Statistical Procedures and the Justification of Knowledge in Psychological Science

Ralph L. Rosnow *Temple University*
Robert Rosenthal *Harvard University*

ABSTRACT: Justification, in the vernacular language of philosophy of science, refers to the evaluation, defense, and confirmation of claims of truth. In this article, we examine some aspects of the rhetoric of justification, which in part draws on statistical data analysis to shore up facts and inductive inferences. There are a number of problems of methodological spirit and substance that in the past have been resistant to attempts to correct them. The major problems are discussed, and readers are reminded of ways to clear away these obstacles to justification.

The exposure and castigation of error does not propel science forward, though it may clear a number of obstacles from its path.

—Medawar, 1969, p. 7

"Think Yiddish, write British" might be an apt slogan for the dominant discursive pattern of the intuitions and inductive inferences that characterize the scientific outlook in psychology during the entire 20th century. As is true in other fields, the inventive ways that psychological researchers think frequently seem to resemble the hunches and intuitions, the illogical as well as logical inferences, of an astute Jewish grandmother. Indeed, it has been observed that the progress of science, as much as the saga of human discoveries in all fields, is not a history of stunning leaps of logic but is often the outcome of "happy guesses" and "felicitous strokes of talent" in ostensibly unrelated situations (Grinnel, 1987, p. 24). The creative process in psychological science, as in all scientific disciplines, might be compared to the energy that excites a neuron in the human nervous system. The energy used to excite the neuron is nonspecific. The same ion flow occurs whether one hits one's finger with a hammer, burns it on the stove, or has it bitten by a dog. As long as the excitation is there, the result will be the same—ignition. In science, it also seems to make little difference as to what circumstances provide the inspiration to light the fuse of creativity. As long as the situation is sufficiently stimulating to excite thought in the scientist, there will be "ignition."

In contrast, the rhetoric of psychological science, the tightly logical outcome of this "thinking Yiddish," tends to be consistent with the traditions of British empiricist philosophy. As much as in all fields of science, journal articles and research monographs that describe the way in which the scientific method was used to open up the psychological world fail to communicate the day-to-day drama of the interplay of discovery and justification, in which speculative ideas based on facts, theories, intuitions, and hunches exert a constant influence on each other (cf. Knorr-Cetina, 1981; Mahoney, 1976; Mitroff, 1974). One reason for this situation may be that language, insofar as it is limited (Polanyi, 1967), imposes limitations on the ability of scientists to justify what they feel that they know. Another plausible reason is that the world's richness of information often exceeds our capacity to process it directly. As a result, the knower's representation of what is "out there" is, like any model of reality, reduced and distorted to fit in with his or her own available schematisms (McGuire, 1986).

In this article, we are concerned with various specific aspects of the rhetoric of justification, which in part draws on the strict logical consequences of statistical data analysis to shore up facts and inductive inferences. Despite the great range of procedures employed, there are some common problems of methodological spirit and methodological substance that although they have been addressed before, nevertheless endure. By exposing these problems again, we hope it may be possible to weaken their influence. In modern philosophy, a nautical analogy may be used to compare the progress of science to a boat that must be reconstructed not in drydock but at sea, one plank at a time. The aspects of statistical data analysis that we discuss might be thought of as the connecting tools that help us hold fast our facts and inductive inferences. In our reliance on statistical data-analytic tools used to reinforce the empirical foundation of psychological science, we want to choose the right tools for the job and to use them properly.

We begin by discussing four matters pertaining to the methodological spirit, or essence, of statistical data analysis. They are (a) the overreliance on dichotomous significance-testing decisions, (b) the tendency to do many research studies in situations of low power, (c) the habit of defining the results of research in terms of significance levels alone, and (d) the overemphasis on original studies and single studies at the expense of replications. We then turn to a consideration of some matters of methodological substance, or form. These are primarily problems in the teaching and usage of data-analytic procedures. The issues to be considered here are the use of omnibus or multivariate tests, the need for contrasts or focused tests of hypotheses, and the nearly universal misinterpretation of interaction effects.

 0003-066X/89/$00.75
Vol. 44, No. 10, 1276-1284

Matters of Methodological Spirit

Dichotomous Significance-Testing Decisions

Far more than is good for us, psychological scientists have for too long operated as if the only proper significance-testing decision is a dichotomous one, in which the evidence is interpreted as "anti-null" if p is not greater than .05 and "pro-null" if p is greater than .05. It may not be an exaggeration to say that for many PhD students, for whom the .05 alpha has acquired almost an ontological mystique, it can mean joy, a doctoral degree, and a tenure-track position at a major university if their dissertation p is less than .05. However, if the p is greater than .05, it can mean ruin, despair, and their advisor's suddenly thinking of a new control condition that should be run.

The conventional wisdom behind the approach goes something like this: The logic begins, more or less, with the proposition that one does not want to accept a hypothesis that stands a fairly good chance of being false (i.e., one ought to avoid Type I errors). The logic goes on to state that one either accepts hypotheses as probably true (not false) or one rejects them, concluding that the null is too likely to regard it as rejectable. The .05 alpha is a good fail-safe standard because it is both convenient and stringent enough to safeguard against accepting an insignificant result as significant. The argument, although not beyond cavil (e.g., Bakan, 1967), affords a systematic approach that many researchers would insist has served scientists well. We are not interested in the logic itself, nor will we argue for replacing the .05 alpha with another level of alpha, but at this point in our discussion we only wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?

Gigerenzer (1987; Gigerenzer & Murray, 1987; Gigerenzer et al., 1989), in discussions that examined the emergence of statistical inference, reminded us that the notion of dichotomous significance testing was initially developed out of agricultural experimentalists' need to answer questions such as, "Is the manure effective?" It is perhaps harder to object to the necessity of an accept–reject approach when the experimental question is phrased in precisely this way. However, the composition of the data base of psychological science, certainly, is substantively different, as would seem to be the phraseology of the research questions that psychological experimentalists try to answer. Indeed, Fisher at one point (largely as a reaction against the criticisms of Neyman and E. S. Pearson) voiced his strong objections to the idea of a fixed, dichotomous decision-level approach and instead argued for a cumulative, more provisional conception of statistical data analysis in science (as discussed in Gigerenzer, 1987, p. 24)—an idea that we will discuss in more detail.

To be sure, determining the particular level of significance of the data at which a null hypothesis will be rejected is essentially a personal decision, and by extension a decision by the field at a given historical moment. It is well known that in other scientific fields there is a strong tradition of rejecting the null hypothesis at an alpha level other than 5%. In using the Bonferroni procedure, scientists further redefine the alpha level so as to protect against post hoc selection of the largest effects (e.g., Harris, 1975; Morrison, 1976; Myers, 1979; Rosenthal & Rubin, 1984). The essential idea at this point in our discussion is that, from an ontological viewpoint, there is no sharp line between a "significant" and a "nonsignificant" difference; significance in statistics, like the significance of a value in the universe of values, varies continuously between extremes (Boring, 1950; Gigerenzer & Murray, 1987).[1]

Working With Low Power

Too often, it seems that psychologists do significance testing with low power as a consequence of ignoring the extent to which, in employing a particular size of sample, they are stacking the odds against reaching a given p value for some particular size of effect. One reason for this situation may be that even though the importance of the implications of the mechanics of power analysis for practice were recognized long ago by psychological statisticians, these mechanics were dismissed in some leading textbooks for a time as too complicated to discuss (e.g., Guilford, 1956, p. 217). However, as a consequence of a series of seminal works by Cohen beginning in the 1960s (e.g., Cohen, 1962, 1965), the concept resurfaced with a vengeance in psychological science.

No matter the reasons why a sense of statistical power was never fully inculcated in the scientific soul of laboratory experimental psychology, it cannot be denied that this situation has led to some embarrassing conclusions. Consider the following example (the names have been changed to protect the guilty): Smith conducts an experiment (with $N = 80$) to show the effects of leadership style on productivity and finds that style A is better than B. Jones is skeptical (because he invented style B) and replicates (with $N = 20$). Jones reports a failure to replicate;

This article was completed while Ralph Rosnow was a Visiting Professor at Harvard University and Robert Rosenthal was a Fellow at the Center for Advanced Study in the Behavioral Sciences. We wish to thank these institutions for the generous support provided us. We also wish to acknowledge the funding received by Rosnow from Temple University through the Bolton Endowment and a National Institute of Health Biomedical Research Support Grant, and that received by Rosenthal from the National Science Foundation and the John D. and Catherine T. MacArthur Foundation. Portions of this article are based on presentations at the annual meeting of the American Psychological Association by Rosenthal in 1987 and by Rosnow in 1988.

Correspondence concerning this article should be addressed to Ralph L. Rosnow, Department of Psychology, Temple University, Philadelphia, PA 19122.

[1] Interestingly, Enrico Fermi, the great physicist, thought $p = .10$ to be the wise operational definition of a "miracle" (Polanyi, 1961), and recent findings would lead us to believe that a similar standard might seem reasonable as a kind of "last ditch threshold" (i.e., before accepting the null hypothesis as true) to many psychological researchers (Nelson, Rosenthal, & Rosnow, 1986).

his t was 1.06, $df = 18$, $p > .30$, whereas Smith's t had been 2.21, $df = 78$, $p < .05$. It is true that Jones did not replicate Smith's p value. However, the magnitude of the effect obtained by Jones ($r = .24$ or $d = .50$) was *identical* to the effect obtained by Smith. Jones had found exactly what Smith had found even though the p values of the two studies were not very close. Because of the smaller sample size of 20, Jones's power to reject at .05 was .18 whereas Smith's power (N of 80) was .60—more than three times greater.

Table 1 helps us to examine this problem more deeply. It shows ratios of Type II to Type I errors for sample sizes from 10 to 1,000. Type I errors may be thought of as inferential errors of gullibility or overeagerness, that is, an effect or a relationship is claimed where none exists. Type II errors may be thought of as inferential errors of conservatism or blindness, that is, the existence of an effect or a relationship that does exist is denied (cf. Axinn, 1966). This table shows what may be conceptualized as the perceived seriousness of Type II to Type I errors for the conventional .05 level of p, the "miraculous" .10 level of p, and levels of r that are frequently characterized as small ($r = .10$), medium ($r = .30$), and large ($r = .50$) in psychological science, following Cohen's (1977) suggestion. For example, if the likelihood of a Type II error = .90 and the likelihood of a Type I error = .10, then the ratio of .90/.10 = 9 would tell us that the error in rejecting the null hypothesis when it is true (Type I error) is taken nine times more seriously than the error in failing to reject the null hypothesis when it is false (Type II error).

Table 1
Ratios of Type II to Type I Error Rates for Various Sample Sizes, Effect Sizes, and Significance Levels (Two-Tailed)

	Effect sizes and signficance levels					
	$r = .10$		$r = .30$		$r = .50$	
N	.05	.10	.05	.10	.05	.10
10	19	9	17	8	13	5
20	19	9	15	6	7	2
30	18	8	13	5	3	1
40	18	8	10	4	2	—
50	18	8	9	3	—	—
100	17	7	3	—	—	—
200	14	6	—	—	—	—
300	12	5	—	—	—	—
400	10	4	—	—	—	—
500	8	3	—	—	—	—
600	6	2	—	—	—	—
700	5	2	—	—	—	—
800	4	1	—	—	—	—
900	3	—	—	—	—	—
1,000	2	—	—	—	—	—

Note: Entries are to nearest integer; missing values <1.

Thus, the generally greater weight attached to the avoidance of Type I errors relative to Type II errors increases the smaller the effect size (i.e., r value), the smaller the N, and of course, the more stringent the p value. Although it might be argued that psychologists working in laboratories usually have plenty of power to detect even small effects because in laboratory experimentation error terms are often very small, we see that working simultaneously with a small effect, a small sample, and a binary decisional $p = .05$ might be compared to trying to read small type in a dim light: It is harder to make out the material. How much power is needed? Cohen (1965) recommended .8 as a convention for the desirable level of power. With a "small" effect (i.e., $r = .10$, $d = .20$), a power of .8 would require us to employ a total N of approximately 1,000 in order to detect various effects at $p = .05$, two-tailed (Cohen, 1977). With a "medium" effect (i.e., $r = .30$, $d = .63$), it would mean a total N of approximately 115 sampling units, and with a "large" effect (i.e., $r = .50$, $d = 1.15$) a total N of approximately 40 sampling units, to detect various effects at $p = .05$, two-tailed.[2] Given a typical medium-sized effect (Brewer, 1972; Chase & Chase, 1976; Cohen, 1962, 1973; Haase, Waechter, & Solomon, 1982; Sedlmeier & Gigerenzer, 1989), it would appear that psychological experimenters seemingly choose to work, or are forced to work by logistic constraints, in "dimly lit" rather than in "brightly lit" situations. This is not universally true in all fields, as we will show.

Defining Results of Research

The example of Jones and Smith would lead us to believe (quite correctly) that defining the results of research in terms of significance levels alone fails to tell the whole story. In his classic *Design of Experiments,* Fisher (1960) stated further that

> convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1% we do not . . . ever need to lose sight of the exact strength which the evidence has in fact reached, or to ignore the fact that with further trial it might come to be stronger or weaker. (p. 25)

He did not give specific advice on how to appraise "the exact strength" of the evidence, but the use of statistical power analysis, effect-size estimation procedures, and quantitative meta-analytic procedures (to which we refer later) enables us to do this with relative ease.

We have looked into power, and we now take another look at significance testing and effect-size estimation in the framework of a study with plenty of power overall. Before turning to this illustration, it may be worth reviewing the logic that insists that effect sizes be computed not only when p values in experimental studies are viewed as significant but also when they are viewed as nonsig-

[2] Small, medium, and large effects of d are conventionally defined as .2, .5, and .8, respectively, but we see that in actuality a somewhat larger effect of d is required when claiming correspondence with a medium or large effect of r (cf. Rosenthal & Rosnow, 1984, p. 361).

nificant. There are two good arguments for this recommended practice.

First, computing population effect sizes guides our judgment about the sample size needed in the next study we might conduct. For any given statistical test of a null hypothesis (e.g., t, F, χ^2, Z), the power of the statistical test (i.e., the probability of not making a Type II error) is determined by (a) the level of risk of drawing a spuriously positive conclusion (i.e., the p level), (b) the size of the study (i.e., the sample size), and (c) the effect size. These three factors are so related that when any two of them are known, the third can be determined. Thus, if we know the values for factors (a) and (c), we can easily figure out how big a sample we need to achieve any desired level of statistical power (e.g., Cohen, 1977; Kraemer & Thiemann, 1987; Rosenthal & Rosnow, 1984).

Second, it is important to realize that the effect size tells us something very different from the p level. A result that is statistically significant is not necessarily practically significant as judged by the magnitude of the effect. Consequently, highly significant p values should not be interpreted as automatically reflecting large effects. In the case of F ratios, a numerator mean square (MS) may be large relative to a denominator MS because the effect size is large, the N per condition is large, or because both values are large. On the other hand, even if considered quantitatively unimpressive according to the standards defined earlier, it could nevertheless have profound implications in a practical context.

The following example serves to illustrate that a test of significance without an effect size estimate gives an incomplete picture: In 1988, a major biomedical research study reported that heart attack risk in the population is cut by aspirin (Steering Committee of the Physicians' Health Study Research Group, 1988). This conclusion was based on the results of a five-year study of a sample of 22,071 physicians, approximately half of whom (11,037) were given an ordinary aspirin tablet (325 mg.) every other day, while the remainder (11,034) were given a placebo. Presumably, the way that aspirin works to reduce mortality from myocardial infarction is to promote circulation even when fatty deposits have collected along the walls of the coronary arteries. That is, aspirin does not reduce the chances of getting clotting but makes it easier for the transport of blood as the arteries get narrower. Part of the results of this study are shown in Table 2.

The top part of Table 2 shows the number of participants in each condition who did or did not have a heart attack. We see that 1.3% suffered an attack, and this event occurred more frequently in the placebo condition (1.7%) than in the aspirin condition (0.9%). Testing the statistical significance of these results yields a p value that is considerably smaller than the usual .05 decision cliff relied on in dichotomous significance testing, χ^2 (1, N = 22,071) = 25.01, $p < .00001$. This tells us that the results were very unlikely to be a fluke or lucky coincidence. However, when we compute the effect size (as a standard Pearson correlation coefficient), the result (r = .034) is so small as to be considered quantitatively unimpressive by methodological convention in our field.

Table 2
Aspirin's Effect on Heart Attack

Condition	MI absent	MI present
Presence or absence of MI in aspirin and placebo conditions		
Aspirin	10,933	104
Placebo	10,845	189
Binomial effect-size display of r = .034		
Aspirin	51.7	48.3
Placebo	48.3	51.7
Total	100.0	100.0
Fatal and nonfatal MIs in aspirin and placebo conditions		
	Nonfatal MI	Fatal MI
Aspirin	99	5
Placebo	171	18

Note. MI = myocardial infarction.

Nevertheless, the implications are far from unimpressive, and we see this more clearly when we recast this magnitude of effect into the language of a binomial effect-size display (Rosenthal & Rubin, 1979, 1982). In such a display, the results are translated for simplicity into dichotomous outcomes such as success versus failure, improved versus not improved, or in this case, myocardial infarction (MI) present versus MI absent. Because discussions of this technique are already available (e.g., Rosenthal & Rubin, 1979, 1982; Rosenthal & Rosnow, 1984; Rosnow & Rosenthal, 1988), it will suffice to note that its use to display the increase in success rate due to treatment more clearly communicates the real-world importance of treatment effects than do the commonly used effect-size estimators based on the proportion of variance accounted for. The middle part of Table 2 provides us with a binomial effect-size display that corresponds to the r = .034 effect size computed on the results in the top part. It suggests that approximately 3.4% fewer persons who would probably experience a myocardial infarction (i.e., given the particular conditions of this investigation) will not experience it if they follow the regimen as prescribed in the aspirin treatment condition.

The bottom part of Table 2 shows a small subset of the sample that participated in this investigation, consisting of those persons who actually suffered a heart attack during the five-year period of observation. In the aspirin condition 4.8% had a fatal heart attack, whereas in the placebo condition 9.5% had a fatal heart attack. It appears that mortality from myocardial infarction decreased by approximately one half as a result of aspirin taken every other day. When we compute the effect size, we find it to be more than twice the size (r = .08) of that

computed for the overall results. However, even though the effect size for the smaller subset is more impressive than the effect size for the entire sample, were we to rely on dichotomous significance testing for a yes-or-no-decision we would be led to not reject the null hypothesis. That is because χ^2 (1, N = 293) = 2.06, p = .08 for the results in the bottom part of Table 2. Inasmuch as the sample size is relatively small, as seen in the context of the "small" magnitude of effect, we are operating with much less power than we were in the top part of Table 2. What is the lesson? Given the low level of power (<.4), this aspect of the investigation should be continued with a larger sample size before deciding that nothing happened.

Before leaving this section, it will be instructive if we briefly discuss the limitations of the findings in this study to underscore the idea that strength of effect is very context dependent. First, the sample in this study consisted entirely of male physicians, and the statistical results may not generalize in exactly the same way to the population at large. Furthermore, in prescribing aspirin, the physician would want to know about the medical history of the patient because the effects of aspirin could be dangerous to persons with ulcers, high blood pressure, kidney problems, or allergies to aspirin or who are about to undergo surgery. Thus, a further lesson is that, like a word or phrase framed by the context in which it is situated, it is important not to strip away the context from the content of a research study as we attempt to frame particular implications of the results.

Second, there is a growing awareness in psychology that just about everything under the sun is context dependent in one way or another (e.g., Gergen, 1973; Hayes, 1987; Hoffman & Nead, 1983; Jaeger & Rosnow, 1988; Lerner, Hultsch, & Dixon, 1983; McGuire, 1983; Mishler, 1979; Rosnow, 1978, 1981; Rosnow & Georgoudi, 1986; Sarbin, 1977; Smith, 1988; Veroff, 1983). Strength of effect measures are no exception, and it is therefore important to recognize how the study characteristics might influence the size as well as one's interpretation of the magnitude-of-effect estimate (e.g., Murray & Dosser, 1987; Rosenthal & Rosnow, 1984; Rosenthal & Rubin, 1979, 1982; Strube, 1988).

Overemphasis on Single Studies

The final matter of methodological spirit to be discussed concerns the importance of replication, a concept to which psychological journal editors, textbook writers, and researchers pay considerable lip service. In practice, however, the majority of editors, as much as most researchers, seem to be biased in favor of single studies at the expense of replications. Sterling (1959) found not a single replication in his classic review of experimental articles in four psychological journals during one year, and this practice does not appear to have changed much in more recent years (Mahoney, 1976).

Is it possible there are sociological grounds for this monomaniacal preoccupation with the results of a single study? Might those grounds have to do with the reward system of science, in which, in our perceptions, as much as in the realities of many academic institutions, merit, promotion, and the like depend on the results of the single study, which is also known as the "smallest unit of academic currency"? The study is "good," "valuable," and above all, "publishable" when $p < .05$. Our discipline might be farther ahead if it adopted a more cumulative view of science. The operationalization of this view would involve evaluating the impact of a study not strictly on the basis of the particular p level, but more on the basis of multiple criteria, including its own effect size as well as the revised effect size and combined probability that resulted from the addition of the new study to any earlier studies investigating the same or a similar relationship. This, of course, amounts to a call for a more meta-analytic view of doing science.

The name, meta-analysis, was coined by Glass (1976) to refer to the summarizing enterprise, although the basic quantitative procedures for combining and comparing research results were known some years earlier (Mosteller & Bush, 1954; Snedecor, 1946). Because numerous texts and articles are available on this subject (e.g., Cooper, 1984; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Mullen & Rosenthal, 1985; Rosenthal, 1984), we will mention only two, more or less secret, benefits to the research process of conducting meta-analytic reviews of research domains: the "new intimacy" and the "decrease in the splendid detachment of the full professor."

First, this new intimacy is between the researcher and the data. We cannot do a meta-analysis by reading abstracts and discussion sections. We have to look at the numbers and, very often, compute the correct ones ourselves. Meta-analysis requires us to cumulate *data,* not *conclusions.* Reading an original-study article is quite a different matter when one needs to compute an effect size and a fairly precise significance level—often from a results section that provides no information on effect sizes or precise significance levels. The *Publication Manual of the American Psychological Association* (American Psychological Association [APA], 1983) insists that when reporting inferential statistics, authors give the symbol, degrees of freedom, value, and probability level. The APA manual does not require that an exact significance level or the estimated effect size be reported, but what a boon it would be for meta-analysts if all journal editors required that authors also report all of their analyses and findings to even this limited extent.

Second, closely related to the first benefit is a change in *who* does the reviewing of the literature. Meta-analytic work requires careful reading of research and moderate data-analytic skills. One cannot send an undergraduate research assistant to the library with a stack of 5 × 8 cards to bring back "the results." With narrative reviews that seems often to have been done. With meta-analysis the reviewer must get involved with the data, and that is all to the good because it results in a decrease in the splendid detachment of the full professor.

There are other benefits of replications that are well

known to scientists. The fact that the results can be repeated ensures the robustness of the relationships reported. The results also can be repeated by uncorrelated replicators (i.e., truly independent experimenters) in different situations, which ensures the further generality of the relationships. In spite of the recognized methodological and epistemological limitations, the importance of replications is supported by quite different methodological theories as essential in a pragmatic sense (e.g., Bakan, 1967; Brewer & Collins, 1981; Houts, Cook, & Shadish, 1986; Rosenthal & Rosnow, 1984; Rosnow, 1981).

Matters of Methodological Substance

Omnibus Tests

The first problem of methodological substance concerns the overreliance on omnibus tests of diffuse hypotheses that although providing protection for some investigators from the dangers of "data mining" with multiple tests performed as if each were the only one considered, do not usually tell us anything we really want to know. As Abelson (1962) pointed out long ago in the case of analysis of variance (ANOVA), the problem is that when the null hypothesis is accepted, it is frequently because of the insensitive omnibus character of the standard F-test as much as by reason of sizable error variance. All the while that a particular predicted pattern among the means is evident to the naked eye, the standard F-test is often insufficiently illuminating to reject the null hypothesis that several means are statistically identical.

For example, suppose the specific question is whether increased incentive level improves the productivity of work groups. We employ four levels of incentive so that our omnibus F-test would have 3 dfs in the numerator or our omnibus chi square would be on at least 3 dfs. Common as these omnibus tests are, the diffuse hypothesis tested by them usually tells us nothing of importance about our research question. The rule of thumb is unambiguous: Whenever we have tested a fixed effect with $df > 1$ for chi square or for the numerator of F, we have tested a question in which we almost surely are not interested.

The situation is even worse when there are several dependent variables as well as multiple degrees of freedom for the independent variable. The paradigm case here is canonical correlation, and some special cases are multiple analysis of variance (MANOVA), multiple analysis of covariance (MANCOVA), multiple discriminant function, multiple path analysis, and complex multiple partial correlation. Although all of these procedures have useful exploratory data-analytic applications, they are commonly used to test null hypotheses that are scientifically almost always of doubtful value (cf. Huberty & Morris, 1989). Furthermore, the effect size estimates they yield (e.g., the canonical correlation) are also almost always of doubtful value. Although we cannot go into detail here, one approach to analyzing canonical data structures is to reduce the set of dependent variables to some smaller number of composite variables and to analyze each composite serially (Rosenthal, 1987).

Contrast Analysis

Whenever we have $df > 1$ for chi square or for the numerator of an F-test, we would argue that contrasts become the appropriate data-analytic procedure given the usual situation of fixed effect analyses (Rosenthal & Rosnow, 1984, 1985; Rosnow & Rosenthal, 1988). Briefly, contrasts are 1 df tests of significance for comparing the pattern of obtained group means to predicted values, with predictions made on the basis of theory, hypothesis, or hunch. Among the practical advantages of contrasts are that they can be easily computed with a pocket calculator, can be computed on the data in published reports as well as with original data, and most important, usually result in increased power and greater clarity of substantive interpretation.

Writing over 25 years ago, Abelson (1962) made a strong case for the method of contrasts and its wide range of varied uses. Why this method, which goes back virtually to the invention of ANOVA, had not previously received a comprehensive, unified treatment was a mystery. He speculated that "one compelling line of explanation is that the statisticians do not regard the idea as mathematically very interesting (it is based on quite elementary statistical concepts) and that quantitative psychologists have never quite appreciated its generality of application" (p. 2). Later, a number of issues at the heart of Abelson's thesis were picked up by other authors working in quite different areas of psychology, but these efforts did not have any definite practical impact on the teaching and usage of data-analytic procedures.

For example, Hale (1977) demonstrated the utility of carrying out contrasts in the area of developmental research. He computed a contrast F-test to reanalyze a portion of another investigator's published data concerning the effects of a vigilance distractor on recall of relevant and irrelevant information. In the published study, 40 children per grade in the first, third, fifth, and seventh grades were instructed to attend to one element in each of several two-element pictures in order to perform what was represented as a memory game. Half of the participants were tested under distraction conditions in which a melody of high notes on a piano was interrupted periodically by single low-pitch notes. Incidental learning was assessed by asking the children which elements appeared together in each picture. The mean scores for the distraction and no distraction conditions at these four grade levels, respectively, were distraction, 2.6, 2.3, 2.5, and 1.7; no distraction, 1.8, 2.4, 2.2, and 2.7. In the original published report the developmental change in treatment effect was tested by an omnibus F for the interaction of age by treatment, which the investigator found to be nonsignificant, that is, $F(3, 152) = 1.9$, $p = .13$. Hale reanalyzed the results by carving a focused F or contrast analysis between treatment and age trend out of the interaction sum of squares, which he found statistically significant, that is, $F(1, 152) = 4.3$, $p = .04$.

Discussions of contrasts have been primarily within the context of ANOVA, but their use is not restricted to this situation (cf. Bishop, Fienberg, & Holland, 1975; Rosenthal, 1984; Rosenthal & Rosnow, 1984, 1985). For example, Donald Rubin (in Rosenthal & Rosnow, 1985, pp. 48–49) has shown how contrasts can also be used when the obtained values are cast as frequency counts in a 2 × C contingency table, in which the classes in one classification are ordered and the classes in the other classification are expressed as a proportion (see also Snedecor & Cochran, 1967, p. 247). Although most current textbooks of statistics describe the logic and the machinery of contrast analysis, one still sees contrasts employed all too rarely. That is a real pity given the precision of thought and theory they encourage and (especially relevant to these times of publication pressure) given the boost in power conferred with the resulting increase in .05 asterisks.

Interaction Effects

The final matter to be discussed concerns what are probably the universally most misinterpreted empirical results in psychology, the results of interaction effects. A recent survey of 191 research articles employing ANOVA designs involving interaction found only 1% of the articles interpreting interactions in an unequivocally correct manner (Rosnow & Rosenthal, 1989). The mathematical meaning of interaction effects is unambiguous, and textbooks of mathematical and psychological statistics routinely include proper definitions of interaction effects. Despite this, most of the textbooks in current usage and most psychological researchers reporting results in our primary journals interpret interactions incorrectly. The nature of the error is quite consistent. Once investigators find significant interactions they attempt to interpret them by examining the differences among the original cell means, that is, the simple effects. However, it is no secret that these condition means are made up only partially of interaction effects; main effects may contribute to simple effects even more than interactions (e.g., Lindquist, 1953). The origin of the problem, as Dawes (1969) suggested, may in part be a consequence of "the lack of perfect correspondence between the meaning of 'interaction' in the analysis of variance model and its meaning in other discourse" (p. 57). Whatever its etiology, however, the error of looking only to the uncorrected cell means for the pattern of the statistical interaction is deeply rooted, indeed.

Because we have discussed the treatment of this problem in some detail recently (Rosenthal & Rosnow, 1984; Rosnow & Rosenthal, 1989), we merely note here that if investigators are claiming to speak of an interaction, the exercise of looking at the "corrected" cell means is absolutely essential. Of course, this should not be viewed as an argument against comparing cell means (i.e., simple effects tests), as it often makes sense to focus on a comparison of means using planned contrasts and to deemphasize the traditional main and interaction effects when they are based only on omnibus *F*-tests. Our point here is that the interaction effect is defined basically in terms of the residuals, or leftover effects, after the lower order effects have been removed from the original cell means. This is true even though the mean square for interaction in the ANOVA can be viewed as variability of the differences between the (uncorrected) cell means for the various rows of the table of overall effects. That is, the mean square for interaction will have a nonzero value if the difference between any two cell means in any row differs from the corresponding difference in any other row. Nonetheless, in focusing attention only on the original cell means, one is essentially ignoring the form and degree of relationship of the interaction itself. Like peeling away the skins of an onion, we need to peel away the lower order effects in order to separate the effects of the interaction from the main effects.

The problem is compounded because users of SPSS, SAS, BMDP, and virtually all data-analytic software are poorly served in the matter of interactions. Almost no programs provide tabular output giving the residuals defining interaction. The only exception to that, of which we are aware, is a little-known package called Data-Text, developed by Armor and Couch (1972) in consultation with leading statisticians including William Cochran and Donald Rubin. Researchers claiming to speak of an interaction must avoid the pitfall described in the anecdote of the drunkard's search. A drunk man lost his house key and began searching for it under a street lamp, even though he had dropped the key some distance away. When he was asked why he did not look where he had dropped it, he replied, "There's more light here!" This principle teaches that looking in a convenient place but not in the right place will never yield the key that will answer the question.

A Final Note

We have examined a number of aspects of the rhetoric of justification, which in part depends on statistical data analysis to shore up facts and inductive inferences. In particular, we have exposed several problems of methodological spirit and substance that have become deeply rooted in psychological science. Because of the unifying influence of the institutionalization of the classical procedure, we have sought in this discussion to review some ways of improving it rather than to argue for an alternative procedure for statistical inference (e.g., Goodman & Royall, 1988). Much of what we have said has been said before, but it is important that our graduate students hear it all again so that the next generation of psychological scientists is aware of the existence of these pitfalls and of the ways around them.

REFERENCES

Abelson, R. P. (1962). *Testing a priori hypotheses in the analysis of variance.* Unpublished manuscript, Yale University, New Haven, CT.

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: Author.

Armor, D. J., & Couch, A. S. (1972). *Data-text primer: An introduction to computerized social data analysis.* New York: Free Press.

Axinn, S. (1966). Fallacy of the single risk. *Philosophy of Science, 33,* 154–162.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation.* San Francisco, CA: Jossey-Bass.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts.

Brewer, J. K. (1972). On the power of statistical tests in the *American Educational Research Journal. American Educational Research Journal, 9,* 391–401.

Brewer, M. B., & Collins, B. E. (Eds.). (1981). *Scientific inquiry and the social sciences: A volume in honor of Donald T. Campbell.* San Francisco, CA: Jossey-Bass.

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61,* 234–237.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1973). Statistical power analysis and research results. *American Educational Research Journal, 10,* 225–229.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cooper, H. M. (1984). *The integrative research review: A systematic approach.* Beverly Hills, CA: Sage.

Dawes, R. M. (1969). "Interaction effects" in the presence of asymmetrical transfer. *Psychological Bulletin, 71,* 55–57.

Fisher, R. A. (1960). *Design of experiments* (7th ed.). Edinburgh, Scotland: Oliver & Boyd.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26,* 309–320.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Kruger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution* (Vol. 2, pp. 11–33). Cambridge, MA: Bradford/MIT Press.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge, England: Cambridge University Press.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health, 78,* 1568–1574.

Grinnell, F. (1987). *The scientific attitude.* Boulder, CO: Westview Press.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29,* 58–65.

Hale, G. A. (1977). On use of ANOVA in developmental research. *Child Development, 48,* 1101–1106.

Harris, R. J. (1975). *A primer of multivariate statistics.* New York: Academic Press.

Hayes, S. C. (1987). A contextual approach to therapeutic change. In N. Jacobson (Ed.), *Psychotherapists in clinical practice: Cognitive and behavioral perspectives* (pp. 327–387). New York: Guilford.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hoffman, R. R., & Nead, J. M. (1983). General contextualism, ecological sciences and cognitive research. *Journal of Mind and Behavior, 4,* 507–560.

Houts, A. C., Cook, T. D., & Shadish, W., Jr. (1986). The person–situation debate: A critical multiplist perspective. *Journal of Personality, 54,* 52–105.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105,* 302–308.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Jaeger, M. E., & Rosnow, R. L. (1988). Contextualism and its implications for psychological inquiry. *British Journal of Psychology, 79,* 63–75.

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

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Beverly Hills, CA: Sage.

Lerner, R. M., Hultsch, D. F., & Dixon, R. A. (1983). Contextualism and the character of developmental psychology in the 1970s. *Annals of the New York Academy of Sciences, 412,* 101–128.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston, MA: Houghton Mifflin.

Mahoney, M. J. (1976). *Scientist as subject: The psychological imperative.* Cambridge, MA: Ballinger.

McGuire, W. J. (1983). A contextual theory of knowledge: Its implications for innovation and reform in psychological research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 16, pp. 1–47). New York: Academic Press.

McGuire, W. J. (1986). A perspectivist looks at contextualism and the future of behavioral science. In R. L. Rosnow & M. Georgoudi (Eds.), *Contextualism and understanding in behavioral science: Implications for research and theory* (pp. 271–301). New York: Praeger.

Medawar, P. B. (1969). *Induction and intuition in scientific thought.* Philadelphia, PA: American Philosophical Society.

Mishler, E. G. (1979). Meaning in context: Is there any other kind? *Harvard Educational Review, 49,* 1–19.

Mitroff, I. (1974). *The subjective side of science: A philosophical inquiry into the psychology of the Apollo moon scientists.* New York: Elsevier.

Morrison, D. F. (1976). *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method* (pp. 289–334). Cambridge, MA: Addison-Wesley.

Mullen, B., & Rosenthal, R. (1985). *BASIC meta-analysis: Procedures and programs.* Hillsdale, NJ: Erlbaum.

Murray, L. W., & Dosser, D. A., Jr. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34,* 68–72.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston, MA: Allyn & Bacon.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299–1301.

Polanyi, M. (1961). The unaccountable element in science. *Transactions of the Bose Research Institute, 24,* 175–184.

Polanyi, M. (1967). *The tacit dimension.* London, England: Routledge & Kegan Paul.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis.* New York: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* Cambridge, England: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Psychology, 9,* 395–396.

Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rosenthal, R., & Rubin, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology, 76,* 1028–1034.

Rosnow, R. L. (1978). The prophetic vision of Giambattista Vico: Implications for the state of social psychological theory. *Journal of Personality and Social Psychology, 36,* 1322–1331.

Rosnow, R. L. (1981). *Paradigms in transition: The methodology of social inquiry.* New York: Oxford University Press.

Rosnow, R. L., & Georgoudi, M. (Eds.). (1986). *Contextualism and understanding in behavioral science: Implications for research and theory.* New York: Praeger.

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology, 35,* 203–208.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Sarbin, T. R. (1977). Contextualism: A world view for modern psychology. In J. K. Cole & A. W. Landfield (Eds.), *Nebraska symposium on motivation* (Vol. 24, pp. 1–41). Lincoln, NE: University of Nebraska Press.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an impact on the power of studies? *Psychological Bulletin, 105,* 309–316.

Smith, M. B. (1988). Beyond Aristotle and Galileo: Toward a contextualized psychology of persons. *Theoretical and Philosophical Psychology, 8,* 2–15.

Snedecor, G. W. (1946). *Statistical methods.* Ames, IA: Iowa State College Press.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames, IA: Iowa State University Press.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318,* 262–264.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Strube, M. J. (1988). Some comments on the use of magnitude-of-effect estimates. *Journal of Counseling Psychology, 35,* 342–345.

Veroff, J. (1983). Contextual determinants of personality. *Personality and Social Psychology Bulletin, 9,* 331–343.