political liability for the actions of the researcher in this relationship. The researcher who enters into a binding contract, in return, has an obligation (contractual and ethical) to fulfill his or her commitment to the sponsor. This includes following the evaluation design of the study (unless amended or modified), pursuing research and presenting findings with the sponsor's interests guiding the research, and being fiscally, administratively, and academically accountable to them. In a Weberian sense, these relationships force one to conclude that "the occupational structure of modern science makes research, ethically speaking, a 'political vocation' " (Klockars 1979:264).

In conventional ethnography, for example, it is not unusual to scratch one's line of inquiry and select another topic and mode of investigation based on informants' information. This usually occurs when the anthropologist is alerted that there is a more pressing or appropriate research concern in the area. In contract research, however, the sponsor and researcher establish the topic and mode of inquiry before entering the field and leave little room for alteration. This is not to say that the study design is cast in stone. Information gathered from field experiences is taken into consideration and may suggest that alternative methods are required to answer the study's policy questions. Field information, no matter how compelling, however, is rarely considered sufficient to drop one's topic of investigation—political pressures are the most powerful force in this regard.

This is not a call for blind obedience or an abdication of one's responsibilities to ensure that research is conducted properly regardless of political pressures. Nor is this discussion aimed at absolving the researcher from a commitment to informants and colleagues. This discussion is presented to stress an obligation that receives little attention at best, and outright condescension at worst.

Fieldwork in the Inner City

Another problem for the anthropologist is urban fieldwork. Fieldwork in the inner city poses many challenges morally and physically. Poverty, powerlessness, political corruption, racial tensions, violent assault, and vandalism represent the backdrop of fieldwork in the inner city. The pressure of these daily activities alone generates considerable personal stress in an urban field-worker. This stress can affect one's judgment regarding data collection and ethical decision making. When researchers are confronted with such activities as police corruption, large-scale drug transactions, burglary, and extortion, they are forced to make serious ethical decisions. These decisions can be guided by a cost or risk-benefit approach (Reynolds 1979:69-84), a respect-for-persons ethic (Mead 1969), or a simple pragmatic manner. A few cases drawn from my urban fieldwork, as well as others', are presented below. The examples are followed by a brief discussion of one or two of the plethora of ethical issues involved in each case.

Guilty Knowledge and Dirty Hands: The Front

During the early period of fieldwork it is important to establish rapport with informants. This involves presenting oneself and one's aims in as honest and direct a manner as possible. In addition, it involves time. The ethnographer must spend time with people, participating in their daily activities, working with them, joking with them, and in some cases, participating in illegal activities. Evaluating a school for dropouts requires an intimate knowledge about dropouts and their activities. Often their activities lend themselves to extralegal and periodically illegal activities. In this regard, I would concur with Polsky's sobering position:

> If one is effectively to study adult criminals in their natural settings, he must make the moral decision that in some way he will break the law himself. He need not be a 'participant' observer and commit the criminal acts under study, yet he has to witness such acts or be taken into confidence about them and not blow the whistle. That is, the investigator has to decide that when necessary he will 'obstruct justice' or have 'guilty knowledge' or be an accessory 'before or after the fact, in the full legal sense of those terms' (1967:139-140).

The following is a case in point.

I became close friends with one of the students participating in the school under study. When I first met him he divided his life into two worlds—school and the street. The latter dominated his interests and activities. He was proud of his ability to thrive in the street. He had a "reputation" and was respected in his community. One of the most important elements of cultural knowledge in the street is knowing where to "cop dope." Illegal drugs and various other commodities are exchanged in the inner city in places called "fronts." Fronts are stores that sell legitimate goods such as records, health foods, shirts, and so on, as a front for routine illegal drug transactions.

My friend introduced me to this element of street life with a "hands on" approach. He invited me out for a bite to eat after school. We walked down the main street of the inner city for a few blocks until he pointed to a health food store. He said he thought that I would want to eat there since I was from California. We entered the establishment and my friend asked the clerk to give me a granola bar. I said thanks and I reached for the bar. The patron handed it to me with a smile and a small envelope underneath it. I looked down at a "nickel" bag of marijuana. My first reaction was "how am I going to look and how is ethnography going to look to my company if their eth-

nographer is busted for drug possession on one of his first site visits." My discomfort was compounded by two policemen walking by viewing the exchange. The policemen saw the transaction, smiled and continued walking. When I asked my friend why they didn't bust us he said, "they don't need the money right now." I asked him to clarify his response and he explained:

> They only bust you if they need the money. They get paid off regular. But if they're hurtin' for money then well, that's another different story. They'll come right in and bust ya, take money out of the cash register and take your dope too. If they're on a run and they gotta show that they mean business then they'll bust your ass. Otherwise they just look the other way.

My informant's words echoed a modern version of William Foote Whyte's racketeer in his classic *Street Corner Society:*

> The cops are paid off. They call it the "union wage." The patrolman gets five dollars a month for every store on his beat that sells numbers. The plain clothes men get the same, but they can go anywhere in Cornerville (1943:123).

After being initiated by this brief encounter with criminal activities and official corruption, I continued to learn a great deal about the environmental pressures that affect the dropouts' behavior. My conclusions paralleled Whyte's when he reported:

> Observation of the situation in Cornerville indicates that the primary function of the police department is not the enforcement of the law but the regulation of illegal activities (1943:138).

Moreover, I was faced with a number of ethical dilemmas involving guilty knowledge (incriminating information made privy to the field-worker) and dirty hands (a situation from which the field-worker can not emerge innocent of wrongdoing), which required a series of immediate decisions. First, a researcher must decide whether the research merits involvement in criminal activities. Students of deviant behavior must discriminate among the range of activities involved and decide which specific activities justify their involvement. These preliminary considerations are routinely based on a utilitarian ethic:' Do the ends justify the means? For example, the "in the name of science" position would argue that the insights gained during this involvement contributed to knowledge, which outweighs the short-term legal and moral transgressions. Soloway and Walters (1977:171-172) described a fieldwork episode in which one of them was made an unwitting party to the execution of an armed robbery. This behavior was considered too cavalier for some colleagues; however, his research was "breaking invaluable ethnographic ground" in the study of heroin addicts. The pursuit of research, however, is not above the law.' The researcher must be willing to suffer the consequences of such involvement. Personally, the researcher must balance the potential significance of the research against the severity of the criminal behavior involved. This is a useful guide in moral decision making. Alone, however, this overly rationalistic risk-benefit approach is at best off target when making moral decisions in the field.

A second question that emerges from this experience is, What is the researcher's civic responsibility after observing or inadvertantly being involved in criminal behavior? In this case, three illegal acts were involved: selling illegal drugs, buying illegal drugs, and police corruption. The researcher technically has a conflicting set of responsibilities to the student in this case. This student is a former dropout who has reentered "the system." Condoning his behavior in this case may represent a criminal type of benign neglect (see Yablonsky (1965). Protecting the student from himself, however, is condescending at best and a breach of confidence at worst. Similarly, although the researcher has an obligation as a citizen to report illegal activities, informing on the drug dealer involved would have constituted a breach of confidence—and posed a considerable threat to one's plans for an extended longevity. Moreover, the researcher must acknowledge that like prohibition, the punishment may not fit the crime and a form of nonviolent civil disobedience may be appropriate. Fundamentally, however, the respect-for-persons position overrides all of these risk-benefit considerations. The respect-for-persons position is essentially a code that "holds that there are certain means which are deontologically repulsive and in se wrong"" (Klockars 1979: 267). In this case, a breach of confidence would constitute "deontologically repulsive and in se wrong" means.

Finally, the third act, police corruption, is of some significance. The idea of tackling such a problem may appear insurmountable. In addition, like the other acts, it is at least outside the scope of the work that must be accomplished in a short period of time. It is important to explore such important variables in the social equation; however, the researcher must maintain some boundaries on the research endeavor if the task is to be completed.

Risk-benefit analysis, respect-for-persons ethic, and basic pragmatism are all appropriate approaches that must be taken into consideration when making moral decisions during fieldwork. As Klockars has stated, however,

> the good end of the dirty means . . . is not the long term good of science, nor the potential value of the particular research at hand, and certainly not the worldly benefits continuation of that research may have for the researcher's career. It is the immediate, morally unquestionable, and compelling good end of keeping one's promise to one's subjects. In particular, it is the keeping of that minimal promise which every fieldworker makes explicit or implies to deviant subjects in the process of gaining first-hand access to their deviance (1979:275-276).

My response to this experience was not to intervene. I recorded the event in detail to provide background material regarding the various inner-city pressures operating on the students. I chose this route because it was early in the research endeavor and much more information was needed to understand how the community operated, and how my actions would affect all participants. In addition, a more active role would have constituted a breach of confidence—a confidence which in and of itself I was obligated to uphold. In turn, this breach would have served as a barrier to all future communications. This parallels Pelto's position that

> any interference by the fieldworker [in this type of situation] would mean that he would have to violate the confidences of his informants, and this would seriously jeopardize his work (1970:222).

A description of police corruption, however, was printed in the report to provide the environmental context. The matter was also discussed with city officials on the researcher's own time. I temporarily separated research from activism. My reactions were based on timing, a trust, a professional responsibility to respect the environmental norms or rules and regulations until the dynamics were understood, and a responsibility

to complete my objectives.' (See Beattie 1965; Klochars 1974; and Wax 1971 for discussion of similar guilty knowledge experiences.)

Dirty Hands and Guilty Knowledge: Burglary and Extortion

Urban fieldwork requires both direct and indirect involvement with criminals. Polsky explained that

> in doing field research on criminals you damned well better *not* pretend to be "one of them" because they will test this claim out . . . [moreover,] before you tell a criminal who you are and make it stick, you have to know this yourself, know where you draw the line between you and him (1967:124–125).

During one of my site visits to these alternative high schools for dropouts my car was burglarized and my clothes and notes stolen. The burglars then attempted to sell me my stolen possessions. In this case the line was easy to draw between the researcher and the criminal because the criminals were neither acquaintances nor participants in the study. They were simply criminals. This episode provided another case for intervention in the field (Gallin 1959; Gallagher 1964; Gearing 1973; Holmberg 1958; McCurdy 1976; Spradley 1976). The potential for producing deleterious results has been documented (Horowitz 1965; Sahlins 1967; Berreman 1969; Holmberg 1954); however, this instance illustrates how intervention with dirty hands can provide useful data for the research endeavor.

The event began at the end of a long day of interviewing at the school. I had just completed an extra interview to get ahead of my self-imposed timeline for the week and was satisfied with the week's work. I said goodbye to everyone for the day and walked down the block to my car. The window was broken, the battery removed, my suitcase and my briefcase stolen. My briefcase contained my notes and slides of my work for the two preceding weeks on site, as well as a completed paper to be presented at a professional meeting and a paper in progress.

I was stunned. Neighbors in the community who knew me came out of their houses to see the damage. One woman said her daughter had seen the burglars: two young men who had "been terrorizing the neighborhood for months." I asked for their names and Corina (pseudonym) declined to respond, explaining:

> My kids, they go to that school. They would be put in danger. I try to run a good Christian home but I'm afraid of the revenge for my girls. They could get hurt by the other kids. You know.

I explained that I understood. I called for assistance from the neighborhood grocery store. No cab would come to the area so I had to wait for the tow truck to pick up the car and take me out of this part of the city. I stayed with the car to protect it from the car parts "vultures" until dark. Corina invited me into her house at dusk explaining, "It gets bad at night, especially since you're White and all. You'd be safer in here with us till the man comes." I immediately accepted her invitation and we talked about the community for a few hours. She explained how these "thugs" had held a gun to her friend's head and stolen her stereo. She explained:

> They had the gall to do that and tell her when she got another one they'd be back for that one. A year later, sure enough. She moved about a block away and they came back and stuck a gun to her head again and said it wouldn't be the last time.

Corina also told me about arson-for-hire in the neighborhood. She told me about the time she

> woke up to a phone call at two in the morning. The man over the phone said to be out of the house in 15 minutes because it was going to burn. That's what they do when it's arson, they call you just like that at two in the morning. I had my rollers on and I was in my bathrobe, that's all I had. I was on the second floor and my grandma she was on the third. I can still remember seein' the flames all around her in her wheelchair. I tried to get her out but I couldn't. I had rheumatic fever, you know, so I'm weak. She was so heavy I just couldn't. I got my babies out but she was so heavy. I just watched her die. I still go to the county [psychiatrist] even now. I dream about it. It still frightens me. I couldn't save her.

Her moving story was cut short by the arrival of the tow truck. The burglary experience had already provided an opportunity to learn more about the community and develop a rapport with another member of the community. I met with Corina the next day to continue our talk about life in the inner city. She said she would be willing to serve as a mediator between the young men (burglars) and myself. She knew their mother from the PTA and agreed to meet with her to "rescue" my papers. Corina and her husband frequented one of the burglar's homes in an attempt to come to an agreement regarding my materials. During negotiations, however, one of the little girls in the burglar's home opened a curtain dividing the rooms and Corina's husband saw his color television set—stolen from them six months before. They forgot about my problem and "blew up" at the mother for condoning this behavior. Needless to say neither their television nor my materials were recovered. That evening, however, Corina volunteered to serve as a witness if I wanted to go to the police. I thanked her and told her I would have to ask a few other people in the community before contacting the police. I had been in the community off and on for over a year and a half, and feared police reprisals if the police were asked to become involved. I discussed the matter with neighbors, community action groups, members of clergy, and city officials directly associated with the community before taking any action. They unanimously agreed that "something must be done to stop these punks from having the run of the community." They suggested that I involve the police and I agreed. I contacted the police and their first response was "forget it . . . it will just end with a bullet anyway." I later learned that burglary was a low priority in an area where murder, rape, and arson were the norm. Later they said that if I felt it was necessary, I should pursue it myself.

One of the burglars then contacted Corina and told her he was willing "to negotiate" with me. I was told to wait in the school at night until he called. I observed much about night life in the inner city while I waited for his call. A crowd of young men drinking and smoking gathered outside the school, growing and dwindling in size and volume all night long. I had to check on my locked (replacement) car every 15 minutes to prevent it from being stripped to the frame. A well-dressed young man in a new Cadillac, however, did not have the same concerns. He parked his car in the middle of the street with the motor running and the radio playing loudly, while he disappeared into the darkness of the school playground with a small box under his arm. He came back empty handed 15 minutes later and drove away. No one had touched his car.

Later I learned that he was known by everyone in the community and "no one crosses the man." I was introduced to him later in the study and learned that he ran the "underworld" portion of the community. This experience provided numerous insights into the students and dropouts in the community. The burglar finally called that night and offered to sell me my materials at $15 a folder (20 folders). I agreed on a trial basis: 1 folder at a time. Corina served as "go between." The venture failed. The burglar took the money and kept the folder. We set up another series of phone negotiations to recover the goods, also unsuccessful. I eventually called them and told them I knew who they were and where they lived and if the materials were not returned in an hour I would call the police.

I waited for two hours—no response. I called the police and explained that I had decided to prosecute. They said they would not go in at night and would pursue the matter in the morning. In brief, I had to orchestrate the entire event: secure the deposition from the witness, find the exact location of the burglars, and bring the police to the location. The burglars were arrested and prosecuted with the "blessing of the community." During the booking proceedings, when the police officers left the room for a minute, one of the burglars leaned over and whispered to me "we've heard about what you're doin' and we know that you're trying' to help the brother so we'll try to get the book stuff back to you after this is all over."

They later explained that they wanted to get busted. One of them said:

> We're hot now, ya see. So if we just chill out for a month or two somebody else is in the spotlight, ya see. And then we can go along with our business with no more trouble. There's just a little too much heat on us right now, don't ya know.

I am still negotiating with them, however. I do not anticipate recovering the goods.

This experience demonstrates that intervention can yield positive results and what steps were required before such behavior is appropriate. "The dilemma of the fieldworker . . . is not *whether* to interfere in the local cultural scene, but *how much* to interfere" (Pelto 1970:223). This experience required intensive involvement. I had been in the field long enough to know the members of the community in depth. In addition, I understood most of the repercussions resulting from police involvement in community affairs. Moreover, I consulted with various members of the community, such as neighbors, clergy, city officials, before making a decision to intervene. I also took a series of time-consuming and potentially hazardous steps toward resolving the matter by negotiating directly with the burglars. My final decision was my own; however, it was influenced by these sources of information and approval in the community.

The decision to have the burglars arrested was required after discussing the matter with (and receiving the "go-ahead" from) various community leaders to fulfill my citizen-obligation as a special guest-member of the community. It may appear odd to sound apologetic for having burglars arrested; however, had the "hard core"" burglars been the focus of my study (with explicit or implicit trust established) these same actions would have been inappropriate, if not immoral. Pragmatically, I wanted my notes and slides back and I had taken all of the conceivable steps required short of this final decision. The risk-benefit approach was inconsequential at this point given that the portion of the community involved in the study had decided to risk any retaliation for the "greater good of the community." The respect-for-persons ethic was inappropriate to apply to the burglars given that no bond of trust had been established with them. The respect-for-persons ethic was applicable, however, to the traditional segment of the community, given that a strong bond had been created with religious and social leaders, teachers, students, and various families in the community. The decision appeared logical and appropriate; however, there are "no hard and fast rules to be laid down [for these types of moral dilemmas in fieldwork]; these are matters of conscience rather than of science" (Beattie 1965:55).

At a recent professional meeting, I was asked whether I thought there was an ethical problem regarding the use of my uncensored fieldnotes by outsiders in this case. I explained:

> If you had asked me what I thought if I had delivered uncoded, uncensored fieldnotes to the federal government I would agree there would have been a problem of breach of trust or confidence. However, the case of fieldnotes being stolen during fieldwork from a locked car is another matter. Given the fact that the notes were stolen, not deliberately disseminated, the fact that the burglars had no use for the materials (exept extortion) and the lengths I went to retrieve the notes, I do not feel that an ethical dilemma exists regarding this facet of the incident.

The experience of being burglarized and extorted provided me with an insight into the turbulence that most of the neighbors experience daily. Moreover, deciding to take an activist role extended my understanding of the community simply by expanding my contact with the community. Intervention in this case provided a number of extremely important data bases that were tapped throughout the study. The staff and students in the school were upset about the experience and generously offered their assistance. A number of students with street contacts helped me to identify the location of the burglars. The positive reaction of the staff and students in the school to this dilemma served to strengthen rapport. The cost of these insights, however, was extremely high. The cost is human suffering, which "is the lowest price that decent human beings must be willing to pay in order that they stay competent at the vocations of policework and fieldwork" (Klockars 1979: 277).

The Ethnographic Report

Fieldwork conducted in highly political settings can be more dangerous than fieldwork in the streets of the inner city (Diamond 1964; Peattie 1968). One of the most common mediums for interaction in the political realm is the report. An ethnographic report rich in detail is as potentially dangerous as it may be helpful, depending on how the material is presented and who uses the information. Tobin's Ph.D. dissertation, for example, "The Resettlement of the Enewetak People: A Study of a Displaced Community in the Marshall Islands" (1976), represents a classic case of misused information. Tobin's study was used by the Air Force as a resource document for preparing a misleading environmental impact statement regarding the Pacific Cratering Experiments (PACE) project. This area was

the site of numerous nuclear tests. The PACE project planned to use this area for further high-explosive testing and used parts of Tobin's work to support their position. Tobin responded,

> I did not give you permission to do this and it is protected by copyright as clearly indicated in the early part of my dissertation. Parts of this work that would have helped the people of Eniwetok against the PACE program were not quoted in the draft environmental statement.
>
> I am biased against the PACE program as I have told Mr. _____ [the director of PACE] as I feel it is against the best interests of the Eniwetok people and it is against their expressed wishes (Department of the Air Force, 1973:56).

The ethnographer's moral obligation, in this example, required a written response to protect the interests of the people of Eniwetok and the use of his own publication.

Serious ethical dilemmas emerge, however, when one's role makes one privy to confidential information that requires exposure. Ibsen's *An Enemy of the People* (1959), Solzhenitsyn's *For the Good of the Cause* (1972), and Daniel Ellsberg's *Papers on the War* (Pentagon Papers) (1972) dramatically illustrate this type of double bind. In one of my studies, this type of double bind was confronted on every level. A few of those encountered in the street have already been discussed. The school setting provided numerous cases. For example, substituting for a sick teacher presented no serious difficulty; however, substituting for a frequently tardy or alcoholic teacher presented a number of difficulties. Should the researcher condone such behavior and administrative laxness by substituting for the teacher and not reporting the incident in his or her report? Or, should the researcher simply look at the practical side—the students need a teacher for that period? From a research perspective, serving as a teacher-researcher provides an invaluable insight into the program. Moreover, the problem of managerial laxness can be demonstrated in other manners. In this case, a risk-benefit approach was extremely useful in moral decision making. The risks of reporting the incident for the individual teacher's reputation and the program's survival outweighed the benefits, given that the matter could be resolved with less drastic measures (informally bringing the problem to the attention of the school administrator). The matter would have required publication if administration had not resolved the problem immediately, because the risk to the student population (of dropping out again) and to the staff (lowering morale) would have been greater than the benefits of protecting one teacher and administrator's positions. Discretion, in any case, must be exercised in the case of reporting observed indiscretions. For example, reporting a rare occurrence such as a fight or an affair between a student and a staff member on school grounds can unfairly distort a picture of program operations. Moreover, the consequences of reporting such behavior "may not match the crime"; for example, the entire program could be closed down for such activities. (See Deloria 1980 for a discussion of a larger social context of research and role of researcher.)

Another problem that must be confronted is the power of numerous vested interests. The pressures of various vested interest groups often impinge on the ethnographer's ability to produce a fair and balanced report of study findings. For example, in the study discussed above the staff wanted me to record and document the implementation difficulties in the report as a means of solving their programmatic problems. The disseminators, however, took a different position. They commented on a draft of one of the reports that the ethnographic study was "a scholarly approach," however, they were concerned with the presentation of the findings.

> Certainly, [the disseminating agency] has gleaned a great deal of knowledge during the demonstration which we are applying to future replication approaches. [The research corporation] has been very helpful in this regard. However, we are down to the wire in terms of the presentation of the final results to society at large. Certainly, [the disseminating agency] has a vested interest in the [program] being presented in the final reports in the best possible light. I am sure that others such as _____, _____ [federal agencies], and [the research corporation] feel the same. . . . [Program]expansion in the future faces an uncertain future in this age of shrinking financial resources and competitive and political realities, etc. We need to present the most accurate, fair, and balanced picture of the replication which, hopefully, proves that [the program] merits continuation and expansion. I trust that you will consider the same.

Their message was clear. I was sympathetic to the political realities; however, I was obligated to include some negative findings to present the most accurate picture of program operations. For example, along with numerous positive findings I included serious implementation problems such as high staff turnover rates and managerial incompetence and/or lack of appropriate qualifications. These problems had a serious impact on program operations. The negative impact of the federal government and the evaluators was also discussed to provide a picture of the extrinsic forces that negatively affected the program and resulted in unfavorable site descriptions (Fetterman 1981a, 1981c). This was an example of "studying up" in the stratification system (Nader 1969). Ignoring these problems would have done little for knowledge development in the area of implementation and distorted the readers' view of program operations. This would have represented an abdication of my responsibility to the staff, taxpayers, and my colleagues. A basic misconception that was dispelled in this regard is that ethnographers are always co-opted by their informants and always present the most positive side (their key informant's side). The duty of the ethnographer, like any scientist, is fundamentally to accurately record and report his or her observations and interpretations. In this case, the observations were primarily positive but the findings were not exclusively placed in a positive light.

Dissemination of Findings

The dissemination of the draft report was also problematic. The code of ethics explains that the findings of research must be shared with clients and sponsors. This guide, however, does not prepare the researcher for dealing with many levels of administration and protocol. In the study under discussion there was a rivalry between the parent organization disseminating the alternative high school program and some of the local affiliates directly responsible for managing the programs. The parent organization was the central conduit for draft reports. The evaluators were informed, however, that one site would not receive the draft for comments because they had new management and staff and would be demoralized by the descriptions of past strife. In addition, the new program would not have the background required to critique the work. The evaluators were also informed that another site would not

receive the report, according to the parent organization, because they misused it the last time; they revealed portions of the confidential draft report to various sources out of context. In the first case, it was true that the report referred to the old staff and would not have been productive reading for the new staff. In the second case, the evaluators would have fed the fire of this rivalry if it were to circumvent the system of protocol by sending the drafts to the sites directly. However, they would not be fulfilling their obligation if they allowed the parent organization to control the distribution of the report.

A compromise was made. All the copies were sent to the parent organization to follow protocol and avoid charges of favoritism. A provision was made, however, that site comments would be requested directly by the evaluators by the end of the month. Any report lost in the mail would then be sent directly to the site by the evaluators. This placed a check on the distribution of the drafts without compromising the evaluator's role or neglecting the significance of protocol.

The presentation of findings to the public is a political activity. The manner in which research findings are presented influence how the information will be used or abused. The researcher who plays the role of politician while conducting and presenting findings, however, is likely to be used as a pawn by various vested interests. The dissemination of findings after the research has been conducted is a separate matter. The evaluators disseminated the generally positive findings to appropriate individuals in government and quasi-governmental institutions. Future funding for the program was dependent on the dissemination of the evaluation findings and the recommendations of various agencies. In addition, the evaluators prepared a Joint Dissemination Review Panel Submission that was substantially based on the ethnographic findings to improve the program's credibility and potential to secure future funding. (This was accomplished in the face of significant resistance because it was politically hazardous to favor social programs during this period.) These actions were in accord with Mills's position:

> There is no necessity for working social scientists to allow the potential meaning of their work to be shaped by the "accidents" of its setting, or its use to be determined by the purposes of other men. It is quite within their powers to discuss its meanings and decide upon its uses as matters of their own policy (1959:177).

The evaluators agreed that they had a moral responsibility to serve as an advocate for the program based on the research findings. As James has discussed: "Advocacy on behalf of social change is the final step in the use of ethnography. It is also the only reasonable justification for probing the life-styles of these human beings" (1977:198). There is a difference between being an academic and an activist; however, academic study does not preclude advocacy. In fact, often anything less represents an abdication of one's responsibility as a social scientist (see Berreman 1968; and Gough 1968). It should be acknowledged, however, that the researcher functions as a public relations person or politician in this arena rather than as a researcher.

Job Stress and Burnout

Finally, the ethnographic evaluator faces one of the most common but least discussed hazards in the profession: job stress and burnout. The job-related stress that an ethnographic evaluator or field-worker experiences has been discussed throughout this review. Job burnout involves the complete loss of interest or motivation in pursuing the individual employment tasks required to satisfactorily function in one's role. This is often the result of prolonged exposure to the pressures of the job. This can severely cripple the most able researcher. Judgment, determination, and stamina (all critical qualities for a field-worker) are all affected by job stress and burnout. Fieldwork in contract ethnography must be conducted at an accelerated pace in a much shorter period of time than traditional fieldwork. This is both physically and mentally demanding. Continuous immersion in the personal and professional problems of informants can be emotionally draining as well. Stories of arson-for-hire, a mother stabbing her daughter's boyfriend, an administrator harassing a staff member, graft, and racism are part of the everyday lives of many informants; however, this continual immersion into hundreds of individual lives can take its toll on the ethnographer. Wax (1971) provided a detailed picture in this regard of "shooting, beating and murder" and the resultant turmoil she experienced in a Japanese-American relocation center. Kobben reported of his Surinam fieldwork that

> since an ethnographer studies people and not insects, his fieldwork also causes emotions in himself. Personally, I lived under great psychological stress and felt little of the proverbial peacefulness of "country life." Few books touch on the subject, but I know that the same is true of quite a number of other fieldworkers. Perhaps it is even a *sine qua non* for fieldwork (1967:46).

The theory, research, and intervention practices related to job stress and burnout in human services occupations are discussed in detail in Cherniss (1980) and Paine (1982).

This experience is compounded by the "father confessor" or mea culpa compression effect. Contract research requires in-depth immersion in a site for short periods of time at regular intervals throughout the year. Informants realize the ethnographer will only be on-site for a week or two and rush to communicate pressing problems. The nature of the visits structures the informant's response. An effort must be made to take this phenomenon into consideration—to balance one's perspective of the site's operations. Once a rapport is established with a few key informants and the ethnographer learns who must be listened to with a grain of salt this problem can be ameliorated.

The fieldwork experience is made more stressful by a demanding travel schedule. One- to two-week site visits throughout the country can keep a researcher away from home for over a month at a time. Life on the road has all the hazards faced by old-time salesmen: road food, empty motels, and the routine separation from family—in this case every three months. Allan Holmberg (1969) provided vivid illustrations of the physically draining side of fieldwork (also see Wax 1960: 175). A few survival tips learned in the field to cope with this type of stress include maintaining regular contacts with family, spending time with friends in the field in relaxing or entertaining settings, or meeting relatives or colleagues during weekends or "break periods" while on the road. Also, attending professional meetings during these free periods serves to recharge oneself while in the field. Pelto emphasizes the value of brief vacations during the fieldwork experience.

> A number of fieldworkers have noted that brief vacations away from

the research community can be excellent tension relievers—for both informants and researchers. After all, at least in small communities the ubiquitous presence of "the man with the notebook and a thousand questions" can be very taxing for the local inhabitants. They must surely wish that for once they could enact a small bit of local custom without having to explain it all to the anthropologist. A few days away—or even longer—in the city, at the beach, hiking in the mountains, or visiting a nearby game reservation—can give the fieldworker time to dissipate his anxieties and hostilities, get some needed physical rest, and perhaps restock his supplies. At the same time, the research community itself gets a rest. Often the return of the fieldworker after even a brief vacation is an occasion for a warm welcome, a reaffirmation of friendships. He may be treated like a returning relative, and a few slightly reluctant informants may have been opened up a bit in their willingness to give information (Pelto 1970:225).

One of the few redeeming features of this work life-style, aside from meeting new people, is that it enables you to step back from the field experience to gain perspective and then back in to test one's hypotheses throughout the year. This is an advantage over traditional fieldwork where it is much easier to "go native," or lose touch with the primary research task at hand.

Conclusion

Moral decision making is a tortuous process, since each event is a convoluted and almost endless labyrinth of considerations and commitments. A simple shift in perspective or an unexpected twist of fate can alter one's entire set of responsibilities and obligations. Guilty knowledge and dirty hands are at the heart of the urban fieldwork experience. Recognition of this fact is essential if a field-worker is to function effectively and morally. Awareness of the context of research can prevent paralysis as well as overzealousness in the field.

Ethical decisions in fieldwork must continuously be discussed and reviewed. This is not to suggest that we must institute sanctions against ethical wrongdoing, for "the cost of emphasizing punishment as a means of regulation and control of occupational deviance is that it suppresses the kind of candid moral discourse which is necessary to make genuine moral maturity possible" (Klockars 1979:279).

Field-workers will continue to encounter numerous personal and professional hazards in contract research. They may range from fieldwork conducted in an accelerated fashion to reporting in a highly political atmosphere. Many of these pressures affect one's judgment while in the field—whether in the streets of the inner city or in plush conference rooms with governmental officials in Washington, D.C. Ethnographers can adapt to most of these environmental pressures if they are aware of them.

> There have been few times in the past century when it has been so important for fieldworkers to involve themselves in processes of ethical decision making. As we do so, we are well advised to temper our instincts for self-preservation and self-determination with a realistic sense of the full range of contexts which impinge on contemporary research activities. Two seemingly opposite images come to mind. The first is an image of a world breathing down our necks, and the second is an image of a world ignoring us entirely (Chambers 1980:341).

Participation in the art of moral decision making may not prevent the world from "breathing down our necks" or from "ignoring us," but it will ensure that we do not forget our own multiple sets of responsibilities.

To improve the level of fieldwork practice, investigators must examine the moral dilemmas particular to this type of research, discover the appropriate ethical principles, and learn how best to apply them. If it is not done, regulation will become an elaborate and expensive charade, useful only in assuaging the sensibilities of legislators, who can convince themselves that they did their best to legislate morality without ever having bothered to examine just what moral standards are appropriate to a particular scientific method (Cassell 1980:38).

This exploration into the hazards and ethical dilemmas that arise from urban fieldwork and contract research has attempted to examine the appropriateness of certain moral standards to the ethnographic method. It is hoped that this probing will be reflexive, stimulating other field-workers in anthropology and other disciplines to examine themselves in their pursuit of knowledge.

NOTES

[1] For further details regarding the role of the ethnographer in educational evaluation, see Britan 1977, 1978; Burns 1975, 1978; Clinton 1975, 1976; Colfer 1976; Coward 1976; Everhart 1975; Fetterman 1980, 1981a, 1981b, 1981c; Firestone 1975; Fitzsimmons 1975; Hall 1978; Hord 1978; Mulhauser 1975; and Wolcott 1980.

[2] It should be emphasized that this involves working with colleagues from different disciplines and potentially conflicting paradigms in a multidisciplinary effort.

[3] Weber's "Politics as a Vocation" (1946) is a study of the moral hazards of a political career. It emphasizes the use of morally dubious means in the attainment of "good ends." The parallel between the context of contemporary research and the political environment that Weber discussed highlights this moral hazard for contract research.

[4] Weber's term was an "ethic of responsibility" (1946:120).

[5] In the Soloway and Walters case no law was broken, according to the Pennsylvania penal code (see Soloway and Walters 1977:172–174). The moral issue remains, and in other states the legal status of the event might differ significantly. It is inappropriate, however, to second-guess the legitimacy of a field-worker's actions in hindsight. There are a multitude of factors influencing behavior in the field at any given moment. Moreover, serendipity more closely characterizes even the most diligent efforts at structuring ethnography. Soloway and Walter's case indirectly emphasized the unpredictability of fieldwork.

[6] The respect-for-persons ethic is usually applied to situations in which a researcher is contemplating deceit in order to secure information from a subject. The respect-for-persons ethic can also be applied to situations in which the researcher considers breaching a trust. These two examples demonstrate the role of "different levels of analysis" in ethical decision making.

[7] This experience differs from what Wax describes as "when the fieldworker's overblown sense of his ability to offend or injure his hosts may so paralyze him that he cannot carry on his work"(1971:274). This type of problem can occur at the early stages of fieldwork when the ethnographer is overly sensitive to informants. Pauline Kael's solution, as noted in Wax (1971) is useful in this regard, "a mistake in judgment is not necessarily fatal, but that too much anxiety about judgment is." Nevertheless, although there are similarities of inaction, the problem Wax describes is more of a methodological problem related to the early stages of fieldwork, while the problem discussed in this review is an ethical problem related to the respect-for-persons ethic in the process of conducting fieldwork.

[8] In the study under discussion, most of the students involved in crime were involved in dope dealing, pimping, and petty theft; few were involved in "hard core" burglary. The "hard core" group was known in the community to have its own rules, sanctions, and social structure. This experience signaled to the "hard core" group what my role and position was regarding the burglary group in the community.

The experience also provided an insight into who the program could and could not serve in the inner city.

REFERENCES CITED

Beattie, J.
1965 Understanding an African Kingdom: Banyoro. New York: Holt, Rinehart & Winston.

Becker, H. S.
1976 Whose Side Are We On? Social Problems 14(3):239-247.

Berreman, G.
1962 Behind Many Masks. Monograph 4. Ithaca, N.Y.: Society for Applied Anthropology.
1968 Is Anthropology Alive? Social Responsibility in Social Anthropology. Current Anthropology 9:391-396.
1969 Academic Colonialism: Not So Innocent Abroad. The Nation, November 10.

Britan, G. M.
1977 Public Policy and Innovation: An Ethnographic Evaluation of the Experimental Technology Incentives Program. Washington, D.C.: National Academy of Sciences.
1978 The Place of Anthropology in Program Evaluation. Anthropological Quarterly 51(2):119-128.

Burns, A.
1975 An Anthropologist at Work. Anthropology and Education Quarterly 6(4):28-34.
1978 On the Ethnographic Process in Anthropology and Education. Anthropology and Education Quarterly 9(4):18-34.

Cassell, J.
1980 Ethical Principles for Conducting Fieldwork. American Anthropologist 82(1):28-41.

Chambers, E.
1980 Fieldwork and the Law: New Context for Ethical Decision Making. Social Problems 27(3):330-341.

Cherniss, C.
1980 Staff Burnout: Job Stress in the Human Services, Vol. 2. Beverly Hills, Calif.: Sage.

Clinton, C. A.
1975 The Anthropologist as Hired Hand. Human Organization 34:197-204.
1976 On Bargaining with the Devil: Contract Ethnography and Accountability in Fieldwork. Anthropology and Education Quarterly 8:25-29.

Colfer, C. J.
1976 Rights, Responsibilities, and Reports: An Ethical Dilemma in Contract Research. *In* Ethics and Anthropology. M. A. Rynkiewich and J. P. Spradley, eds. Pp. 32-46. New York: Wiley.

Coward, R.
1976 The Involvement of Anthropologists in Contract Evaluations: The Federal Perspective. Anthropology and Education Quarterly 7:12-16.

Deloria, V.
1980 Our New Research Society: Some Warnings to Social Scientists. Social Problems 27(3):265-271.

Diamond, S.
1964 Nigerian Discovery: The Politics of Fieldwork. *In* Reflections on Community Studies. Vidick, Bensman and Stern, eds. Pp. 119-154. New York: Wiley.

Ellsberg, D.
1972 Papers on the War (Pentagon Papers). New York: Simon & Schuster.

Everhart, R. B.
1975 Problems of Doing Fieldwork in Educational Evaluation. Human Organization 34(3):183-196.

Fetterman, D. M.
1980 Ethnographic Techniques in Educational Evaluation: An Illustration. Journal of Thought, Special Ed. (December): 31-48.
1981a Study of the Career Intern Program. Final Report. Task C: Program Dynamics: Structure, Function and Interrelationships. Mountain View, Calif.: RMC Research Corporation.
1981b New Perils for the Contract Ethnographer. Anthropology and Education Quarterly 12:71-83.
1981c Blaming the Victim: The Problem of Evaluation Design, Federal Involvement, and Reinforcing World Views in Education. Human Organization 40:67-77.
1982a Ethnography in Educational Research: The Dynamics of Diffusion. Educational Researcher (March):17-29.
1982b Ibsen's Baths: Reactivity and Insensitivity (A Misapplication of the Treatment-Control Design in a National Evaluation). Educational Evaluation and Policy Analysis 4(3):261-279.

Firestone, W. A.
1975 Educational Field Research in a "Contract Shop." American Educational Research Association, Division Generator 5(3): 3-11.

Fitzsimmons, S. J.
1975 The Anthropologist in a Strange Land. Human Organization 34(2):183-196.

Gallagher, A., Jr.
1964 The Role of the Advocate and Directed Change. *In* Media and Educational Innovations. W. C. Meierhenry, ed. Lincoln: University of Nebraska Press.

Gallin, B.
1959 A Case for Intervention in the Field. Human Organization 18(3):140-144.

Gearing, F.
1973 The Strategy of the Fox Project. *In* To See Ourselves: Anthropology and Modern Social Issues. Pp. 438-441. Glenview, Ill.: Scott, Foresman.

Gough, K.
1968 World Revolution and the Science of Man. *In* The Dissenting Academy. T. Roszak, ed. Pp. 135-158. New York: Random House.

Gouldner, A. W.
1968 The Sociologist as Partisan: Sociology and the Welfare State. American Sociologist 3(1):103-116.

Hall, G.
1978 Ethnographers and Ethnographic Data, An Iceberg of the First Order for the Research Manager. Austin: Research and Development Center for Teacher Education, University of Texas.

Holmberg, A.
1954 Adventures in Culture Change. *In* Method and Perspective in Anthropology. R. F. Spencer, ed. Pp. 103-116. Minneapolis: University of Minnesota Press.
1958 The Research and Development Approach to the Study of Change. Human Organization 17(1):12-16.
1969 Nomads of the Long Bow: The Survivors of Eastern Bolivia. Garden City, N.Y.: The Natural History Press.

Hord, S.
1978 Under the Eye of the Ethnographer: Reactions and Perceptions of the Observed. Austin: Research and Development Center for Teacher Education, University of Texas.

Horowitz, J.
1965 The Life and Death of Project Camelot. Trans-action (December).

Ibsen, H.
1959 An Enemy of the People. *In* Four Great Plays by Henrik Ibsen. Pp. 129-216. New York: Dutton.

James, J.
1977 Ethnography and Social Problems. *In* Street Ethnography: Selected Studies of Crime and Drug Use in Natural Settings. R. S. Weppner, ed. Pp. 179-200. Beverly Hills, Calif.: Sage.

Kiste, R.
1976 The People of Enewetak Atoll vs. the U.S. Department of Defense. *In* Ethics and Anthropology. M. Rynkiewich and J. Spradley, eds. Pp. 74-75. New York: Wiley.

Klockars, C. B.
1977 Field Ethics for the Life History. *In* Street Ethnography: Selected Studies of Crime and Drug Use in Natural Settings. R. S. Weppner, ed. Pp. 210-226. Beverly Hills, Calif.: Sage.
1979 Dirty Hands and Deviant Subjects. *In* Deviance and Decency: The Ethics of Research with Human Subjects. C. B. Klockars and F. W. O'Connor, eds. Pp. 261-282. Beverly Hills, Calif.: Sage.

Kobben, A.
1967 Participation and Quantification: Fieldwork among the Dyuka. *In* Anthropologists in the Field. D. G. Jongmans and P. Gutkind, eds. Pp. 35-55. New York: Humanities Press.

McCurdy, D.
1976 The Medicine Man. *In* Ethics and Fieldwork: Dilemmas in Fieldwork. M. Fynkiewich and J. Spradley, eds. Pp. 4-16. New York: Wiley.

Mead, M.
1969 Research with Human Beings: A Model Derived from Anthropological Field Practice. *In* Experimentation with Human Subjects. P. Freund, ed. Pp. 152-177. New York: Russell Sage Foundation.

Mills, C.
1959 The Sociological Imagination. New York: Oxford University Press.

Mulhauser, F.
1975 Ethnography and Policymaking: The Case of Education. Human Organization 34:311.

Nader, T.
1969 Up the Anthropologist—Perspectives Gained from Studying Up. *In* Reinventing Anthropology. D. Hymes, ed. Pp. 284-311. New York: Vintage Press.

Paine, W. S., ed.
1982 Job Stress and Burnout: Research, Theory and Intervention Perspectives. Beverly Hills, Calif.: Sage.

Peattie, L.
1968 The View from the Barrio. Ann Arbor: University of Michigan Press.

Pelto, P.
1970 Anthropological Research: The Structure of Inquiry. New York: Harper & Row.

Polsky, N.
1967 Hustlers, Beats, and Others. Chicago: Aldine.

Reiman, J. H.
1979 Research Subjects, Political Subjects, and Human Subjects. *In* Deviance and Decency: The Ethics of Research with Human Subjects. C. B. Klockars and F. W. O'Connor, eds. Pp. 35-57. Beverly Hills, Calif.: Sage.

Reynolds, P. D.
1979 Ethical Dilemmas and Social Science Research. San Francisco: Jossey-Bass.

Sahlins, M.
1967 The Established Order: Do Not Fold, Spindle, or Mutilate. *In* The Rise and Fall of Project Camelot: Studies in the Relationship between Social Science and Practical Politics. J. Horowitz, ed. Pp. 71-79. Cambridge, Mass.: MIT Press.

Soloway, I., and J. Walters
1977 Workin' the Corner: The Ethics and Legality of Ethnographic Fieldwork among Active Heroin Addicts. *In* Street Ethnography: Selected Studies of Crime and Drug Use in Natural Settings. R. S. Weppner, ed. Pp. 159-178. Beverly Hills, Calif.: Sage.

Solzhenitsyn, A.
1972 For the Good of the Cause. D. Hoyd and M. Hayward, transls. New York: Praeger.

Spradley, J.
1976 Trouble in the Tank. *In* Ethics and Fieldwork: Dilemmas in Fieldwork. M. Rynkiewich and J. Spradley, eds. Pp. 17-31. New York: Wiley.

Thorne, B.
1980 You Still Takin' Notes? Fieldwork and Problems of Informal Consent. Social Problems 27(3):284-297.

Tobin, J. A.
1967 The Resettlement of the Enewetak People: A Study of a Displaced Community in the Marshall Islands. Ph.D. dissertation, University of California, Berkeley.

Wax, R.
1960 Twelve Years Later: An Analysis of Field Experience. *In* Human Organization Research. R. Adams and J. Preiss, eds. Pp. 166-178. Homewood, Ill.: Dorsey.
1971 Doing Fieldwork: Warnings and Advice. Chicago: University of Chicago Press.

Weber, M.
1946 From Max Weber's Essays in Sociology. H. Gerth and C. W. Mills, transls. and eds. New York: Oxford University Press.

Whyte, W. F.
1943 Street Corner Society: The Social Structure of an Italian Slum. Chicago: University of Chicago Press.

Wolcott, H.
1975 Criteria for an Ethnographic Approach to Research in Schools. Human Organization 34(2):111-127.
1980 How to Look Like an Anthropologist without Really Being One. Practicing Anthropology 3(2):6-7, 56-59.

Yablonsky, L.
1965 Experiences with the Criminal Community. *In* Applied Sociology. A. W. Gouldner and S. M. Miller, eds. Pp. 55-73. New York: Free Press.

42

Thinking Strategically About Private Sector Evaluation

The Key Issues

Luis Ma. R. Calingo, Robert Perloff, and Fred B. Bryant

Starting from pleas for formally assessing the value or worth of human service programs, evaluation research has matured to a major field of professional activity and scientific inquiry. As a professional discipline, the practice of program evaluation/evaluation research was for some time characterized as a "growth industry" itself (Rossi & Freeman, 1982). The increasing requests for program evaluations coming from government agencies and not-for-profit organizations commanding large resources for social research stimulated much of this growth. Thus program evaluation has become an integral component of program management in the public, not-for-profit sector.

Little is known, however, about the extent of evaluation activities in the American industrial enterprise. Business organizations also undertake a wide variety of programs aimed at effecting social and organizational change. If we view any comprehensive evaluation as consisting of performance monitoring, impact evaluation, and economic efficiency analysis (Rossi & Freeman, 1982) clearly these activities (particularly the third) should also be relevant to private, for-profit organizations. While performance measures in the private sector are traditionally viewed as "harder" than those of human service programs, organizational researchers have increasingly recognized the fuzzy nature of performance evaluation in industry (Birnberg, Turopolec, & Young, 1981; Ouchi, 1977). The program evaluator's facility in dealing with ill-structured social problems, among others, is definitely helpful in improving the practice of evaluation research in the private sector.

From Luis Ma. R. Calingo, Robert Perloff, and Fred B. Bryant, "Thinking Stategically About Private-Sector Evaluation: The Key Issues," original manuscript.

Authors' Note: This paper was presented at the joint meeting of the Evaluation Network and the Evaluation Research society, Chicago, October 1983. Portions of this paper were based on Robert Perloff's keynote address at the sixth annual conference of the Eastern Evaluation Research Society, New York, June 1983.

The decline in the level of public-sector resources available for program evaluation also contributes to the desirability of identifying evaluation research opportunities in the private sector. Within the last two years, the reduction of budgets and programs in the human services area has resulted in dramatic reductions in public dollars and manpower for evaluation research. It could be argued that this decrease in demand for evaluation will be only a temporary budgetary and ideological aberration. However, equally plausible is the argument that this represents a more fundamental structural change that threatens the survival of the evaluation research profession. Prudence on the part of the evaluation research community demands that the latter scenario be accepted as more likely and be used to guide our future actions. Moreover, it suggests that evaluators search for sponsors and users of evaluation research beyond the public sector.

Apart from the realities of decreasing public-sector expenditures for program evaluation, there are other strong reasons that evaluators should consider engagements in the private sector. First, it is very likely that the private sector pays more attractively than the public sector. Second, more diverse activities and challenges exist in the more heterogeneous private sector than are available to evaluators in the public sector. Third, due to the lesser degree of bureaucracy in the private sector (Fottler, 1981), it is not unreasonable to expect a quicker turnaround time between the planning of an evaluation project and the utilization of its findings. Finally, by working in a corporate setting and by dealing with a different set of criterion measures (e.g., profit), evaluators can gain not only fresh, new insights, but also new methodological and technical skills.

Evaluators should not, however, fall into the trap of aggressively soliciting private-sector engagements without first understanding the characteristics of this new "market" for evaluation research. Indeed, evaluators contemplating entry into the private sector stand to benefit by viewing themselves as new competitors entering an uncertain market. How should they position themselves in relation to quasi-evaluators (e.g., industrial psychologists) who are already entrenched in program evaluation in industry? The problem is compounded by the general lack of understanding about the size of the "evaluation market" in the private sector. While some work has focused on activities related to evaluation research in industry, this literature is small and uses a terminology unfamiliar to evaluators. Worse still, the available literature often cites the difficulty of having effective evaluations done in the private sector (Grant & Anderson, 1977; Murphy, 1980).

This paper represents an initial attempt to reduce this knowledge gap by identifying the key strategic issues involved in the decision to enter the private sector. Using a generally accepted framework for organizational analysis, the paper identifies the major evaluation opportunities in the private sector, their skills requirements and their relative attractiveness. We then analyze the capabilities of evaluators in the light of the skills required to successfully pursue the new evaluation opportunities. Finally, the paper suggests entry guidelines

that maximize the use of our existing strengths in order to take advantage of evaluation opportunities in the private sector.

THINKING ABOUT PRIVATE SECTOR ENTRY: THE KEY STRATEGIC QUESTIONS

To help us in developing strategies for entry into the private sector, a series of questions can provide a sequence and frame of reference for our thinking. The questions that make up this framework serve as aids in developing strategies for entry, with one question building on another, leading to conclusions regarding the most appropriate strategies.

What Evaluation Opportunities Exist in the Private Sector?

The approach begins with the question, What evaluation opportunities exist in the private sector? These opportunities represent potential "niches" or evaluation areas that evaluators can enter. Since management is the ultimate sponsor and user of such evaluations, these evaluation opportunities need to be defined within the context of the management function in industry.

One way to help understand the managerial task is to look within business organizations at various levels—strategic, coordinative, and operating (Kast & Rosenzweig, 1970; Thompson, 1967). There are basic differences in the orientation of managers at these different levels. In turn, the nature of evaluation issues is contingent upon these contextual differences.

Managers operating at the strategic level are concerned with how the organization chooses to relate to its environment, leading to decisions about *strategy:* In which products or markets should it compete? How should it compete within each product or market segment? The organization's strategy is, therefore, the appropriate unit of analysis for evaluations at this managerial level. Our the years, three streams of research and practice have emerged in the business strategy field which are oriented to the evaluation requirements at the strategic level. The first deals with the *ex ante* appropriateness of the organization's strategy, an evaluative task prior to strategy selection. The second deals with the *ex post* evaluation of the effectiveness of the selected strategy in achieving the organization's goals and objectives. The third evaluation-oriented stream is the relatively embryonic area of *strategic control.* This deals with the ongoing evaluation of the appropriateness of the organization's existing strategy in the light of posterior information about the company (i.e., its strengths and weaknesses) and its environment (i.e., the opportunities and threats facing the company).

Management at the coordinative level is concerned with the integration of the organization's internal activities in order to implement organizational strategy effectively. Such integration occurs primarily in the form of programs or projects that are undertaken in the different fucntional areas of the corporation such as production, marketing, and human resources. Evaluation activ-

ities at the coordinative level include (a) *outcome monitoring*—whether the program produces the desired level of outcomes at the right time; (b) *summative evaluation* of programs through either, or a combination of, impact evaluation or benefit-cost analysis. Clearly, much of the work traditionally ascribed to evaluators deals with this level.

At the operating level, management is concerned with the efficient and effective execution of organizational tasks. The managerial orientation at this level is more microscopic in the sense that the focus is on the management of processes that combine to form programs or projects. Obvious areas for evaluation at this level include

(a) the *efficiency* of existing processes;
(b) *program implementation*—whether the processes are being undertaken as planned; and
(c) *formative evaluation*—whether, and how, the process can be improved.

While these processes need not be boundary-spanning (i.e., directly related to organizational clients), we expect evaluation requirements at this level to focus on processes related to the company's mission or "service delivery" function, rather than purely administrative tasks. Again, this is one area in which the expertise of evaluators can be tapped.

Table 1 summarizes the preceding discussion by enumerating for each managerial level the corresponding entities that can be evaluated and the major evaluation issues at that level. It should be noted that each evaluation issue or problem is not exclusive to the managerial level under which it is listed. For example, formative evaluation may be equally applicable at both coordinative and operating levels. However, using the managerial levels as a starting point for the identificiation of the evaluation issues ensures the comprehensiveness and hopefully, exhaustiveness of the resulting classes of evaluation opportunities.

Do We Have the Comparative Advantage to Participate in Each Area?

Once the scope of evaluation opportunities in the private sector has been defined, the next question is, Where do we possess differential advantage in comparison with potential competitors among professionals in industry? Answering this question involves first identifying the skills required to participate in each evaluation area in industry. The presence or lack of these skills would spell the difference between success and failure when participating in each evaluation area. The normative implication is that evaluators planning to enter the private sector should consider only those opportunities where they possess better evaluative capabilities. In the long run, evaluators should develop capabilities and skills for participating in evaluation areas where their existing contribution could be weak at best.

TABLE 1
Major Evaluation Issues at Each Managerial Level in Private, For-Profit Organizations

Management Level	*Evaluable Entity*	*Evaluation Issues*
Strategic	Strategy, consisting of —which products/markets to compete in —how to compete in a particular product/market segment	1. Is this strategy appropriate in the light of company strengths and weaknesses, as well as the opportunities and threats the organization faces? *(ex ante* evaluation of strategy) 2. How cost-effective is this strategy? *(ex post evaluation of strategy)* 3. Is this strategy appropriate in the light of new information about company capabilities and environmental factors? *(strategic control)*
Coordinative	Programs/projects for the different functional areas: —production/operations —engineering/research and development —marketing —finance —personnel —public and governmental relations —management information systems	1. Is the program producing desired outcomes as planned? *(outcome monitoring)* 2. Is the program effective on the overall? *(summative evaluation)* a. impact evaluation b. benefit-cost analysis
Operating	Processes for doing both routine and nonroutine activities	1. How efficient are existing processes? *(efficiency evaluation)* 2. Are processes being implemented as planned? *(evaluation of program implementation)* 3. Can the process be improved? *(formative evaluation)*

TABLE 2
Partial List of Evaluation Methods Applicable in the Private Sector

Evaluation Area	*Relevant Methods from Policy Analysis and Evaluation*
Strategic	
Ex ante strategy evaluation	Multiattribute utility (Edwards, 1979; Saaty, 1977)
Ex post strategy evaluation	Dialectical performance assessment (Dunn, Mitroff, & Deutsch, 1980)
Strategic control	Dialectical performance assessment (Dunn, Mitroff, & Deutsch, 1980)
Coordinative	
Outcome monitoring	none
Impact evaluation	Experimental and quasi-experimental designs (Campbell & Stanley, 1966)
Benefit-cost analysis	*Ex post* benefit-cost and cost-effectiveness analyses (Haveman, 1973)
Operating	
Efficiency evaluation	none
Evaluation of program implementation	Participant observation (Glaser & Backer, 1973); Network analysis (Engelberg, 1980); Qualitative evaluation methods (Cook & Reichardt, 1979)
Formative evaluation	Qualitative methods; Nonconventional evaluation methods (Smith, 1981)

A measure of the comparative advantage of evaluators is the appropriateness and uniqueness of the methods they could bring if they participate in each area. A review of the literature on evaluation methodology suggests a number of methods developed in evaluation research that could be tapped by evaluators entering the private sector. Table 2 presents a partial list of these methods.

As shown in Table 2, evaluators can contribute in a methodological sense to the advancement of all but two of the evaluation areas identified in the private sector: monitoring of program outcomes and efficiency evaluation. Since the methods evaluators bring to the private sector were developed primarily for evaluation research applications, their longer experience in using these methods is a source of differential advantage for them. Since management scientists and operations researchers have a longer tradition of involvement in outcome monitoring and efficiency evaluation, these are evaluation areas in industry in which evaluators can benefit from their private-sector counterparts.

The immediate implication of the foregoing analysis seems to be that evaluators should focus their private-sector efforts only on those evaluation areas where they possess a distinctive competence. Table 2 suggests that, except for outcome monitoring and efficiency evaluation, the evaluator *can* enter almost any evaluation area in industry. However, this choice has to be mediated by another important determinant—the relative attractiveness of the proposed evaluation area. This brings us to the next question.

How Attractive Are These Evaluation Opportunities?

To have a more complete understanding of the evaluation opportunities in the private sector, it is also important that an assessment be made of the relative attractiveness of each of the identified evaluation areas. The extent to which an evaluation area represents an opportunity, in the strict sense, for evaluators is a function of the inherent attractiveness of that area. A measure of an evaluation area's attractiveness is the current stage of development of that evaluation area as a field of research and professional activity. A private sector evaluation area that is either in its embryonic or its growth stage would be more attractive to pursue than an area in its muturity stage. The obvious reason is that mature evaluation areas are already saturated with existing industry professionals, such that one evaluator's gain in clientele is another's loss. In the face of this zero-sum game, retaliation from existing industry professionals represents a barrier to entry by ENet/ERS-type evaluators.

Table 3 shows the stage of development of each of the nine identified evaluation areas in the private sector. While authors' license was exercised to some extent in estimating the relative attractiveness of these stages, the information in Table 3 provides a useful starting point for evaluators in deciding where they should navigate in the sea of opportunities afforded by the private sector.

The majority of the nine identified areas can be regarded as being in either the embryonic or growth stage of development. Among these are (a) strategic control, (b) *ex post* evaluation of strategy, (c) evaluation of program implementation, and (d) formative evaluation. A particularly noteworthy area is the summative evaluation of functional-area programs that are not considered to be in the mainstream of private-sector activities. These nonmainstream programs include employee mentoring, organization development programs, employee assistance programs, assertiveness programs for female managers, quality control circles, and corporate arts programs, (through which companies invest in good paintings to be hung in executives' offices).

Evaluation activities for nonmainstream programs are still in the embryonic stage. These programs are, furthermore, very close to the types of programs in which evaluators are traditionally engaged. Therefore, they represent an attractive opportunity for evaluators to make significiant contributions to the private sector. The attractiveness of these nonmainstream programs arises, for the most part, from the noninvolvement in these nonmainstream areas of evaluation-oriented professionals in industry. For example, major mainstream responsibilities of our colleagues in industry (e.g., industrial psychologists) include performance evaluation, employee motivation, and attitude and morale surveys, to name a few.

Which Evaluation Areas Do We Enter?

Once the scope of opportunities has been defined and our comparative advantage has been identified, the next question is, How do we most effec-

TABLE 3
Stage of Development of Each Evaluation Area in the Private, For-Profit Sector

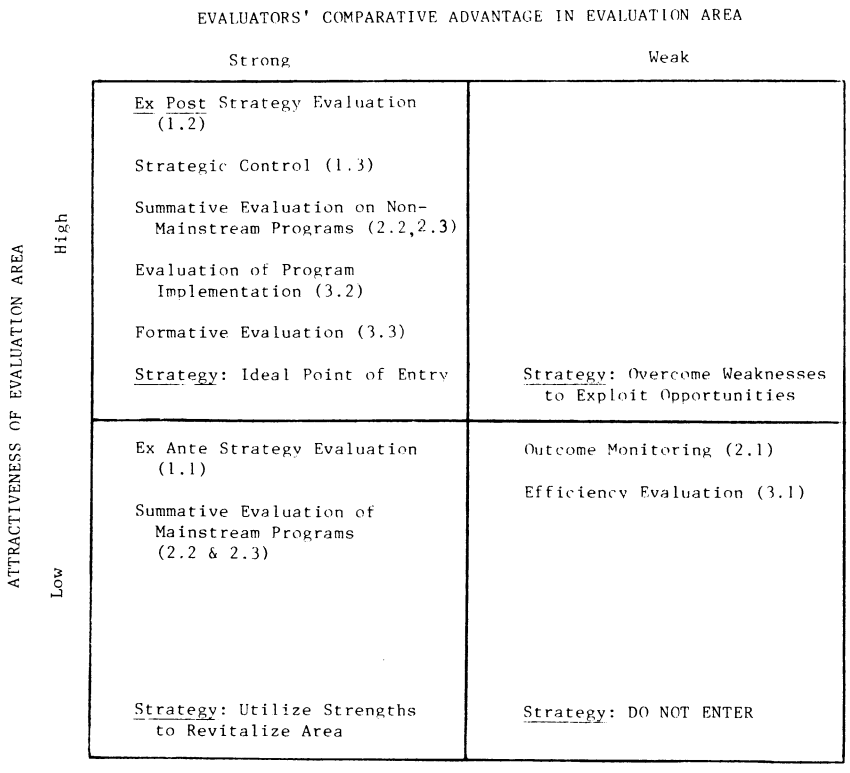
Evaluation Area	*Stage of Development as an Evaluation Area*	*Exemplars (if any)*	
		Evaluation Research	*Business Management*
Strategic			
Ex ante evaluation	Maturity		Tilles (1963)
Ex post evaluation	Growth	Dunn, Mitroff, & Deutsh (1980)	Mitroff, Emshoff, & Kilmann (1979)
Strategic control	Embryonic		Newman (1975)
Coordinative			
Outcome monitoring	Maturity	Sorensen & Elpers (1978)	M.I.S. research (Blumenthal, 1969)
Impact evaluation of programs/projects	Maturity for mainstream programs (e.g., social audit)	Suchman (1967)	Bauer & Fenn (1973)
	Embryonic for nonmainstream programs (e.g., corporate arts)		
Benefit-cost analysis	Maturity for mainstream programs	Haveman (1973)	Dean (1951)
	Embryonic for nonmainstream programs		
Operating			
Efficiency evaluation	Maturity		Charnes, Cooper, & Rhodes (1981)
			Taylor (1981)
Evaluation of program implementation	Growth	Patton (1978)	M.I.S.
Formative evaluation	Embryonic	Scriven (1972)	

tively employ the advantages we have, counter to those of potential competitors in the private sector, and develop or acquire greater advantage?

Knowing that some evaluation areas are inherently more attractive than others and that evaluators possess a distinctive competence in some areas, answering the strategy question involves the simultaneous consideration of two factors. These are the capabilities of the evaluator to compete in a particular evaluation are (as presented in Table 2) and the relative attractiveness of each evaluation area (Table 3). A simple yet insightful way of synthesizing the information in Tables 2 and 3 is to classify the evaluation areas identified in the private sector in terms of a 2 x 2 matrix, where the two matrix dimensions represent the two factors mentioned above.

Figure 1 illustrates the results of this classification scheme. The nine private-sector evaluation areas are classified into four cells or situations. These four situations can be viewed as market "clusters" or "segments," which form the basis of four distinct strategies or guidelines for private-sector entry.

The evaluation areas relegated to Cell 1 are those areas that are inherently attractive (i.e., in either embryonic or growth stage of development) and in



NOTE: Numbers refer to the area's identification number in Table 3.

Figure 1: Private Sector Evaluation Areas Classified by Comparative Advantage and Attractiveness, with Proposed Strategies for Entry

which evaluators possess a comparative advantage. Five evaluation areas can be identified as appropriately belonging to Cell 1, namely, (a) *ex post* strategy evaluation, (b) strategic control, (c) summative evaluation of nonmainstream programs, (d) evaluation of program implementation, and (e) formative evaluation. The appropriate strategy in these areas is clear and simple: Enter from strengths; build and grow in that segment. This involves extending the evaluator's portfolio of services and is obviously the ideal position from which to enter the private sector.

Since too many of us might end up thinking the same way and enter the private sector via this route, three other segments and their corresponding strategies in the private-sector market are worth equal consideration. Cell 2 evaluation areas include those areas that are inherently attractive but in which evaluators do not possess a comparative advantage. None of the nine private-sector evaluation areas could be identified offhand as belonging to this cell. However, it is useful to note that the appropriate strategy for this cell is for evaluators to overcome their weakness or comparative disadvantage in order

to take advantage of these attractive opportunities. Joint ventures with private-sector professionals is a viable means of implementing this strategy.

Cell 3 includes those evaluation areas that are in the maturity stage (and, therefore, less attractive), but in which evaluators possess distinctive competences. The appropriate strategy for these evaluation areas seems to be for evaluators to utilize their strengths or comparative advantage (in the form of the new or improved methods they could bring) in order to revitalize these evaluation areas. Two of the nine evaluation areas can be classified as belonging to this cell, namely, (a) *ex ante* strategy evaluation, and (b) summative evaluation of mainstream programs.

Finally, Cell 4 evaluation areas include outcome monitoring and efficiency evaluation. These are areas already past the maturity stage (and, therefore, not attractive as a point of entry) and in which evaluators do not possess a comparative advantage by way of methodology. The appropriate strategy for these evaluation areas is clear and simple: Do not enter.

This analysis has identified seven major groups of evaluation opportunities and has proposed differing approaches for exploiting these opportunities. While the reader may disagree with the subjective assessments we have made, the approach we advocate ensures that important external elements (i.e., those pertaining to the opportunities themselves), as well as important internal elements (i.e., those pertaining to the evaluators' capabilities), have been explicitly considered.

CONCLUSION

We have attempted in this paper to alert the evaluation research community to the vast evaluation opportunities in the private sector. We have in the process proposed general guidelines as how we evaluators can best utilize our competencies to take advantage of these evaluation opportunities.

The problem now at hand is how to get started. We do not see any substitute for the hard but direct approach of "pounding pavements, wearing out shoe leather, knocking on doors" and dealing with potential clients on a one-to-one basis. Another approach would be for us to "plant the seeds" by writing short essays or articles in trade journals or organizational newsletters, describing ways that evaluations might be of use to organizations. In addition, we can interact with colleagues already working in the private sector (e.g., psychologists, training and development staff). These interactions will acquaint private-sector colleagues with ways their companies may be served by professional evaluators and investigate the possibility of setting up joint ventures to this end. Finally, we can interact with consulting firms who work regularly for organizations to encourage them to expand their service offerings by including the resources and capabilities traditionally ascribed to evaluators.

REFERENCES

Bauer, R. A., & Fenn, D. H. (1973). What *is* a corporate social audit? *Harvard Business Review,* **51**(1), 37–48.

Birnberg, J. G., Turopolec, L., & Young, S. M. (1981). *The organizational context of accounting* (Graduate School of Business Working Paper WP-484). University of Pittsburgh, November.

Blumenthal, S. (1969). *Management information systems: A framework for planning and control.* Englewood Cliffs, NJ: Prentice-Hall.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Charnes, A., Cooper, W. W., & Rhodes, E. (1981). Evaluating program and managerial efficiency: An application of data envelopment analysis to Program Follow Through. *Management Science,* **27,** 668–697.

Cook, T. D., & Reichardt, C. S. (1979). *Qualitative and quantitative methods in evaluation research.* Beverly Hills, CA: Sage Publications.

Dean, J. (1951). *Capital budgeting: Top-management policy in plant, equipment, and product development.* New York: Columbia University Press.

Dunn, W. N., Mitroff, I. I., & Deutsch, S. J. (1980). *The failure of evaluation research: Guidelines for maximizing relevance.* Unpublished manuscript, University of Pittsburgh and Georgia Institute of Technology.

Edwards, W. (1979). Multiattribute utility measurement: Evaluating desegregation plans in a highly political context. In R. Perloff (Ed.), *Evaluator Interventions: Pros and Cons.* Beverly Hills, CA: Sage Publications.

Engelberg, S. (1980). Network analysis in evaluation: Some words of caution. *Evaluation and Program Planning,* **3,** 15–23.

Fottler, M. D. (1981). Is management really generic? *Academy of Management Review,* **6,** 1–12.

Glaser, E. M., & Backer, T. E. (1973). A look at participant observation. *Evaluation,* 1973, **1**(3), 46–49.

Grant, D. L., & Anderson, S. B. (1977). Issues in the evaluation of training. *Professional Psychology,* **8,** 659–673.

Haveman, R. H. (1973). Ex post benefit-cost analysis: The case of public investments in navigation facilities. In *Benefit-cost analysis of federal programs.* U.S. Congress, Joint Economic Committee, 92nd Cong., 2nd Sess., Committee print. Washington, DC: U.S. Government Printing Office.

Kast, F. E., & Rosenzweig, J. E. (1970). *Organization and management: A systems approach.* New York: McGraw-Hill.

Mitroff, I. I., Emshoff, J. R., & Kilmann, R. H. (1979). Assumptional analysis: A methodology for strategic problem solving. *Management Science,* **25,** 583–593.

Murphy, D. E. (1980). *A report from ASTD: Problems and progress.* Paper presented at the American Educational Research Association conference, Boston, April.

Newman, W. H. (1975). *Constructive control: Design and use of control systems.* Englewood Cliffs, NJ: Prentice-Hall.

Ouchi, W. G. (1977). The relationship between organizational structure and organizational control. *Administrative Science Quarterly,* **22,** 95–113.

Patton, M. Q. (1978). Evaluation of program implementation. In M. Q. Patton (Ed.), *Utilization-focused evaluation.* Beverly Hills, CA: Sage Publications.

Rossi, P. H., & Freeman, H. E. (1982). *Evaluation: A systematic approach* (2nd Ed.). Beverly Hills, CA: Sage Publications.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology,* **15,** 234–281.

Scriven, M. (1972). The methodology of evaluation. In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education.* Boston: Allyn & Bacon.

Smith, N. L. (Ed.). (1981). *Metaphors for research: Sources of new methods.* Beverly Hills, CA: Sage Publications.

Sorensen, J. E., & Elpers, J. R. (1978). Developing information systems for human service organizations. In C. C. Attkisson, W. A. Hargreaves, M. J. Horowitz, & J. E. Sorensen (Eds.), *Evaluation of human service programs.* New York: Academic Press.

Suchman, E. A. (1967). *Evaluation research: Principles and practice in public service and social action programs.* New York: Russell Sage Foundation.

Taylor, F. W. (1981). *Scientific management.* New York: Harper & Row.

Thompson, J. D. (1967). *Organizations in action: social science bases of administrative theory.* New York: McGraw-Hill.

Tilles, S. (1963). How to evaluate corporate strategy. *Harvard Business Review,* **41**(4), 111–121.

43

Opportunities for Evaluation in the Next Few Years

Thomas D. Cook

It would be pointless for me to document that evaluation is not now growing at the rate it did during the 1960s and 1970s. Indeed, it may be declining both in the dollar amount spent for activities that can be broadly conceived as evaluative and in the number of persons whose major professional identity is evaluator. Moreover, several of its stellar theorists now seem much less active in the field, and *Evaluation Review* recently let it be known that the flow of manuscripts has declined. While each of these factors has many possible explanations, they do not create the image of a vital and growing field. They speak more to nongrowth and even to contraction. In a situation like this, I would not expect organizations like the Evaluation Research Society (ERS) and the Evaluation Network (ENet) to "go gently into that dark night." Their premier organizational mandate—to maintain their health—demands other responses. One obvious way to maintain organizational health is to seek out new opportunities for the exercise of one's talents, and it was in this context that I interpreted Paul Wortman's request to speak to you today about new opportunities for evaluation.

What I have to say today will be dispiriting to those who believe that a market exists for evaluators in the private sector of the American

---

AUTHOR'S NOTE: This paper was delivered at the 1983 ERS/ENet Convention. Its contents reflect many stimulating conversations with Laura Leviton and William R. Shadish, Jr., neither of whom is responsible for lapses of logic or taste.

From Thomas D. Cook, "Opportunities for Evaluation in the Next Few Years," *Evaluation News*, 1984, 5(2), 20–46.

economy or in those parts of the public sector concerned either with defense or with physical infrastructure—housing, sewers, airports, and roads. I will argue that very few opportunities exist for evaluators beyond the current sphere of social programs. This is not because evaluation is irrelevant in the other contexts. Rather, it is because evaluation has long been practiced there, is widely perceived to be satisfactory, and we in ERS and ENet can add little that the professionals working in these settings could not easily learn from us and adopt into their own practice. Moreover, in cases where evaluators are successful in entering the private sector, the military, or spheres concerned with "hard services," my guess is that they will be asked to deal with issues that go beyond the logic of evaluation, with its emphasis on determining justifiable criteria of merit, standards of comparison, techniques for measuring performance, and methods for judging results. Instead, they may become more like applied social researchers in general, concerned to apply social science methods to practical problems, whatever the relationship to evaluation logic.

Where, then, are new opportunities for evaluators to be found? My judgment is that they are to be found in trying to learn the lessons of evaluation's own brief past and in rapidly reexamining the major reasons for the belief of many that evaluation seems to be less useful than originally hoped. Such an assessment would help those who have to make decisions about future investments in evaluation, since these decisions could then be based more on evaluation as it is today than as it was 10 or 20 years ago, when the first generation of evaluators was busy learning its mistakes—many of which, in my opinion, need not be repeated, but all of which contributed to the current image of a disappointing payoff from evaluation.

GROWTH OPPORTUNITIES IN THE PRIVATE SECTOR

The Contextual Predicaments in Evaluating Social Programs

In the public sector, evaluations serve many functions. One is to summarize the past achievements of social programs and their constituent elements. Another is to provide feedback about ways to

improve a program through the direct observation of program functioning, the deliberate testing of alternative practices, or the dissemination throughout a program of practices that could be more widely adopted at the local level.

To fulfill these functions, evaluators need to subscribe to a logic of evaluation. The one most widespread in the United States is that of Scriven (1983). It emphasizes three tasks to be carried out once a decision has been made about the object to be evaluated (the *evaluand):* (1) criteria of merit have to be determined that specify the outcome variables worth study; (2) standards of comparison have to be determined that specify whether the evaluand is to be compared to a single presumed standard (e.g., whether a speed limit of 55 miles per hour is exceeded) or to an alternative way of bringing about comparable outcomes (e.g., whether a Honda is preferable to a Toyota); and (3) measures have to be made on each of the criteria of merit for each of the standards being compared. We might add another step: (4) The results of the measurement process have to be synthesized across all of the criteria so as to arrive at an assessment of the merit of the evaluand. This logic is extremely general and applies as much to evaluating past achievements as to improving internal operations.

Unfortunately, difficulties arise when trying to implement this logic for the study of social programs. The first difficulty is in selecting criteria. This arises because program goals are often vague, stakeholders differ in the value they accord to different outcomes, some variables are more easily measured than others, and some are more likely than others to show changes within the time frame available to evaluators. Next is the problem that stakeholders differ in what they would like to see an evaluand compared to. Some prefer a comparison with nothing so that absolute efficacy can be assessed; others prefer a comparison to current practice so as to assess marginal improvement over the present; and yet others prefer a comparison with the major competitors so that the currently optimal can be identified. Third is a difficulty much discussed in the past literature on evaluation. It is one thing to specify that measurement should take place on every criterion; it is quite another to specify how the measurement should take place so as to infer causal consequences, detect unintended side effects, and probe the processes that mediated obtained results. Finally, many problems beset anyone who would synthesize results across many outcome variables, for they

will differ in validity, utility, proximity to program activities, and degree of confidence in the results about them. There are techniques of synthesis—for example, cost-effectiveness analysis, multiattribute utility scaling, and other weight-and-sum techniques—but each is fraught with technical and value assumptions.

It would be myopic to view the problems of implementing evaluation logic as purely methodological. When the evaluand is a social program, historical, political, and administrative difficulties arise that are intrinsic to how social programs are begun, maintained, and modified. Administratively, social programs are diffuse entities, mostly providing funds and vague directives from a central source. Local project directors and their staffs often see the program goals and regulations as being of limited relevance to their own professional and personal agendas, and the latter heavily influence both the choice of services and their mode of delivery. Moreover, for many reasons senior program and local project officials change jobs often, and their replacements are often under pressure to modify how things are done even if this radically alters the nature of what is being evaluated. On the political front, it is much easier to gain funds for a social program if a heterogeneous collection of legislators and other opinion makers supports it. This is why program goals are often written to be vague and general. But in espousing generality to gain the greatest number of supporters, it is more difficult to specify evaluation criteria and standards of comparison. Stakeholder groups also seek to maintain the funds for the programs on which their welfare depends, usually making their case on political grounds. Evaluative feedback will tend to be overlooked if it conflicts with the interests of powerful groups that instead resort to power politics.

Another major contributor to decision-making is ideology. Every nation has a history of cherished ideals, and in the United States these include several that influence evaluation practice and utility—for example, the services delivered to persons in need should be under local control; professionals should be as autonomous as possible; public debates about program changes should be held, and all the groups that might be affected by the changes should be invited to make their point of view known. Such historical ideals inevitably influence program design; but they also limit the ways in which evaluation results can be used. The upshot of all this is that when social programs are evaluated, the general logic of evaluation can be only imperfectly implemented in the

administrative, political, and historical context in which social programs currently function.

The utility of social program evaluation is constrained by another contextual factor, this one concerning the role of evaluation within problem-solving logics. Programs are designed and funded to alleviate social problems. One of the arts of policy analysis is to know which potential solution to a problem (i.e., which program) to implement first, since it is often the case that a selection has to be made from among many alternatives. In rational models of problem-solving, each potential solution, once selected, must then be evaluated in terms of the extent to which it solves the original problem, and a decision must subsequently be made about whether the option is to be retained. In such a conceptualization of problem-solving, evaluation is merely one of four stages. The first involves problem selection, the second selecting potential solutions for test, the third involves testing or evaluating potential solutions, and the final stage concerns the retention and use of solutions that reduce the original problem to manageable proportions.

In this conceptualization, the evaluation of social programs could be perfect but useless. This would be the case if a program is designed that addresses a trivial social problem, or if the solutions selected for study are so restricted in range and puny in power that they have no hope of affecting the original problem, or if the persons who might have used the evaluation results prefer instead to rely on political expediency or some similar criterion. Evaluation has its own logic; but this logic is embedded within a larger problem-solving logic over whose three other components—problem selection, solution choice, and utilization—evaluators have little control. The image thus emerges of an evaluation logic that can be imperfectly implemented only when social programs are being studied and that totally depends for its utility on a problem-solving context over which evaluators have little control.

The foregoing analysis further suggests that we should not be surprised if the evaluation of social programs has been disappointing in its results in the past. For evaluation to be adequate requires (1) a defensible theory or logic of valuing; (2) a defensible epistemology for learning about society; (3) a defensible set of practical methods that are consonant not only with preferred theories of knowing, but also with the political and administrative realities of social programs and the resources evaluators typically have at their disposal; and (4) a widespread

understanding of the modest role that evaluation necessarily plays in any scheme for social problem-solving in the public sector, since powerful reasons exist there for leaving social problems poorly defined (Lindblom & Cohen, 1979), for implementing social programs that only marginally differ from current practice (Shadish, in press), and for basing decisions on political criteria of convenience, ideology, and survival rather than a social approach (Weiss & Weiss, 1981).

One obvious appeal of the private sector and the military for evaluators is that these sectors are expanding and so offer new possibilities for work. But if I am correct in the preceding analysis, growth is not their only source of appeal. The political and administrative factors that operate in an open society to make evaluation more difficult operate less in the more closed world of the private sector and the military, where sources of authority are less ambiguous. Consequently, program goals are likely to be clearer, and the interest groups likely to make their point of view known are less disparate. Moreover, traditions of hierarchical organizational structure make local practice more subservient to central directors and operating manuals than is the case with social programs. Evaluation is easier to conduct under these circumstances and is less likely to produce unclear findings that are open to dispute on technical or ideological grounds. I do not want to be misunderstood. All the problems with social programs exist in the private sector and military; only to a lesser degree.

The Professions Already Providing Evaluative Services in the Private Sector and in the Military

Because evaluation is a necessary part of all problem-solving, it should not be surprising to note that the private sector and the military have evaluators. Indeed, the functions of evaluation are so crucial that the private sector had evaluators long before there was an ERS or ENet. (Moreover, few of these evaluators belong to either organization.)

The professions that use evaluation logic to summarize achievements or to suggest improvements in the private sector are accountants, management consultants, marketing experts, and research and development scientists. Consider accountants. In the private sector, independent auditors examine the books of corporations for a given period,

critically assessing past revenues and expenditures and current inventories. They do this to arrive at a "bottom line" summative judgment about the organization's health based on its profitability. Waiting for this information are several groups, including corporate officers, shareholders, and the federal officials who collect taxes or protect the public against fraud and gross mismanagement. In fulfilling this summative function, the private sector is particularly fortunate. First, there is a widely agreed-upon criterion of merit—profitability; next, money has fortunate metric qualities that make measurement relatively easy; and third, a single comparison standard is usually involved—how much more was earned than spent. To be sure, social critics worry about the simplicity of such criteria and standards, and they ask for corporations to be evaluated in terms of their unintended social consequences, such as the social costs of the pollution that accompanies steel production. But for the corporations themselves, these side effects are less important than financial viability, and assessing them remains largely the purview of government officials responsible for social impact assessment.

Accountants sometimes provide financial information on a division-by-division basis. This helps provide knowledge about where problems are occurring in the organization or where performance exceeds expectations. From such information ideas can be generated to improve internal operations. Bookkeepers employed by a corporation also provide evidence about sources of poor or superior performance, and judgments about these issues also arise in conversations between managers and in the reports of external management consultants. Management consultants specialize in improving internal operations through diagnoses of the exact nature of a problem and through offering suggestions about potential solutions often selected from a set suggested by the consultant's prior experience in the sector. Irrespective of whether bookkeepers, managers, or external management consultants are involved, each is dedicated to the evaluative functions of detecting, diagnosing, and remedying problems of operation. In this they are like many members of ERS and ENet who work on primarily formative tasks, either as in-house or as external evaluators.

The private sector has a need to develop and test new products and means of production, and it has professionals who cater to these evaluative needs. Basically these are persons working in the research

department of a firm. Their job is to keep a sharp eye on new developments in basic research in their substantive fields and to try to make these ideas relevant to the business for which they work. They might therefore create new products or modify already existing ones so as to make them more locally appropriate. In the public sector, the function of testing new products is fulfilled largely by demonstration projects designed to test novel ideas, many of which have been generated partly from basic research findings in the social sciences.

In the private sector it is sometimes also necessary to stimulate the use of already proven ideas, often in the satellite offices that report to corporate headquarters. Consequently, the private sector has a vested interest in disseminating information about successful practices in order to get them adopted. Indeed, in businesses that sell directly to the public, financial success depends to such a large degree on how information about products is disseminated and how customers respond to the information that a large industry has developed to market products and measure the effectiveness of marketing campaigns. In some conceptualizations of evaluation, an important function is not only to summarize past achievements and develop new ideas for practice, but also to disseminate more widely what is already known about successful practices.

We see, then, that the evaluative functions of (1) independent, retrospective summary and (2) improving internal operations through (a) on-the-job observation, (b) the creation of new products, and (c) improved dissemination of knowledge about successful practices are so important to the private sector that several professions already exist to meet these evaluative needs.

What Can Members of ERS and ENet Contribute to Evaluation in the Private Sector?

In general, I suspect that we evaluators have little to add to what the private sector already does to evaluate itself. We are—or should be—characterized by extensive knowledge of the logic of evaluation so that we are particularly knowledgeable about selecting criteria and standards, deciding how to measure, and synthesizing the results of any measurement. We are also—or should be—characterized by extensive knowledge

of social science methods, of the historical, administrative, and political realities that influence how social programs operate, and by an awareness of the larger problem-solving context that makes the utility of evaluation dependent on processes of problem selection, solution generation, and decision-making over which evaluators have little or no control.

I find evaluation to be most distinct from the other professions previously mentioned because of its past experiences in open society contexts and because of its commitment to the methods of the social sciences for the measurement of program activities and consequences. Unfortunately, our sensitivity to open system constraints is less needed in the private sector or the military; and what perplexes me about the use of social science methods is that evaluators do not have unique access to them, for they are widely available in textbooks. Yet accountants, management consultants, research and development specialists, and marketing experts do not routinely use these methods, with formal sampling techniques being an occasional exception. It may be that the practitioners in these fields have tried and rejected such methods; or they may have tried them and modified them heavily in the crucible of experience; or it may even be that the methods are not known. I am dubious about this last possibility, though. The businesses involved in providing evaluative services to the private sector often hire people with training in the social sciences but do not use their cutting-edge skills. Moreover, the businesses are in active competition with each other; and anyone who believed that the social science methods of academe give a competitive edge could have tried them by now. But to my knowledge few firms advertise to potential clients in terms of their ability to do more sophisticated participant-observation, interviewing, sampling, or experiments. They seem to advertise more by reputation, price, and sectoral experience.

My guess is that the professions that provide evaluative services to the private sector do not appreciate the marginal gains in bias reduction and precision that the more refined social science methods achieve, for they may achieve this at a cost—in time, flexibility, intelligibility, and perhaps even expenditures. By and large, the professions in question have adopted other strategies that enhance credibility, and perhaps even validity. For example, accountants rely heavily on a set of professional standards that are more detailed and specific than those available to

evaluators, so that it is relatively more easy in accountancy to know what are and are not standard practices. Moreover, the practices in accountancy and related fields that have survived the test of time and usage may have done so because clients are willing to live with whatever biases or imprecision remain. To be viable in the private sector or the military, evaluators have to make the case not that their methods are superior, but that their superiority promises a consistent, practical payoff.

Management consultants seem to derive their credibility from repeated experience in a particular business or sector so that they accumulate considerable tacit or practitioner knowledge that quickly guides them to the identification of key issues. Also, their comparative knowledge of other businesses and key new developments is supposed to put them in a unique position to suggest practical ways to improve practice that are likely to be effective because they have been implemented with apparent success elsewhere. It seems to me that by comparison, social program evaluators are more likely to flit from site to site, from program to program, and even from sector to sector, making it difficult to acquire the site-specific insider knowledge that many teams of management consultants eventually gain. As far as I can tell, the corporations that use management consultants do not seem obviously dissatisfied with what they get and do not seem to be actively searching for alternatives—certainly not for alternatives based on more advanced social science methods that were developed to promote scholarship by making each study as valid as possible.

In many parts of the private sector it is alien to think in terms of the perfect research study. Consider, in this connection, how market research firms deal with the dissemination questions of their largest corporate clients. Once a product has been tentatively developed, market researchers bring together focus groups composed of purposively sampled but heterogeneous individuals who sample the proposed product and discuss their reactions to it. After this, the product may be modified. But then other individuals, again purposively sampled but heterogeneous, are asked to take the product home, use it, and report on their experiences with it. The product may then be revised again before being test-marketed at a small number of purposively selected but heterogeneous sites. Here, sales and customer reactions will be noted and plans will be made for further modifications, or perhaps even for

national distribution. Involved here is sequential, programmatic research. There is no reliance on a single, perfect study; sampling designs are not of the kind advocated in textbooks on sampling; data analyses are more impressionistic than the statistical tests used by evaluators who prefer quantitative methods or the self-consciously critical analyses of evaluators who have been schooled in qualitative techniques. The market researchers' conception of research would be recognized as more sophisticated today than 10 years ago, given the growing advocacy of programs of research, recent critiques of the inevitable limitations of a single study and of statistical tests, and given also recent work that emphasizes how external validity can depend on replication across heterogeneous instances as well as random sampling and how practitioner knowledge is partially disciplined by trial-and-error tests.

Some theorists of evaluation—particularly Wholey and Rossi—see one of evaluation's greatest contributions as making policymakers aware of the theory behind a social program as well as providing them with knowledge about effects and internal operations. This type of theoretical knowledge is usually achieved while planning an evaluation, for one way evaluators can arrive at criteria and standards of comparison is by describing what is supposed to happen in a program or local project. Once this is known, a measurement framework can be generated to estimate in provisional fashion the likelihood of each of the links occurring that seem necessary if a program is to effect a particular outcome.

Evaluability assessment of this kind is an important consciousness-raising device for managers, providing them with an overview of the program and of the assumptions on which its efficacy is thought to depend. My guess is that most evaluators will be able to do a better job of evaluability assessment than most management consultants. However, management consultants do try to lay bare the theory and assumptions behind a factory or corporation, and they also try to examine whether the sequence of planned activities is logically ordered and consonant with what is already known. Consequently, the evaluator's advantage in analyzing the theory behind an evaluand may be real but not large. Moreover, such analysis is more of a preevaluative than an evaluative task, casting the evaluator more in the role of a policy analyst or applied social researcher than an evaluator.

It should be clear from the above that I am not entirely sure what evaluation researchers have to offer companies in the private sector, since they are already served by professions that provide evaluative services without calling them this. I have the same doubts about evaluation in the military, for firms already exist to evaluate tanks, personnel, and early warning defense systems, and traditions of field trials, test firings, and simulations have developed from a long history of practice designed to evaluate the capability of new and existing systems. However, much of the evaluation done for the military is carried out by the very firms that produce the systems. They are supposed to be monitored in their evaluation by professional officers. However, the officers' careers partly depend on producing successful systems, and this may bias them toward obtaining apparently successful evaluation outcomes. In this system, no one can afford failure or indifferent performance. Outside perspectives are clearly needed. Could we in ERS and ENet provide them?

A special difficulty arises here because most of the persons in ERS and ENet are social scientists by training, and many of the evaluands in which the military is interested speak more to the substantive knowledge of engineers. There are some social programs in each service, though, such as drug abuse or safety programs. However, each service currently is associated with companies that regularly provide evaluations, few of whose employees belong to ERS or ENet. Most seem to be ex-officers, psychometricians, or industrial or organizational psychologists. Anthropologists, educators, and social workers are rare, as are even sociologists and social psychologists. Thus the range of social scientists is restricted. And even if possibilities for contracts were to open up in the military despite my pessimism, there would be an unknown number of members of ERS and ENet who would find it unpalatable to collaborate in evaluating weapons systems or even social programs in the military. So I suspect that if the military wanted outside evaluators other than those in the firms they currently patronize, they may be more inclined to go to engineers or economists. Evaluators with such background rarely join ERS and ENet.

Why Cannot Other Professions Adopt What Evaluation Has to Offer?

If my pessimism is wrong and members of ERS and ENet do indeed have valuable skills to offer the private sector and the military, nothing

prevents the firms that currently supply evaluative services to these sectors from incorporating into their practice those parts of our practice that their clients find useful. This would immediately reduce the uniqueness of evaluation's contribution and, though flattering, would restrict the amount of work potentially available to members of ERS and ENet. Any scheme for entering the private and military sectors has to deal head-on with the possibility that the useful services we provided could be assimilated into the standard practice of other professions. There is little evaluators could do to prevent this, except perhaps to argue that management consulting firms would do better to add evaluation specialists than to add new skills to the repertoires of their current employees.

I do not want to be unduly pessimistic. A few individual evaluators have already stuck their foot in the door. If their work turns out to be useful, they will be used again, slowly paving the way for more widespread realization of evaluation's relevance and utility. However, I believe that such evaluators will be welcomed more for their general skills in policy analysis or for their command of particular social science methods than for their skills in evaluating options designed to contribute to solving important problems. If so, the pressure to find work may incline evaluators in ERS and ENet to accept contracts for applied settings that reduce their distinctiveness. At present, the majority of evaluators in ERS and ENet have been trained in education, social work, psychology, anthropology, and sociology and have too little contact with evaluators trained in economics, political science, and policy studies. This last group of social scientists works with a broader range of programs than the human service sector in which members of ERS and ENet specialize, and they tend to belong to professional associations like the Association for Public Policy and Management rather than ERS and ENet.

OPPORTUNITIES FOR IMPROVING EVALUATION'S IMAGE AND PRACTICE

If members of ERS and ENet cannot expect much expansion beyond the domain of social programs with a social welfare or educational emphasis, what should they do now to create more opportunities for

their talents in the future? I would suggest that instead of trying to expand, we should try to consolidate what we know for when a new administration takes office that cares more about social welfare or management by results than is the case with President Reagan's administration. I suggest that consolidation take place along two lines. The first involves confronting two coarse but widespread beliefs: that "social programs don't work" and that "evaluations are so insensitive that true effects of modest magnitude cannot be detected." The second involves asking what we have learned from the last two decades of evaluation about the types of research that have the greatest payoff and should be advocated for the future.

Strategies for Probing Some Widespread Beliefs that Reduce the Perceived Utility of Evaluation

The perceived utility of evaluation is reduced by the widespread and undifferentiated belief that social programs do not work. If nothing works, the argument goes, evaluation is unnecessary and the resources it consumes might be used more profitably to redefine social problems, develop substantive theories, or design novel options for practice. Another response to "nothing works" is to claim that past evaluations were generally so insensitive (relative to the size of expected effects) that they failed to detect real gains of modest magnitude. It would be extremely useful, I believe, if evaluators were to spend the next few years appraising these beliefs, for as long as they are widely accepted it will be difficult to justify evaluation.

In particular, we need to bring together the research evidence on the programs, types of local projects, and elements of practice that seem most useful for (1) reaching clients, (2) implementing services, (3) affecting program beneficiaries in the short term, and (4) having broader societal impacts in the longer term. Studies with such goals would be more helpful if they were structured around syntheses of past evaluations rather than around the detailed examination of individual studies. This is because I am personally convinced that most of the treatments we implement are not very bold deviations from the status quo and are not well informed by well-tested theories, which specify the conditions under which clients should and should not benefit from program services. Moreover, evaluations are implemented with considerable

slippage—and anyway, the methods used in nearly all cases are coarse grained. Since we should not expect effects to be large, but should expect margins of error to be large, it follows that modest gains are more likely to be detected in syntheses than in individual studies. (For a partial review of the results, especially meta-analyses, see Light, 1983.) I do not want to suggest that we should all run out and uncritically meta-analyze everything. Like nearly all research tools, meta-analysis is a delicate instrument, easily abused by those who forget its crucial assumptions (Cook & Leviton, 1980). However, many social agencies in Washington have recently commissioned meta-analyses and other forms of synthesis in the hope that their programs and procedures will be shown to be modestly effective. Such studies are already beginning to revise the coarse belief that nothing works and may do so even more in the future.

The syntheses have had a second important consequence. Most of us now set our expectations lower, not only about what a single social program can achieve but also about what a single evaluation can do. The reduced expectations stem from a stronger awareness of the inevitable contextual and methodological limitations of a single study. While the results of meta-analyses have made it difficult to argue that the methods used by quantitatively inclined evaluators in the past were totally insensitive, they have definitely indicated that the methods need improving if much weight is to be assigned to the vast majority of individual studies. The methods we now use are, I think, improved over those used in the evaluations conducted two decades ago whose results have done so much to create the pessimistic climate about the utility of evaluation. We now ask a broader range of questions than those concerning effects alone; we probe more at the theory behind a program; we use a more catholic array of techniques; we integrate evaluations better into past research; and we conduct surveys, experiments, and observational studies that are more sensitive to the practical constraints that were identified by the first generation of evaluators who ran headlong into them. I would like to see the organizations responsible for the welfare of evaluation not only stimulating syntheses about effective practices, but also beginning to summarize some of the mistakes of the past that we now know how to avoid or overcome.

It may also help create a more positive political climate for evaluation if studies are conducted of how evaluation results have been used in the

past. Boruch and Cordray (1980) completed such a study within education, and I was personally impressed by the levels of utilization reported (see also Leviton & Boruch, 1983). They are at obvious variance with the simplistic belief that nearly all evaluations lie on shelves and are not used, although there may still be many (indeed, too many) evaluations like this. My own experience with the Food and Nutrition Service of the Department of Agriculture suggests that evaluation reports are routinely discussed in Congress in deliberations about food stamps, the School Lunch Program, and the Women and Infant Children Feeding program, while work on the effects of television violence routinely gets exposure in newspapers and congressional hearings. My strong hunch is that disappointment at low levels of utilization is more commonplace among evaluators than anyone else and has led to an unnecessary and self-defeating overreaction. To combat such self-defeating, public pessimism, we badly need surveys of utilization, sector by sector.

Among the issues that should be explored in such surveys, four deserve special scrutiny. One deals with when the evaluation in question was conducted. My hypothesis is that disappointment with utilization developed largely out of the first generation of studies—of Head Start, Follow Through, the New Jersey Negative Income Tax Experiment, the Kansas City Patrol Experiment, and so one. These were the pioneering studies in which mistakes came to be identified, many of which were avoided in later studies of which I am aware. Since it would be sad if perceptions of evaluation's utility were based on studies whose inadequacies do not fully reflect more recent practices, I would like to see utilization studies relate use to the date of study, among other things.

The second issue worth special mention involves the sponsorship of evaluation. My impression is that studies where Congress mandates the questions produce results that are frequently used in policy deliberations, especially if the evaluations deal with options about ongoing programs. Also more often used, in my experience, are studies designed by federal departments without specific congressional demand. Less often used, I suspect, are studies conceived at the level of a local project (e.g., school district, drug abuse center, community mental health center). This seems to me especially the case when there has been a congressional mandate to evaluate but no specific evaluation questions have been set, no additional funds have been voted for evaluation, and

no perception has been generated that Congress has a real interest in the evaluation results (see Cook & Shadish, 1982, for efforts in mental health; David, 1978, for education; and Feeley & Sarat, 1980, for criminal justice).

The third issue concerns the definition of utilization. It should be defined in multiple ways to include evaluation results (1) constituting decisions; (2) being discussed in specific policy debates without constituting a decision; (3) increasing enlightenment through providing new perspectives that do not necessarily influence the program under review; and (4) being used to modify professional practice because reports of evaluations are cited either in the textbooks used for training future practitioners or in the in-service training provided to current practitioners.

The final point is related to the foregoing. Evaluation results can be used in many ways by many audiences. More stress has been placed in the past on types of usage than on types of audience. It seems to me myopic to constrain utilization to the efforts of formal decision makers and their staffs. Professional practice at the point of service delivery is also crucial, and practitioners come to hear about research results from a variety of sources. Journalists in all the mass media can hear about and disseminate research results that help create issues and mold public agenda and public opinions. Then, too, evaluation results can be used by scholars for secondary purposes that include theory probing or training students. Just as evaluation findings can be used many ways, they can also be used by many persons.

I am not trying to argue that all types of usage and all audiences are equally important; merely that in describing evaluation's total utilization they all deserve mention and descriptive examination. Nor am I arguing that surveys of utilization conducted by evaluators will be the last word in utilization studies. They will not be, for many commentators will consider presumptively self-serving any study of the use of evaluations done by professional evaluators. All I want to argue is that such studies may help to create a climate in which evaluations are viewed as useful and as being used for making important decisions.

Evaluations with More or Less Payoff

We turn now to the second major thrust I advocate for letting future administrations estimate the gain they could expect from investing in

evaluation. This involves determining the types of evaluation that promise most, and least, payoff. I understand types in two ways. The first concerns institutional arrangements for evaluation and the second touches on the sorts of questions that have most impact when evaluating ongoing social programs.

(a) Institutional Arrangements. Mandating evaluation is a blessing for those who want to be assured of work as evaluators. But I am not sure that all forms of mandated evaluation are useful. In particular I am struck by the frequency with which mandates to evaluate using in-house evaluators have proven to be unproductive. This seems to have been true in community mental health centers where Congress mandated that a given percentage of the budget had to be devoted, as a minimum, to evaluation. (It was 1% at the time, rising to 2% or more later.) But it seems that little new evaluation occurred as a result of the mandate (Cook & Shadish, 1981). Many reasons might explain this: Congress did not vote extra funds for evaluation; the funds per center were generally so small as to permit paying only a junior person and then not always full-time; this person rarely had much clout in the center; center directors rarely made their information needs salient to evaluators; and in rare cases where needs were made clear, they could not readily be answered. Finally, center directors usually wanted information for public relations reasons rather than for improving internal operations. In short, the in-house evaluators were low-power persons with conflicting demands, no clear idea of what they were supposed to do, and no audience waiting for their work. Something of the same order seems to have been the case with congressionally mandated in-house evaluators of local projects in both criminal justice and compensatory education. Rarely do we find empirical support for the guiding idea behind in-house formative evaluation: that in-house evaluators are especially useful because local decision makers trust them, and the evaluators know their projects so well that they can help generate significant research questions and develop realistic data collection techniques.

It is important to keep the above critique in perspective. There are undoubtedly some community mental health centers and school districts where evaluation is wanted, results are welcomed, and evaluators have the necessary technical and political skills. I can even think of some off-hand. However, for the reasons cited above, I do not believe that community mental health centers and school districts can, *in general,*

perform useful evaluative tasks on themselves. Nor do I believe that most of the obstacles to effective evaluation are likely to change in the immediate future.

(b) Focused Evaluands and Research Questions. A major consequence of the condemnation of global evaluation mandates is that we must be more careful in deciding what to evaluate and what we should seek to learn or discover when we do evaluate. In order to comment on where we might place evaluation priorities, some conceptual distinctions are called for.

Most federal and state social programs seem to be organizational umbrellas for managing funds and issuing regulations to the local projects that directly deliver services to the persons whose needs led to funding the program in the first place. Programs often have unclear or latent goals, and regulations are rarely so precise or so well monitored that individuals at the local level are totally constrained in what they can do. Indeed, the tradition of professional autonomy makes it very difficult for federal and state officials to tell physicians, teachers, and police officials what they should do. Try as they might, officials in Washington, D.C. cannot totally determine what happens in the many local projects—in schools, in police departments or in local centers of the Women, Infant, Children (WIC) Nutrition Program—that constitute the programs they administer from afar. Most social *programs* are composed, then, of many local *projects,* all doing somewhat different things in unique settings, enjoying multiple sources of funds, and knowing that few agencies have the resources for continuously monitoring local services. Services differ widely, not only from project to project but also from client to client and practitioner to practitioner within projects. Services are the elements of daily life over which practitioners have some control, and for this reason they are important as evaluands. Yet many elements seem trivial in potential impact when taken alone (e.g., whether we use textbook X or Y; whether pregnant women attend a center twice or three times a week; whether supplementary cheese is provided as part of a food package); consequently, judgments are always called for about the elements that might have more or less of an impact.

Related to the distinction between programs, projects, and elements is a distinction between cases and types. That is, one can select as an evaluand either a program like WIC or the total class of programs that

provide nutritional supplementation to the poor. The more general group would include the National School Lunch Program and food stamps, among others. Likewise, one could select a single project to study, such as a school in Evanston that has a bilingual project, or one could focus on different types of bilingual education (e.g., those that emphasize immersion [no courses in the native language], submersion [tutoring in the foreign language during English but all other classes in the English language], English as a Second Language [all substantive instruction in English but all English classes in the native language], and transition [all courses in the native language with a slow transition into English]). Finally, one can choose to evaluate a single element (e.g, the particular food supplementation package offered in WIC centers) or types of element (e.g., the amount of protein in the diet, which can be provided in many ways). Another example at the element level might be between the single element—doing more homework—or the more general type of element—"time on task." The latter is more general and can be manipulated as more homework, more class hours devoted to a particular topic, more school days, and so on.

The final set of distinctions I would like to make concern different types of evaluation question. Many questions concern who the audience or clients are for a program, project, or element, both the intended and actual clients. Other questions touch on the implementation of services, potentially emphasizing their quantity, quality, and appropriateness for individual clients. Next come questions about the effectiveness of an evaluand, and these questions typically deal with both intended and unintended effects. Related to questions about effects are what I call questions about impacts, these being the effects a program, project, or element has on the systems with which clients interact. Thus a drug abuse project might have effects on the persons who attend but might also have more remote impacts on addicts' families, local employment rates, or even attendance at other social welfare projects. Not to be forgotten, of course, are questions about causal processes—why is it that particular effects and impacts have come about? Indeed, some evaluation theorists have emphasized how knowledge of causal mediating variables can be used to facilitate transfer by permitting one to make sure that the effective causal mediators are present at sites where one is trying to recreate effects obtained elsewhere (Cronbach, 1982). Finally, questions can be asked about what a particular pattern of results means

for policy action and broader social values. To describe whom a program, project, or element reaches, how well it is implemented, and what effects and impacts it has, and to explain why such results occur, does not per se provide judgments about value or action implications. Questions about the meaning of results also need to be broached.

These distinctions lead to the larger issue: Which types of questions are more important with which types of evaluand? To that I will now turn.

The Program Level

Information about programs has a higher potential for leverage than information about local projects because programs reach more people. Information about programs also has a higher potential for leverage because any one client typically receives a mix of services from a program, and this mix promises more impact than a single element from the mix. However, programs rarely die, so that information about programs is rarely used to make stop or go decisions about a program. Even when budgets, administrations, and knowledge bases change, my guess is that programs do not cease; rather, they are modified internally by adding or removing projects or by modifying the funding for particular elements of practice. It is perhaps fortunate that turnover occurs more readily in projects and elements than in programs, for it is probably easier to obtain social and political acceptance for modifying the former than the latter. After all, changes in many elements will not cause heavy time or psychic burdens, and adding a new project does not. The dilemma here is that, *in general,* programs promise most impact if they are phased in or out, but phasing them in or out is not likely; elements promise least impact but most turnover; while projects are usually in-between on both counts.

Since programs consist of local projects, it is logically required that questions at the program level be based on representative samples of local projects. It is logistically complicated and difficult to pull off studies with such samples, and they have little utility if stop or go decisions are rarely made about programs. The information once advocated for making stop or go decisions is, of course, information about effects and impacts at the program level, and it may well be that such questions have little utility. However, information about programs may be used more often if large samples of projects provide information

about the number and type of clients who receive program services and about the quantity and quality of service implementation. Such knowledge can be used to modify program details in the hope of increasing efficiency, particularly if it is also linked to explanations of why the program is reaching some people but not others and why some elements are being implemented well but not others. The hypothesis, then, is that questions about the effectiveness and impact of a social program are less useful than questions that describe the audience and implementation of a program and attempt to explain why it is being implemented as it is. Of the various models for describing and explaining program operations I have been most impressed by Hendrick's (1981) Service Delivery Assessment model because of the required heterogeneity in the sites selected for visit, the multiple perspectives adopted in data collection, the critical way visitors to different sites discuss their observations, and the way impressions and recommendations are developed and defended so as to arrive at new proposals for action.

The Project Level

The foregoing analysis of the heterogeneous structure of programs suggests the utility of evaluations aimed at identifying the more successful types of projects within a program. When new projects are authorized, it might then be possible to specify which types to prefer for funding, thereby slowly influencing the mix of projects within a program through taking advantage of spontaneous turnover at the project level. The same considerations hold in the obverse case when budget contractions force some projects to close. Then, knowledge of successful project types might be used to guide the contraction process. As I am using the term, a "successful" project requires not only that services are delivered to the right targets with adequate quality, but also that—once delivered—the services influence the clients in desired ways and may even affect the social systems with which clients interact. Thus effectiveness is at the heart of evaluation at the project level, even while it is less relevant to evaluation at the program level.

It will not be easy to identify successful project types quickly or inexpensively. Commitment to the task also requires commitment to programmatic evaluation that, in a series of studies, describes the range

of projects within the program, classifies them as far as possible by type, uses correlational methods to identify some types as provisionally effective, and then uses more controlled methods to validate and refine the preliminary correlational evidence. Because of the many subtasks required for validly identifying successful project types, human impatience and government staff turnover will cause difficulties in gaining the necessary funding continuity and time horizons.

It is difficult to estimate the value of explanatory questions about causal process if project types are the focus of an evaluation effort. This is because the major purpose of explanation in applied research is to identify the key components of a treatment that have to be present for an effect to be reproducible. Yet if reliable information about project types has been produced, it can be presumed reproducible, since common effects would have to have been demonstrated across most of the instances in which a particular type of project was implemented. Thus strong inferences about successful project types would require reproducibility despite all the irrelevant differences between the projects constituting a type, thereby reducing the need for explanation. Explanatory evidence seems less crucial when project types are under review than when a whole program is and one wants to explore why the program has achieved a particular kind of audience or pattern of implementation.

The utility to evaluating single projects (rather than types of projects) is much less clear. Each project is typically only a small part of a program and is very local in its treatment delivery and potential effects. For this reason it does not have much of a profile in national or state capitols. Moreover, I am doubtful that feedback of any kind is produced that local officials or practitioners are likely to use. Yet most inferences about a program are based on accumulating data across individual projects, and sometimes the only way to do this is by examining the results from a restricted ad hoc collection of projects where evaluation happens to have been carried out. This is the case in many meta-analyses, the studies I would like to see done more often! However, at issue here are the kinds of evaluation questions that promise most payoff for those who commission evaluations. To fund the evaluation of an ad hoc collection of individual projects seems to me a shotgun approach to evaluation at the program level. Moreover, most of the meta-analyses with which I am acquainted are of project-level evalua-

tions conducted with local funds, often by graduate students in order to gain degrees. They did not come out of resources set aside for evaluation.

The Element Level

The patience and stability of policy agendas required to investigate project types within an ongoing social program are needed less when elements are the evaluands under study. If successful elements can be identified that are also manipulable and can be introduced into many of the projects in a program, this knowledge can be used by service providers to improve project performance, which, in its turn, should improve program performance. Given the marginal nature of most elements, the art is to identify transferrable elements that are not likely to be trivial in the magnitude of expected effects.

One factor that increases the importance of an element relates to the number of persons it can potentially affect. For instance, a change in the form schools use to assess eligibility for free or reduced-price school lunches constitutes a minor modification in the life of any parent or child, is a small savings per child per year, and might be expected to change the percentage of ineligible children who obtain subsidized lunches from about 10% to 5% of the program total. But given the numbers of children in the program, a decrease of this magnitude would result in a national savings of many millions of dollars per year, with the total being larger over the course of a child's career in school (AMS, 1984).

A second factor that increases the importance of a particular element concerns its role in influencing the most desired outcomes. Academic achievement is a valued outcome to most parents and teachers. Within the range it typically varies, the quality of classroom lighting is not likely to be closely linked to achievement—certainly not more so than, say, time on task. Similarly years of experience in teaching may be less directly linked to achievement than the extent to which the curriculum is completed within a school year. The underlying assumption making time on task and curriculum coverage so important is that coverage is presumed to be necessary for learning, while time on task is assumed to be sufficient. Within their usual range, the quality of lighting and years of teaching experience are probably neither necessary nor sufficient for effectiveness. When elements have been chosen very carefully for study

and meet the requirements of being manipulable, transferrable, and potentially impactful, they are well worth making the focus of evaluation efforts, for they constitute the procedures over which treatment deliverers have control at the point of service delivery.

SUMMARY

In this discussion I argued that few new opportunities exist for evaluators in the private and military sectors. This is because professions and companies already exist to provide evaluative services there, few complaints are heard about the quality of the services they provide, and it is not clear what special skills evaluators have that others could not easily adopt.

I then argued that evaluators might do better if they devoted their energies to preparing a case for future administrations in Washington and state capitals that were more sympathetic to the social welfare programs with which most members of ERS and ENet are experienced. In particular, I suggested that evaluators (1) should conduct reviews in the hope of identifying effective program variables and (2) should conduct analyses, sector by sector, of the uses to which evaluations have been put in the past two decades.

I further suggested that evaluators should also reexamine the major experiences we have had over the last 20 years in order to arrive at generalizations about the types of evaluation that seem more or less useful. In this context I suggested that (1) mandated evaluations conducted by in-house evaluators do not seem to bear much fruit, and I further suggested that (2) when the evaluand is a social program, questions about the audience and implementation should be paramount; (3) when the evaluand is at the project level, inferences about project types have greater utility than inferences about individual projects, and that for studying project types questions about effects and impacts should be paramount; and finally I suggested that (4) few elements are worth evaluating unless there is strong evidence that they are manipulable and potentially transferrable and might be presumed important because they reach many persons and have many small effects that might meaningfully cumulate across all the persons and projects in a program.

Not everyone may agree with the lessons I have drawn that emphasize that particular types of questions may differ in importance depending on what is being evaluated. However, I hope that few will disagree about the need to think through the sorts of evaluations most and least worth conducting. If this is done and some sort of consistency results, I believe that future administrations more sympathetic to social welfare will be more likely to view evaluation as a field that has benefited from the mistakes of its first generations and is prepared to proceed with a more refined sense of how to be useful in contributing to the amelioration of social problems. This is our ultimate goal, and I hope that the organizational imperative to maintain health will not let it slip from sight.

REFERENCES

Applied Management Sciences, Inc. (1984, January). *Income verification pilot project. Phase II: Results of quality assurance evaluation, 1982-83 school year.* 962 Wayne Ave., Silver Spring, MD 20910.

Boruch, R. F., & Cordray, D. S. (Eds.). (1980). *An appraisal of educational program evaluations: Federal, state, and local agencies.* Report to Congress. Evanston, IL: Northwestern University.

Cook, T. D., & Leviton, L. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48,* 449-472.

Cook, T. D., & Shadish, W. R., Jr. (1982). Meta-evaluation: An evaluation of the CMHC Congressionally-mandated evaluation system. In G. Stahler & W. R. Tash (Eds.), *Innovative approaches to mental health evaluation* (pp. 221-253). New York: Academic Press.

Cronbach, L. J. (1982). *The design of educational evaluations.* San Francisco: Jossey-Bass.

David, J. L. (1978). *Local uses of Title I evaluations.* Report prepared for Office of Assistant Secretary for Planning and Evaluation, DHEW. Menlo Park, CA: SRI International.

Feeley, M. M., & Sarat, A. D. (1980). *The policy dilemma: Federal crime policy and the Law Enforcement Assistance Administration 1968-1978.* Minneapolis: University of Minnesota Press.

Hendricks, M. (1981). Service delivery assessment: Quantitative evaluations at the cabinet level. In N. L. Smith (Ed.), *New directions for program evaluation: Federal efforts to develop new evaluation methods.* San Francisco: Jossey-Bass.

Leviton, L. C., & Boruch, R. F. (1983). Contributions of evaluation to educational programs and policy. *Evaluation Review, 7,* 563-598.

Light, R. J. (Ed.). (1983). *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Sage.

Lindblom, C. E., & Cohen, D. K. (1979). *Usable knowledge: Social science and social problem solving.* New Haven, CT: Yale University Press.

Scriven, M. (1983). *The logic of evaluation.* Monterey, CA: Evergreen Press.

Shadish, W. R., Jr. (in press). Policy research: Lessons from the implementation of deinstitutionalization. *American Psychologist.*

Weiss, J. A., & Weiss, C. H. (1981). Social scientists and decision makers look at the usefulness of mental health research. *American Psychologist, 36,* 837-847.