MPR Reference No.: 8813-300

MATHEMATICA
Policy Research, Inc.

Nonexperimental Replications of Social Experiments: A Systematic Review

Interim Report/ Discussion Paper

September, 2002

Steven Glazerman
Dan M. Levy
David Myers

Submitted to:

Smith Richardson Foundation
60 Jesup Road
Westport, CT 06880
(202) 222-6222

Project Officer:
Phoebe Cottingham

Submitted by:

Corporation for the Advancement of Policy Evaluation
600 Maryland Avenue, Suite 500
Washington, DC 20024
(202) 484-9220

Principal Investigator:
David Myers

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

ACKNOWLEDGMENTS

This research was supported by grants from the William and Flora Hewlett Foundation and the Smith Richardson Foundation; however, the findings and conclusions do not necessarily represent the official position or policies of the Hewlett Foundation or the Smith Richardson Foundation.

Many people contributed to this research. Phoebe Cottingham of the Smith Richardson Foundation provided useful direction and advice. We received helpful comments and clarifications from authors of the studies we reviewed, including Wang Lee, Ty Wilde, Rob Hollister, Jeff Smith, Howard Bloom, Rob Olsen, Roberto Agodini, and Robert Lalonde. Members of the Campbell Collaboration Methods Group also provided helpful comments. In particular, we thank Jeff Valentine and Harris Cooper for their specific guidance on conducting Campbell reviews, and the participants at a conference organized by Phoebe Cottingham in July 2002, including Larry Orr, Robert Moffitt, Rob Hollister, and Steve Bell.

We would also like to thank our colleagues at Mathematica including Mark Dynarski and Allen Schirm, who read the study carefully, and many others who provided insightful comments. Jay Bhalodia provided able research assistance and Sally Henderson assisted with the many complex literature searches. The report was edited by Carol Soble and Daryl Hall and was produced by Melanie Lynch.

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

CONTENTS

Chapter **Page**

TABLES

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

FIGURES

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

I. BACKGROUND AND OBJECTIVE

Controlled experiments, where subjects are randomly assigned to receive interventions, are desirable but frequently perceived to be infeasible or overly burdensome, especially in social settings. Therefore, nonexperimental (also called quasi-experimental) methods are often used instead. Quasi-experimental methods are less intrusive and sometimes less costly than controlled experiments, but their validity rests on particular assumptions that are often difficult to test.[1]

It is therefore important to find empirical evidence to assess the likelihood that a given method applied in a given context will yield unbiased estimates. The current study is a systematic review of validation research to better understand the conditions under which quasi-experimental methods most closely approximate the results that would be found in a well-designed and well-executed experimental study. We collect and summarize a set of earlier studies that each tried, using convenience samples and one or more quasi-experimental methods, to replicate the findings from a social experiment. Our synthesis aims to give both producers and consumers of social program evaluations a clear understanding of what we know and what we do not know about the performance of quasi-experimental evaluation methods.

A. QUESTIONS GUIDING THIS RESEARCH

The research aims to address the following questions:

- Can quasi-experimental methods approximate the results from a well-designed and well-executed experiment?
- Under what conditions can nonexperimental methods produce unbiased program impact estimates?
- Can the bias from a single nonexperimental impact estimate be cancelled out or offset by aggregating multiple nonexperimental impact estimates?

[1] The terms "quasi-experimental" and "nonexperimental" are used interchangeably here.

The answers to these questions will be useful to policymakers and researchers in two ways. First, they will inform our decisions about whether and how to consider quasi-experimental evidence when we review or synthesize research on the effectiveness of social interventions. This issue holds particular interest for the members of the Campbell Collaboration, an international body of scholars dedicated to the production and dissemination of systematic reviews of such research.

Second, our study will inform those who sponsor or conduct evaluation research. By examining nonexperimental replications of social experiments, we can begin to learn the conditions under which, when random assignment is not feasible, it would be acceptable to rely on a quasi-experimental research design. We can also provide insight into which quasi-experimental approach is most promising in a given context. This information is critical for government agencies, private foundations, and evaluation researchers as they plan future evaluations.

B. STATE OF THE FIELD BEING REVIEWED

For decades, vigorous debate has focused on the most appropriate methods for studying the effectiveness of social programs, with experimental design pitted against a wide array of alternatives. Most of the discussion, however, has remained theoretical. (See Campbell and Stanley 1966; Cook and Campbell 1979; Chalmers et al. 1983; Wortman 1996; Troia 1998; Burtless 1995; and Heckman and Smith 1995).

The more recent phenomenon is the accumulation of *empirical* evidence that might address the questions we raised above. Researchers have used two types of empirical evidence to assess nonexperimental methods: between-study comparisons and within-study comparisons (Shadish 2000). The current study synthesizes the evidence from within-study comparisons, but we describe between-study evidence as background.

Between-study comparisons look at two or more research studies that each tried to estimate the same parameter (e.g., the effectiveness of a program) using different research designs and study samples. By comparing results from all the studies that used experimental designs with those that used nonexperimental designs and methods, researchers try to estimate the relationship between the research design and the program impact.

A good source of between-study design comparisons are meta-analyses. A common practice in meta-analysis is to report the effect size separately for studies that used random assignment to assign subjects to treatments and for studies that used nonrandomized comparison groups or some other design.[2] Some meta-analysts have made even finer distinctions in research design to explore more systematically the relationship between study design and overall findings (Shadish and Ragsdale 1996), but detailed information on study designs beyond experimental/nonexperimental or beyond idiosyncratic quality rating scales is rare.

One study (Lipsey and Wilson 1993) conducted a meta-analysis of meta-analyses to address the question of how research design influences findings. The authors found 74 meta-analyses that distinguished between randomized and nonrandomized treatment assignment and showed that the average effect sizes for the two were similar, 0.46 of a standard deviation for experimental design and 0.41 for nonexperimental design. But Lipsey and Wilson examined meta-analyses in a wide range of content domains, spanning nearly the entire applied psychology literature. Their comparison of mean effect sizes masks the variation that might occur both within and across content domains. To illustrate more directly, they graph the distribution of differences between random and nonrandom treatment assignment for each individual meta-

[2] For recent examples, see Cooper et al. (2000), Table 2, or National Research Council (2000), Chapter I, Tables 6–7.

analysis (where each one pertains to a single content domain). They found that while the average difference between findings based on experimental versus nonexperimental designs was close to zero, implying no bias, the range extended from about -1.0 standard deviation to +1.6 standard deviations, with the bulk of differences falling between -0.20 and + 0.40. Their finding implies that for many types of interventions, the average of the quasi-experimental studies gives a slightly different answer from the average of the experimental studies, and, for some, it gives a markedly different answer. This between-study evidence still leaves open the question of whether differences in impact estimates are due to design or to some other factor.

Within-study comparisons are single studies of one intervention whereby researchers estimate a program's impact by using a randomized control group and then estimate the same impact by using one or more nonrandomized comparison groups. We refer to these studies as "design replication" studies. The nonrandomized comparison groups are formed and adjusted by using statistical or econometric techniques designed to estimate or eliminate the bias attributable to self selection of different kinds of people into treatment and comparison conditions. Some design replication studies use multiple comparison groups or the same comparison group with multiple sample restrictions to examine the effect of different comparison group strategies. The nonexperimental estimate is meant to mimic what would have been estimated if a randomized experiment had not been possible. If the nonexperimental estimate is close to the experimental estimate, then the nonexperimental technique is assumed to be "successful" at replicating an unbiased research design.

The advantage of within-study comparisons is that they make clear that the difference in findings between methods is attributable to the method itself rather than to the researcher, the intervention, or the study context. Therefore, any within-study comparison yields a valid estimate of the bias. On the other hand, a disadvantage is that it is more difficult to rule out

chance when comparing a small number of experimental and nonexperimental estimates. It is therefore necessary to conduct several different within-study comparisons in a variety of contexts before drawing any general conclusions.

The corpus of within-study comparisons, or design replications, is just reaching a critical mass. The often-cited early efforts in the area under review are Lalonde's (1986) and Fraker and Maynard's (1987) studies of the National Supported Work demonstration, which cast doubt on the ability of traditional econometric estimators to yield unbiased impact estimates. Lalonde's data have undergone reanalysis twice in the last few years, by Dehejia and Wahba (1999) and by Smith and Todd (2002), each time incorporating more estimators or different criteria for defining the comparison sample. In the meantime, many other researchers (see reference list) have conducted design replication studies by using data from other randomized experiments in job training, welfare-to-work, and education. Most of these studies were completed in the last three years, and each uses slightly different techniques to estimate bias and weigh its statistical and policy significance. At least two more such studies are in progress and should be completed in time for inclusion in a later version of this review.

To our knowledge, no systematic reviews of the literature on within-study comparisons have been conducted to date. Despite some attempts to discuss the history of these studies (Shadish 2000), no one has integrated the studies to draw out their lessons, which is our goal. Nearly every one of the component studies of this review includes a brief summary of the literature that preceded it; to the best of our knowledge, the most comprehensive summary is the most recent one (Bloom et al. 2002), which addresses an important subset of this literature, the portion dealing with mandatory welfare programs. The current review tries to go beyond the literature summaries by casting a wider net, standardizing the units of the bias, standardizing the description of the nonexperimental methods, and making quantitative comparisons across

studies. The present interim report includes preliminary comparisons and discusses the feasibility of a more formal meta-analysis to be completed as a next step.

C. ORGANIZATION OF THIS REPORT

Chapter II lays out the conceptual framework, major hypotheses, and estimation issues. Chapter III describes the methods as well as the data (the source studies) used in conducting the review. Chapter IV provides preliminary findings, to be expanded in a future revision of this report. Chapter V discusses the implications of the preliminary findings for the practical questions raised above and lays out some of the challenges for proceeding with a more rigorous meta-analysis for the final report.

II. FRAMEWORK

Given that our topic is methodological, the current review differs from a typical systematic review of the effects of a social intervention. The "effect size" of interest is not the impact of some intervention on a given outcome but rather the estimated size of the difference between the experimental and the nonexperimental impact estimates. This quantity, an estimate of the bias attributable to the nonexperimental method, will differ for each nonexperimental estimate reported in each study. Thus, each study may report several bias estimates, each pertaining to a unique combination of the comparison group and the method of adjusting for differences between treated and untreated subjects. In fact, for every combination of comparison group and method there are often multiple sample definitions, time periods, subgroups, or study sites, resulting in dozens or hundreds of bias estimates. A model is helpful for understanding the relationship between bias estimates and the potential sources of bias as well as the alternative explanations for why program impact estimates vary. A major challenge in constructing such a model is accounting for the nonindependence among estimates that are generated from the same experiment or even the same comparison group members. For example, ten bias estimates from a single study should receive less weight than ten studies with one estimate each. This chapter provides a framework that formalizes these ideas and allows us to generate and test hypotheses about the magnitude of bias under different conditions.

A. MODELING SELECTION BIAS

The goals and methods of the proposed review can be precisely stated by using formal notation. Let θ represent the parameter of interest for an individual study, the true impact of the intervention on the treated population. The goal of this systematic review is to examine the bias

associated with different nonexperimental estimates of θ and to see how the bias varies under different conditions.

Bias can never be directly observed because a true impact is not known, but this review includes two types of studies that allow us to estimate the bias empirically. The first type of study presents up to K nonexperimental estimators, $\hat{\theta}_k$, where k=1,…K, of the parameter of interest and one experimental estimate $\hat{\theta}_0$ such that $E[\hat{\theta}_0]=\theta$. The second type of study compares average outcomes for a control group $\bar{Y}_0$ with the (adjusted) average outcomes for some comparison group based on nonexperimental method k, $\bar{Y}_k$, often a matched comparison group or regression-adjusted mean outcome for some convenience sample of untreated subjects. The relationship among these variables is shown in equations (1) and (2), where $\bar{Y}_T$ is used to represent the average outcome for the treated group and $B(\hat{\theta}_k)$ is the bias.

$$\hat{\theta}_k = \bar{Y}_T - \bar{Y}_k \tag{1}$$

$$\hat{\theta}_0 = \bar{Y}_T - \bar{Y}_0 \tag{2}$$

Subtracting equation (2) from equation (1) yields two forms of the bias estimate, corresponding to the two types of reporting formats discussed above:

$$\left(\hat{\theta}_k - \hat{\theta}_0\right) = \left(\bar{Y}_0 - \bar{Y}_k\right) = \hat{B}(\hat{\theta}_k) \tag{3}$$

Thus, the two types of studies are equivalent, even though the latter type does not use information from the treatment group.

By using these estimates, we can estimate the bias associated with each of the k estimators, defined as $B(\hat{\theta}_k) = E[\hat{\theta}_k - \theta]$. This formula shows that the bias of the estimator requires the true parameter to be known. Instead, we estimate the bias as the difference between the

nonexperimental and the experimental estimator. If the experiment is well executed, then the estimated bias should itself be unbiased, as shown in equation (4).

$$E[\hat{B}(\hat{\theta}_k)] = E[\hat{\theta}_k] - E[\hat{\theta}_0] = E[\hat{\theta}_k - \theta] = B(\hat{\theta}_k) \tag{4}$$

B. SPECIFICATION OF MULTILEVEL MODEL

The goal of the analysis in this review will be to model $B(\hat{\theta}_k)$ as a function of characteristics and context of the study and its intervention, captured in a vector labeled *Z*, and the characteristics of the estimator itself, captured in a vector labeled *W*. These data allow us to answer the following question: how does the bias vary with the type of estimator employed, the setting, and the interaction between the setting and the type of estimator? Because there are multiple bias estimates clustered within studies, the analysis lends itself to a multilevel model, wherein *j* indexes the study and *k* indexes the estimator within each study.

$$B(\hat{\theta}_{jk}) = f(Z_j, W_k, Z_j W_k) \tag{5}$$

Heckman et al. (1998) use a rigorous definition of bias that allows for heterogeneous treatment effects. In their formulation, the bias is a function of individual characteristics *X*. Their specification is helpful for understanding what is known as "the common support problem" (Lechner 2000) endemic to nonexperimental studies. The common support problem is the difficulty of making comparisons across populations whose distributions of observed background characteristics do not fully overlap. Matching methods typically have to discard or ignore treatment group members outside the region of common overlap, leading to what Heckman et al. (1998) observed was a large component of bias. Moreover, regression adjustment methods typically use linearity assumptions to extrapolate the effect of large individual background differences on outcome differences. Thus, an ideal specification of equation (5) would have

three levels: one each for studies, estimates, and individual study subjects to which the estimates pertain.

Here, we propose to examine only two levels: study and estimate. A practical problem with adding the individual dimension to the proposed review is that the source studies do not report sufficient information, such as treatment effects separately by regions of *X* (with and without common support) or the values of *X* on some common set of measures across studies. We do, however, code and include some aggregate measures of the study population as part of the *Z* vector and some measures of the average background characteristics used to form each comparison group. The specification of (5) as a multilevel model might take the following form:

$$B(\hat{\theta}_{jk}) = \alpha_{0j} + \alpha_1 W_{jk} + \alpha_{2j} W_{jk} Z_j + \varepsilon_{jk} \tag{6}$$

$$\alpha_{0j} = \beta_0 + \beta_1 Z_j + \omega_j \tag{7}$$

where α and β are parameter vectors to be estimated and ε and ω are random disturbance terms, uncorrelated with each other. Equation (6) is the estimator-level equation. Equation (7) models the effects of each estimator on bias as a function of the characteristics of the study from which the estimator was derived. The elements of α will tell us what statistical methods are associated with lower bias. Elements of β will tell us about the conditions under which the bias in the nonexperimental estimator is higher or lower.

C. HYPOTHESES

We apply the framework to the body of evaluation research in the areas of education, training, and welfare programs. Several researchers in these policy areas (Bloom 2000; Dehejia and Wahba 1999; Heckman, et al. 1998, among others) have suggested circumstances under which nonexperimental methods tend to work better (in the sense of producing estimates that are close to the experimental benchmark). These suggestions provide a starting point for what to

include in the *W* and *Z* vectors in empirical tests of different methods. The following list describes some major hypotheses that the literature has suggested:

1. Using longitudinal data for several years of preprogram employment and earnings (which include the well-known "preprogram dip") can improve the accuracy of nonexperimental impact estimators (Bloom 2000; Ashenfelter and Card 1985; Dehejia and Wahba 1999).

2. Collecting data for the treatment and comparison groups in the same manner (i.e., same questionnaire, timing, and so forth) can help avoid one source of bias (Heckman et al. 1997). Differences in measurement instrument represent an important disadvantage of using nationally representative datasets to construct nonexperimental comparison groups.

3. Controlling properly for the observables may remove most (but not all) of the selection bias in nonexperimental estimators (Heckman et al. 1998).

4. Comparing groups of individuals from the same or similar labor markets may improve the accuracy of nonexperimental impact estimators (Friedlander and Robins 1995; Bell et al. 1995; Heckman et al. 1998).

5. Using nonexperimental estimators for mandatory programs and those that use objective eligibility criteria, is more likely to be accurate because these programs are likely to have less selection bias (Bloom et al. 2002).

To test these hypotheses, it is important to describe specific nonexperimental estimators. While work by Heckman and Hotz (1989) and Heckman et al. (1998) is useful for categorizing methods in a general way, we prefer to avoid forcing the methods into mutually exclusive categories because many estimators we identified used multiple approaches. Instead we describe each estimator by a vector of characteristics that pertain to the source of the comparison group and the analytic techniques used to adjust for differences between the comparison group and the treatment population.

Figure II.1 shows the coding scheme used to classify nonexperimental estimators. We noted the techniques used to adjust for differences between the treatment and comparison groups, and

FIGURE II.1

ILLUSTRATION OF THE CODING SCHEME USED TO CLASSIFY NONEXPERIMENTAL ESTIMATORS

1. What was the source of the comparison group?
    a. national dataset
    b. control group from another site
    c. other [additional codes divide "other" into eligible nonapplicants, etc.]

2. Was the comparison group drawn from the same labor market, school district, or relevant geographic area?
    a. same specific area (e.g. same labor market or school)
    b. same general area (e.g. same state or school district)
    c. not matched geographically

3. Were covariates used in estimating program impacts (i.e. were they regression-adjusted)?
    [followup questions code the richness of the regressors]

4. Was matching used in estimating program impacts?
    [followup questions code the matching technique]

5. Were pre-intervention measures of the outcome used in estimating impacts?
    [followup questions distinguish between simple difference-in-differences, fixed effects, or other technique in this class of estimators]

6. Was some technique used to adjust for unobservable differences between the treatment and comparison groups?
    [followup questions determine the econometric technique and the nature of exclusion restrictions or instrumental variables]

7. Was this estimator subjected to a specification test that could have identified it as valid, a priori?
    a. Yes, failed test
    b. Yes, passed test
    c. No specification test conducted

also the source of the comparison group, whether it was drawn from the same location (labor market or school district) as the treated population and whether it was measured in the same way as the treated population. For the estimation method, we coded the richness of the regressors, if any, used to adjust for observable background differences in the outcome equation as well as the richness of the background characteristics, if any, used for matching. For those estimators that used matching, we coded various aspects of the matching procedure. We also coded whether and how sample members were trimmed (dropped from the analysis) if they appeared to lie outside the region of common support.

D. ESTIMATION

Several issues arise in the estimation of models such as those in equations (6) and (7). It is important to: express the dependent variable in units that are meaningful and comparable across studies; account for sampling error in the variables and the parameter estimates in the model; identify appropriate units of analysis; account for nonindependence among estimator-level observations; and ensure sufficient variation in the explanatory variables to identify the relevant parameters.

First, the dependent variable in our analysis, which is the estimate of bias, is not expressed in the same metric for all the studies. The impact or bias estimates are often reported in natural units, such as dollars per year for earnings outcomes. Because nearly all of the studies use earnings as an outcome, we will convert all effects into two types of units: (1) dollars per year adjusted for inflation as relevant and (2) standardized effect sizes, formally expressed as the number of natural units of the outcome divided by the standard deviation in the same units of the outcome for the treated population. Analyses using the first type of units will be restricted to studies with earnings outcomes. Analyses using the standardized effect sizes will make use of the full set of studies where the effect size can be calculated.

A second estimation issue pertains to accounting for estimation and sampling error in the variables and parameter estimates of the model. The dependent variable is a value that is known to be estimated with error because it is an *estimate* of the bias. This estimate of the bias has its own sampling variance determined by the variance of the experimental estimator, the variance of the nonexperimental estimator, and the covariance between the two. Model (6) can be further specified in a way that allows us to estimate a variance component that captures the variability associated with each type of estimator. One way of thus specifying the model is to estimate a new parameter α_{1k} as a random coefficient on the research design variable W_{jk}, with a mean and variance, both of which have useful interpretations. If the mean of α_{1k} differs from zero, it implies systematic bias associated with that nonexperimental approach. If the variance is high relative to those of the other approaches, it implies that the nonexperimental approach is less efficient than the others. For this interim report, we analyze the absolute value of the bias in order to focus on how different nonexperimental statistical methods and comparison group strategies relate to the overall magnitude of bias. Thus we can interpret α_1 as the effect of technique W_k on bias. A negative value means the technique reduces bias. A future version of this report will include the more general analysis.

A third issue pertains to the units of analysis and accounting for nonindependence among units. The multilevel model allows us to account for the clustering of multiple bias estimates (level 1) within studies (level 2). Estimates derived from the same study will be allowed to share a study-specific variance component or fixed effect. Another source of nonindependence derives from estimators within the same study that use essentially the same data and very similar techniques. For example, 10 independent flavors of propensity score–matching estimators do not

represent 10 different replications of propensity score–matched designs. To deal with this problem in the preliminary analysis, we first aggregate within studies (see Chapter IV).

Finally, a particularly difficult issue pertains to whether the parameters of interest to the study can even be identified by using the sample of estimates currently available in the literature. In fact, we suspect model (6) will be under-identified because of the small number of nonexperimental estimates relative to the number of known sources of variation described above. In other words, more than one explanation will plausibly fit the data. Nevertheless, even a meta-analysis where the parameters are under-identified will be useful for at least two reasons: (1) it will provide a formal framework for discussion of the different factors that may explain the variation in bias associated with different nonexperimental estimators; and (2) it will set up the problem so that the review can be easily updated when future studies are incorporated.

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

III. DATA AND REVIEW METHODS

The data for this study consist of information extracted from 18 primary studies. To generate the data set, we first developed criteria to identify studies appropriate for inclusion in the review. We then designed a search strategy for finding studies that met the criteria and assembled the set of studies. Finally, we used a coding protocol form to systematically extract information specific to each study and to each bias estimate within each study. The resulting database is described at the end of this chapter.

A. STUDY INCLUSION CRITERIA

To be included in the review, each study had to meet the following criteria:

- ***A randomized control group was used to evaluate a program, and a comparison group was available for computing at least one nonexperimental estimate of the same impact.*** Given that some studies estimated the bias directly by comparing comparison and control groups, the criterion did not require the presence of a treatment group.
- ***The experimental and nonexperimental estimates had to pertain to the same intervention in the same site(s).*** This criterion excluded a World Bank-sponsored study of education programs that were part of the Bolivian Social Investment Fund. That study (Newman et al. 2002) compared findings from an experimental design in one region of Bolivia with findings from a nonexperimental design in a different region. Such a study can potentially confound regional differences with differences in study design.
- ***The comparison was based on estimates from the same study.*** This criterion excludes the between-study literature cited in Chapter I.
- ***The intervention had to affect one of the policy areas of education, training, employment services, and welfare-to-work programs.*** This criterion is somewhat arbitrary but strikes a balance between having a sufficiently broad scope while maintaining some similarity in the likely processes that govern program participation. An important area excluded by this criterion was health-related interventions (for example, MacKay et al. 1995 and 1998). We imposed the criterion because health interventions are qualitatively different from other social interventions. Models of program participation, the key factor in sample selection bias, might be similar among education-, training-, and employment-related interventions but are likely to differ markedly for a medical or community health intervention. Furthermore, the outcomes would typically be very different.

- ***The bias estimates must pertain to the impact of a social intervention.*** This criterion excluded one study that used a nonexperimental technique to adjust for study attrition (Grasdal 2002). The role of nonexperimental methods in adjusting for missing data or nonresponse is important, but outside the scope of our study.

B. SEARCH STRATEGY

Identifying true design replication studies is challenging. There is no common language for describing what we call design replication studies. We were looking among a large volume of research for a small number of studies that met the above criteria; furthermore, a study's title or abstract does not necessarily indicate whether the study satisfies our criteria for review. Many of the studies we were aware of would not have turned up in a keyword search.

We started out with a list of known studies that included both published results and in-progress evaluations. We also used these studies' reference lists. We then conducted searches of electronic databases that index statistical or policy-related publications, publication lists of evaluation firms and government agencies, and working paper lists of selected economics and public policy institutions. To capture research in progress, we searched the recent programs of major social science and public policy research conferences and made numerous queries to researchers working in the field.

In total, we assembled dozens of candidate studies and narrowed them down to 33 for closer examination. Of the 33, only 18 met the search criteria and were thus included in the database of studies as of this interim report. Two of the 18 are included in the database even though they are still in progress and have not yet yielded results.[3]

[3] One of these studies included the results as an appendix to a government report on the effectiveness of a job training program, and the author is drafting a manuscript that highlights the experimental-nonexperimental comparison. In the other study, the authors have shared with us a prospectus that proposes in some detail to conduct the type of analysis that would meet our search criteria.

Of the 15 papers not included in the review, many simply discussed the tradeoffs between experimental and nonexperimental methods but did not provide both types of estimates. One such paper (Abadie and Imbens 2001) contained simulations, no empirical evidence. Other excluded studies (Reynolds and Temple 1995; Pradhan et al. 1998; Newman et al. 2000) involved both types of estimates but reported between-study comparisons by using results from different studies or, in some cases, different study sites. We did not impose study quality criteria or require the study to have undergone peer review to be included in our database.

C. PROTOCOL AND CODING

Two coders read the 16 completed studies. One coder took responsibility for physically coding each study, but discussed all questionable cases with the other coder. Both coders coded two studies to ensure a common understanding of the coding instrument and process. The coders occasionally asked a research assistant to search for a particular item in a paper to resolve a question about whether something they were unable to find was actually present. We also contacted authors of the source studies to obtain clarification and, sometimes, additional information.[4]

D. DESCRIPTIVE STATISTICS

The set of studies available for analysis at this stage includes 16 reports, but they are not necessarily independent replications. Four studies use data from the same experiment, the National Supported Work (NSW) demonstration. Two of the NSW studies (Dehejia and Wahba 1999; Smith and Todd 2002) are explicit re-analyses of a third study (Lalonde 1986) that used

[4] These procedures, including the coding form, are described in a formal protocol submitted to the Campbell Collaboration for review. The Campbell Collaboration will make this protocol available on its website in the future.

the same data set, although they each vary in the nonexperimental methods and sample definitions used. The fourth (Fraker and Maynard 1987) included a different treatment population and examined different matching methods than the other three. While the authors do not reach the same conclusion, the findings can be combined into a single, large virtual study for the sake of conducting meta-analysis.[5]

Two other studies in our database (Friedlander and Robins 1995; Hotz et al. 1999) also use data from a common experiment, the Work INcentives (WIN) demonstration conducted in four states. These two studies are not quite overlapping, as they focus on different outcomes and vary on the sample definitions and nonexperimental methods used.[6] The remaining 10 studies were each carried out using data from a unique social experiment. Table III.1 contains descriptive characteristics of the studies and associated interventions.[7]

Most of the interventions under study (eight) provided employment services, often as welfare-to-work demonstrations. Three were educational interventions, and one was a job

[5] A fifth study (Heckman and Hotz 1989) also used NSW data to address the same research question as we do here, but its approach was to specify and conduct specification tests to narrow down the choices among nonexperimental estimators rather than to test new ones. We plan to include results from their specification tests as descriptive variables pertaining to Lalonde's estimates, but we did not explicitly include their study in our database.

[6] In particular, Friedlander and Robins (1995) calculate within-state and cross-state estimates, whereas Hotz et al. (1999) focus only on comparisons across states. In addition, Hotz et al. (1999) use "Difference in Difference" as a nonexperimental method, whereas Friedlander and Robins do not use this method.

[7] Most studies included in this review are very recent. In fact, half of the available studies were published or released in the last two and a half years. Only two of the studies reviewed here were published or released between 1988 and 1997.

training program.[8] Half of the interventions took place in the 1990s, and only one (NSW) occurred before 1980.

In terms of their geographic location, two of the interventions were single-site programs (Arizona State University and Bergen, Norway); four were multisite interventions in a single state (California, Tennessee, Florida, and Indiana); and the remaining six were multistate interventions in the United States. Participation was mandatory for about half of the interventions of interest, and voluntary for the other half.

As with any systematic review, issues related to the comparability of studies arise. First, given that outcomes were not all expressed in the same units, we would ideally like to convert all estimates of bias into a common metric. Since post-program earnings were the key outcome measure in most studies, we converted this measure into constant (inflation-adjusted) dollars of a common base year (1996). For this interim report, we restrict most statistical analyses to the 11 studies that used post-program earnings as a key outcome. For the remaining 5 studies, we were able to convert the bias estimates of 4 of them into "effect size" units. An effect size is obtained by expressing the impact as a percentage of a standard deviation of the outcome in the target population. For many studies in this review, we did not have sufficient information to estimate the effect size.[9]

---

[8] These figures count interventions, not studies, and do not include the two studies in progress.

[9] Some studies reported bias as a percentage of the experimental impact. While such a measure can be computed for nearly every study and would appear to be an attractive unit-free measure, it is nonetheless somewhat arbitrary in that we hope to draw methodological lessons across studies where a program was found not to be effective (zero impact). In those cases, for example, Olsen and Decker (2001), the point estimate of the impact is very small, so that any bias, even a small one, appears dramatically (and misleadingly) large. A preferable alternative would be to express the bias as a percentage of the impact deemed just large enough to change a policy decision. There is no uniform or objective way, however, to extract such a benchmark from the studies reviewed.

TABLE III.1

DESCRIPTIVE STATISTICS FOR STUDIES USED IN THIS REVIEW

Characteristic	Number of Studies
Study Characteristics	
Publication source	
Peer reviewed book, chapter, or article	9
Contractor report to government or foundation	2
Working paper	5
Year of publication/release	
1986	1
1987	1
1988-1994	0
1995	2
1996-1997	0
1998	2
1999	2
2000	1
2001	3
2002	4
Intervention Characteristics (Number of Interventions in Parentheses)	
Type	
Employment services (e.g., welfare to work)	12 (8)
Employment training	1 (1)
Education related	3 (3)
Timing	
1970s	4 (1)
1980s	6 (5)
1990s	6 (6)
Location	
Single site	2 (2)
Multi-site, single state	4 (4)
Multi-state	10 (6)
Program participation rules	
Mandatory participation	7 (6)
Voluntary participation, strong incentives	1 (1)
Voluntary participation, weak incentives	8 (5)

TABLE III.1 *(continued)*

Characteristic	Number of Studies
Comparability Issues	
Main outcome	
Earnings[a]	11
Employment	1
Receipt of public assistance	1
Achievement (student test scores)	2
Persistence (school dropout)	1
Multiple time periods	7
Multiple subgroups	3
Multiple sites	5
Multiple outcomes	5
Multiple comparison groups	9

[a]Some of these studies also examined employment, but of those, all measured earnings as well.

A second issue related to comparability is that the studies varied considerably in terms of number of time periods, subgroups, sites, outcomes, and comparison groups. For example, half of the studies reported estimates separately by time period, whereas one study (Bell et al. 1995) used seven time periods. Seven studies used one comparison group, whereas two studies (Lalonde 1996; Smith and Todd 2002) used two comparison groups by applying several sample definition criteria, which resulted in eight or more bias estimates for each proposed nonexperimental method. Not surprisingly, the large variation in these dimensions led to a different number of bias estimates we used from each study. In fact, the number of estimates used from a single study ranged from 4 (Aiken et al. 1998) to 498 (Bloom et al. 2002). The analyses presented in the next chapter try to account for the large number of nonindependent estimates while still exploiting the information they contain.

E. STUDY QUALITY

An important feature of the research we review here is the quality of the evidence presented in each study. Three dimensions of study quality are most important for our purposes: (1) the quality of the randomized trial used as a benchmark; (2) the quality of the nonexperimental estimators; and (3) completeness of the information reported in the source studies.

1. Quality of the Experiments

We have assumed that the experimental estimators presented in the studies under review are themselves unbiased.[10] Common threats to the validity of the experimental estimator include differential attrition or nonresponse, randomization bias, spillover effects, substitution bias, John Henry effects, and Hawthorne effects.[11] Experimental results would also be biased if data are not collected uniformly for the treatment and control groups or if random assignment is not carried out and monitored very carefully. Large amounts of noncompliance with treatment assignment, even if monitored and well documented, can threaten the experiment's ability to answer interesting policy questions.

[10] It is less important for our purposes that the experimental estimator be externally valid or that it represent one policy parameter in particular (such as the effect of the treatment on the treated or the local average treatment effect), as long as the nonexperimental estimator that we are trying to assess purports to measure the same thing.

[11] Randomization bias results when the treatment group's experience is influenced by the presence of a randomized evaluation. Spillover effects result when the control group's experience is influenced by the presence of a treatment group. Substitution bias results when control group members are given an alternative treatment they would not have received absent the experiment. John Henry and Hawthorne effects result from members of the control and treatment group, respectively, behaving differently because they are aware they are part of an experiment.

The 12 experiments used in the design replications we reviewed were generally of high quality. Most were well-funded and were carried out by large evaluation research firms with established track records in random assignment and data collection. In four of the experiments, the Manpower Demonstration Research Corporation (MDRC) oversaw random assignment. Abt Associates oversaw random assignment in three of them, and Mathematica Policy Research (MPR) conducted two. The remaining three experiments were overseen by university-based researchers. Many of the details of the experimental designs and their implementation were not reported in the replication studies, but must be extracted from source documents from the evaluations themselves. This data extraction is not complete, but spot checks indicate that most of the experiments had relatively low crossover and attrition rates and the amount of attrition and nonresponse was typically unrelated to treatment status. For the final report, we intend to include a more thorough description of the experimental designs and will conduct sensitivity analysis to see if excluding studies with suspected threats to validity will change our overall findings.

2. Relevance of the Nonexperimental Approaches

A "high quality" nonexperimental approach in a design replication study is one that is realistic in the sense that it might have been tried in the absence of random assignment. This is an easy criterion to satisfy, since researchers have employed diverse approaches to quasi-experimentation. In fact, some nonexperimental estimators may be useful to test even if they were not justified and implemented to the highest standards, as long as we carefully describe what was done. In other words, we are interested not only in best research practices, but prevailing practice.

Some estimators, however, are less plausible than others and therefore less useful. Many of the studies under review (Friedlander and Robins 1995; Hotz et al 1999; Hotz et al. 2000; Lee

2002; Wilde and Hollister 2002; Bloom et al. 2002) tested estimators based on non-randomized comparison groups that are in fact randomized control groups for another study or another study site. The resulting replication test tells us whether the control group in one site can be made to look like the control group in the site where policy interest is focused. These comparisons are only useful to the extent that an evaluator has access to a group like the control group from a reference site to use in conducting evaluations. To isolate these cases we have coded the source of the comparison group and used this variable in our analyses.

It is also important that the estimators tested in the design replication studies we identified are representative of methods used in prevailing practice. Because of the nature of design replication studies, some estimators are more likely to be empirically testable than others. We focus here primarily on nonrandomized ("nonequivalent") comparison group designs that use regression-adjustment, matching (cell matching or propensity score matching), and related techniques (difference-in-differences, fixed effects) to estimate program impacts. Some of the estimates we extracted were based on econometric selection correction methods. One was labeled as a regression discontinuity design. Other important design differences have to do with the source of the comparison group and the nature of the background data used to adjust for differences between the comparison group and the treated population. Some of the studies tried only a few nonexperimental estimators while others examined a wide range of estimators. To avoid misinterpretation, we code and label the designs and techniques used in the papers we review. Different weighting schemes will be used with a sensitivity analysis to examine whether or how the interpretation is changed by focusing on specific estimators or classes of estimators that have more *a priori* credibility.

3. Completeness of Information

One dimension of study quality is completeness of the information presented. The studies we reviewed varied in the amount of detail in which they presented their methods and findings. In particular, the authors paid uneven attention to the problem of error estimation. A concern with methods that use a single scalar variable to correct for selection bias – "index sufficient" methods – is that the index function (e.g. the propensity score) is itself estimated with error (Heckman et al. 1998). Ignoring such estimation error can make the nonexperimental impacts or the bias estimates appear more precise than they really are. Some authors (Heckman et al. 1998; Agodini and Dynarski 2000; Smith and Todd 2002; Wilde and Hollister 2002) used bootstrap methods to calculate the sampling variance of the impact or bias estimates. Others reported analytic standard errors that most likely did not account for the estimation error in computing the propensity score. Regardless of which methods were used, not all studies reported the standard error (or related statistic such as p-value) of the bias estimate, which would be a useful statistic for testing the hypothesis that the prediction error (or bias) equals zero. Instead, we treat information that is not reported or standardized across studies as missing data. In a future version of this report, we will deal with the problem through imputation and sensitivity analysis.

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

IV. INTERIM RESULTS

Individually, each of the design replication studies is potentially difficult to interpret. The authors differed in how extensively they probed issues related to the statistical and policy significance of their results. Some acknowledged that their study was essentially a case study; others made broader statements praising or condemning a nonexperimental method. Here, we use the authors' own words to examine their conclusions and then try to re-examine their findings, drawing together the multiple sources. We plan to combine the studies quantitatively in a formal meta-analysis, although the present draft reports only illustrative findings.

A. WHAT DID THE STUDIES CONCLUDE?

The 16 completed design replication studies offered a range of conclusions about the value of the nonexperimental methods they examined and whether those methods produced findings that were similar to those of randomized experiments (see Table IV.1). Five studies concluded that nonexperimental methods performed well. Three studies found evidence that some nonexperimental methods performed well while others did not. The remaining eight studies found that nonexperimental methods did not perform well (or found insufficient evidence that they did perform well).

Four of the five studies that found positive results (evidence of small bias) qualified their conclusions, indicating that the researcher needs detailed background data (particularly prior earnings), overlap in background characteristics, or intake workers' subjective ratings of the applicants they screened. Four of those interventions where smaller biases were reported were welfare-to-work demonstrations and one was a college remedial writing program with very small samples.

TABLE IV.1

AUTHORS' CONCLUSIONS FROM THE STUDIES REVIEWED

Author(s) and Year of Publication	Type of Intervention[a]	Methods Examined[b]	Comparison Sample(s) Used[c]	Study Conclusion (Verbatim)
Nonexperimental Estimators Performed Well				
Aiken et al. 1998	College remedial writing	"Regression discontinuity" (OLS)	Ineligible nonparticipants	"The regression discontinuity design yielded a *pattern* of effect sizes that was similar to that from [the experimental design]."
Bell et al. 1995	WTW	OLS, Instrumental variables	Withdrawals, screenouts, No-shows	"We believe that the evidence presented here is generally encouraging with regard to the use of applicant-based impact methods when experiments cannot be implemented…The screenout-based approach proved much less reliable than the no-show-based model during the in-program period, but yielded similar results in the postprogram period...Screenouts may well provide the comparison group for future nonexperimental evaluations...The addition of [intake workers' subjective] ratings consistently moved the withdrawal and screenout based estimates (though not the no-show-based estimates) closer to the experimental norm."
Dehejia and Wahba 1999	Supported work	PSM, DD	CPS, PSID	"When the treatment and comparison groups overlap, and when the variables determining assignment to treatment are observed, propensity score methods provide a means to estimate the treatment impact."
Hotz, Imbens, and Klerman 2000	WTW	OLS, DD	Control groups from other sites	"The results presented here are encouraging for the ability of non-experimental methods to reproduce the results of experimental methods, if enough detailed information on individual characteristics (e.g. histories of employment, earnings, and welfare receipt) is available."
Hotz, Imbens, and Mortimer, 1999	WTW	OLS, PSM	Control groups from other sites	"We are able to predict the average outcomes for non-trainees fairly accurately, thus eliminating selection bias. Important in achieving this result is the inclusion of pre-training earnings, some personal characteristics, and some measures of aggregate differences across locations…Using control groups from other experimental evaluations appears to lead to more suitable comparison groups in our analyses, even though the experiments are conducted in very different locations and for different training programs."
Mixed Results				
Bratberg et al. 2002	Occupa-tional therapy	OLS, PSM, DD, SC	Eligible nonparticipants	"In our case study we find that nonexperimental evaluation based on sample selection estimators with selection terms that fail to meet conventional levels of statistical significance is highly unreliable. The difference in difference estimator and propensity score matching estimators perform better in our context."
Heckman et al. 1998	Job training	OLS, PSM, DD, SC	Eligible nonpartici-pants; national dataset (SIPP)	"We reject the assumptions justifying matching and our extensions of it. The evidence supports the selection bias model and the assumptions that justify a semiparametric version of the method of differences in-differences."
Olsen and Decker 2001	Job search assistance	OLS, PSM	Comparison group from related study	"The linear regression model produced accurate impact estimates. The matched comparison groups tested in this evaluation produced less accurate impact estimates than the linear regression model. This evaluation provides no evidence that the regression methods used in the WPRS evaluation are unreliable."

TABLE IV.1 *(continued)*

Author(s) and Year of Publication	Type of Intervention[a]	Methods Examined[b]	Comparison Sample(s) Used[c]	Study Conclusion (Verbatim)
Nonexperimental Estimators Performed Poorly				
Agodini and Dynarski 2001	Dropout prevention	OLS, PSM	Comparison group from related study; national dataset (NELS)	"We find no consistent evidence that propensity score methods replicate experimental impacts in our setting. This finding holds even when data available for matching are extensive. Moreover, no patterns are evident in the results to suggest the types of programs for which propensity score methods may be more likely to replicate experimental impacts."
Bloom et al. 2002	WTW	OLS, PSM, DD	Control groups from other sites	"The answer to the question, 'Do the best methods work well enough to replace random assignment?' is probably, 'No.' "
Fraker and Maynard 1987	Supported work	OLS, cell match, SC	CPS	"Nonexperimental designs cannot be relied on to estimate the effectiveness of employment programs. Impact estimates tend to be sensitive to both the comparison group construction methodology and to the analytic model used."
Friedlander and Robins 1995	WTW	Statistical matching, DD	Control groups from other sites	"Our findings illustrate that estimates of program effects from cross-state comparisons can be quite far from true effects, even when samples are drawn (as ours were) with the same sample intake procedures and from target populations defined with the same objective characteristics… Our results suggest that statistical matching or a specification test alone will be unable to reduce markedly the uncertainty surrounding that kind of nonexperimental estimate. When we switched the comparison from across states to within a state we did note some improvement, but inaccuracies still remained."
Lalonde 1986	Supported work	OLS, DD, SC	CPS, PSID	"Many of the econometric procedures do not replicate the experimentally determined results."
Lee 2001	WTW	OLS, PSM	Control groups from other sites	"We find no evidence that propensity score methods replicate experimental impacts consistently."
Smith and Todd 2002	Supported work	OLS, PSM, DD	CPS, PSID	"We find little support for recent claims that traditional, cross-sectional estimators generally provide a reliable method of evaluating social experiments. Our results show that program impact estimates generated through propensity score matching are highly sensitive to the choice of analysis sample. Among the estimators we study, the difference in differences matching estimator is the most robust."
Wilde and Hollister 2002	Class size reduction	OLS, PSM	Control groups from other sites	"The nonexperimental estimates were not very 'close' and therefore were not reliable guides as to what the 'true impact' was."

[a] WTW = welfare-to-work program

[b] Abbreviations for methods used: OLS = regression-adjusted difference in means; PSM = propensity score matching; DD = difference in differences; SC = parametric or nonparametric selection correction

[c] Abbreviations for national datasets: SIPP = Survey of Income and Program Participation; NELS = National Educational Longitudinal Study of 1988; CPS = Current Population Survey; PSID = Panel Study of Income Dynamics

It is important to probe further than the present discussion of authors' conclusions allows. The study authors used different standards to assess the size of the bias. In some cases, different authors reached different conclusions with the same data. Furthermore, the studies should not be treated equally. Some were more realistic replications of what would have been done in the absence of random assignment than were others. Within studies, some of the estimators or comparison groups were more or less likely to have been used than others, absent an experimental benchmark. Some estimates were based on smaller samples than others.

B. ANALYZING POOLED DATA

Chapter II presented a formal model relating bias to the characteristics of the nonexperimental design (consisting of a comparison group strategy and a method for adjusting for differences between the treatment and comparison groups) and the characteristics of the study. In this chapter, we report the estimates from a simplified version of that model by using a subset of the studies in our database and a pared-down list of explanatory variables. The aim is to develop an approximation of the types of analysis we hope to conduct for the final report while illustrating the basic direction in which the data point and the nature of the challenges for more systematic analysis.

1. Units of Analysis

For this interim report we have extracted 1,349 separate bias estimates from 16 studies, an average of about 84 estimates per study. The analysis we report here uses the 11 studies for which the outcome, and hence the bias, can be expressed in terms of annual earnings. The 11 studies contain 1,027 bias estimates, including estimates disaggregated by time, subgroup, study site, and so on. Simply counting each estimate separately would tend to arbitrarily over represent findings that were disaggregated. Therefore, we averaged the bias estimates within

each research design (i.e., each unique combination of design attributes) within each study.[12] For example, Olsen and Decker (2001) reported data for computing 10 separate bias estimates: 1 pertaining to a design with no matching or regression adjustment, 4 pertaining to a design with both matching and regression-adjustment, 4 pertaining to a design with matching only, and 1 with regression only. We used the average of the absolute value of the bias within each of these four combinations. The result across the 11 studies was 67 design-by-study combinations, or about 6 design types per study. These 67 observations form the analysis sample for the regressions below.

Table IV.2 shows the number of estimates we extracted from each study and the number of design types for which we computed an average bias estimate. Table IV.2 also shows the average bias overall for each study as well as the standard deviation of the bias estimates across design type.

2. Explanatory Variables

We created the design types by using seven dummy variables describing the source of the comparison group and the method used to adjust for differences between the comparison group and the treated population. The analysis below then uses these same dummy variables as explanatory variables. Three of the variables pertain to the comparison group:

1. MATCH = 1 if the comparison group is geographically matched to the treatment/control population. In the current analysis, MATCH is defined as drawn from the same state or school district. A future refinement will further distinguish between close geographic match (same labor market or school) and weak geographic match (different labor market in the same state or different school in the same district).

[12] In the future, we plan to implement weighting schemes that more explicitly account for unequal sample sizes and unequal variances in the estimates.

TABLE IV.2

DESCRIPTIVE STATISTICS OF BIAS ESTIMATES FOR STUDIES
WITH EARNINGS AS AN OUTCOME

Study	Absolute Value of Bias Annual Earnings in 1996 Dollars		Number of Estimates	Number of Types of Estimates
	Average	Standard Deviation		
Bell et al. 1995	$614	$88	63	2
Bloom et al. 2002	654	290	498	12
Bratberg et al. 2002	2,853	4,438	13	5
Dehejia and Wahba 1999	4,621	8,048	58	4
Fraker and Maynard 1987	1,027	283	48	4
Heckman et al. 1998	2,973	3,150	45	18
Hotz, Imbens, and Klerman 2000	585	218	36	2
Hotz, Imbens, and Mortimer 1999	371	160	32	4
Lalonde 1986	5,026	5,103	72	6
Olsen and Decker 2001	1,161	724	10	4
Smith and Todd 2002	4,511	4,260	152	6
Total	**$2,447**	**$3,567**	**1,027**	**67**

2. NATIONAL = 1 if the comparison group is drawn from a general-purpose national data set such as the Current Population Survey (CPS), the Panel Study of Income Dynamics (PSID), the Survey of Income and Program Participation (SIPP), or the National Education Longitudinal Study (NELS).

3. OTH_CONT = 1 if the comparison group is drawn from a control group in another site. We distinguish between this type of comparison group and naturally occurring comparison groups, such as eligible nonapplicants or program no-shows, to capture the realism of the design replication. Control groups from other sites or studies are available only if another site is conducting an experiment, but an assessment of nonexperimental methods should not necessarily assume that experimental data are available.

The other four variables pertain to the nonexperimental method:

4. REGRESSION = 1 if the impact is estimated in a regression whereby some covariates are used to adjust for observable differences. The future analysis will distinguish among those regression estimators that use more or less detailed sets of background characteristics as covariates.

5. MATCHING = 1 if the impact is estimated by using any matching methods, such as cell-matching (Heckman et al. 1998), statistical matching (Fraker and Maynard 1987; Friedlander and Robins 1995), or propensity-score matching (Dehejia and Wahba 1999).[13] A future analysis will make finer distinctions among matching methods, such as whether the cases were matched by using one-to-one or one-to-many matching and how the researcher addressed the common support problem.

6. DIFFDIFF = 1 if the impact is estimated by using pre-intervention data on the outcome variable. The variable includes difference-in-difference models and fixed-effect models as well as simple regression models that include pre-program earnings (if earnings is the outcome).

7. SELECTION = 1 if the impact is estimated by using an explicit econometric model of selection bias, such as Heckman's two-step estimator or its nonparametric counterpart (Heckman et al 1998).

The above design attributes are not mutually exclusive. The vector of seven binary attributes potentially defines $2^7 = 128$ nonexperimental designs, although the comparison group types are not completely independent of each other and many combinations are simply unlikely. Design replication studies typically investigate only four to six unique combinations, as shown in Table IV.2. Table IV.3 shows the fraction of design types in our sample that satisfies each criterion.

[13] Cell-matching compares sample members with the exact same background characteristics. Statistical matching compares members whose background characteristics are "similar" based on some distance measure. Propensity score matching is a version of statistical matching where the distance function is estimated as the probability of being in the treatment group.

TABLE IV.3

DESCRIPTIVE STATISTICS OF MAIN VARIABLES USED IN STATISTICAL ANALYSES

	Mean	Standard Deviation
Dependent Variable (Earnings, 1996 Dollars)		
Absolute value of bias	$2,447	$3,567
Main Explanatory Variables (Proportion of Estimator Types)		
Method dummies		
OLS	0.55	0.50
Matching	0.45	0.50
Selection correction	0.04	0.21
Difference in difference	0.28	0.45
Other dummies		
Geographic matching	0.37	0.49
National dataset used	0.37	0.49
Other control groups used as comparisons	0.27	0.45

C. REGRESSION RESULTS FOR EARNINGS OUTCOMES

To examine the effect of research design on bias, we estimated several regressions with the absolute value of the bias in annual earnings as the dependent variable and the design attributes and study dummies as explanatory variables (see Table IV.4). As noted above, these regression results are meant to be illustrative. Nevertheless, the results largely confirm the received wisdom. Outcomes for the various nonrandomized comparison groups available to evaluators are not good approximations to the counterfactual outcomes if left unadjusted. The intercept in the regression with no study fixed effects represents the bias associated with raw mean differences, estimated at over $4,000 in annual earnings (see row 1). The $4,000 is the expected bias if one did not make any adjustments to the "average" comparison group in our sample.

TABLE IV.4

PRELIMINARY RESULTS SHOWING THE EFFECT OF NONEXPERIMENTAL APPROACH ON BIAS IN EARNINGS IMPACTS

	Model Specification					
Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	4,431 ***		4,687 ***		6,173 ***	
	(755)		(1,231)		(1,221)	
Statistical Method						
Regression	-2,116 **	-1,976 **	-1,916 **	-1,868 **	-4,460 ***	-4,384 ***
	(835)	(818)	(780)	(811)	(1,107)	(1,145)
Matching	-1,432	-2,119 **	-1,561 **	-2,280 ***	-3,921 ***	-4,530 ***
	(829)	(854)	(774)	(849)	(1,102)	(1,159)
(Regression) x (Matching)					3,218 **	3,368 **
					(1,443)	(1,472)
Difference-in-differences	-1,158	-1,309	-1,146	-1,338	-4,307 ***	-4,163 ***
	(914)	(930)	(848)	(920)	(1,430)	(1,554)
(Regression) x (Diff-in-diffs)					3,728 **	3,271 *
					(1,626)	(1,672)
(Matching) x (Diff-in-diffs)					2,294	2,100
					(1,599)	(1,777)
Selection correction	3,445	1,880	3,099	2,183	3,976 **	2,973
	(2,057)	(2,072)	(1,931)	(2,055)	(1,846)	(1,974)
Comparison Group Strategy						
Geographic match			-815	-1,158	-704	-966
			(1,101)	(1,265)	(1,024)	(1,182)
National dataset			1,300	1,856	1,623	1,741
			(1,232)	(1,803)	(1,156)	(1,680)
Control group from another site			-1,778	-1,187	-1,922 *	-995
			(1,092)	(3,339)	(1,024)	(3,111)
Study dummies included	No	Yes	No	Yes	No	Yes
Number of studies	11	11	11	11	11	11
Number of bias estimate types (cells)	67	67	67	67	67	67

Note: Dependent variable is the absolute value of the bias in annual earnings, expressed in 1996 dollars. Standard errors are in parentheses; all explanatory variables are dummy variables.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

The definition of a substantively large bias depends on the program and the policy decision at stake, but it is safe to conclude that for disadvantaged workers, even a $1,000 difference in annual earnings is important. For example, in a benefit-cost study of Job Corps (McConnell and Glazerman 2001), a steady-state impact on annual earnings of just over $1,000 was used to justify the program's expenditure levels, one of the highest per trainee (about $16,500) for any federal training program. A reduction of $1,000 in the annual earnings impact estimate would have completely changed the study's outcome and might have led to a recommendation to eliminate rather than expand the annual $1.4 billion program. For less costly programs, such as the Job Training Partnership Act (JTPA) and the various welfare to work programs captured in our data, where both the program costs and the impacts on earnings are likely much smaller than in the Job Corps example, a bias of $1,000 seems large enough to change a policy conclusion in many cases.

The entries in the next two rows suggest that using background data as either covariates or matching variables is about equally effective, with each technique reducing the bias by about $1,900 to $2,300 once we account for the studies fixed effects. While the reduction cuts the bias associated with raw mean differences roughly in half, it still leaves policymakers with a sizable margin of error.

Combining the methods is only marginally better than applying them individually. Models (5) and (6) include an interaction term whose large positive coefficient suggests that the bias reduction from these two methods is not additive. That is, regression and matching serve as near substitutes with some possible increased benefit from combining methods. In model (5), for example, the bias from raw differences in means, represented by the intercept, is $6,173. This value is reduced to $1,713 if only regression is used and to $2,252 if only matching is used. If

matching and regression are both used, then they somewhat reinforce each other, with the bias reduced to $1,010.

The coefficients on the difference-in-difference indicator show that using baseline measures of the outcome is important, as reported in the literature. For the simpler models in (1) and (2), difference-in-difference estimators reduce the bias by about $1,200 in annual earnings, a slightly smaller reduction than was achieved with the other estimators. The interaction terms of difference in differences with the regression and matching (see models (5) and (6)) indicate that these methods are also partially offsetting. More detailed analyses planned for the next phase of the research will reveal more clearly the tradeoffs between using detailed background information, such as that obtained from a survey, versus more detailed earnings history, such as that obtained from administrative records.

The one estimator that did not reduce bias at all was the selection correction estimator. We suggest caution, however, in interpreting this finding. Few estimates in our data were based on econometric methods such as the two-step estimator, and of the few that were based on econometric methods, one study (Bratberg et al. 2000) rejected the specification based on a hypothesis test but still reported the bias estimate, which was particularly large.[14]

The regression results also indicate that the comparison group strategies identified in Chapter II were sound. By itself, use of a comparison group that is matched to the same labor market or geographic area reduced bias by about $1,000. Funders of evaluation research would

[14] The study population for Bratberg et al. (2002) differs from the populations targeted in the other studies under review not only because the population comprised Norwegians, but also because the sample members were not disadvantaged workers. The larger bias estimates would apply to a larger earnings base and therefore not be as substantively important as a similarly sized bias found in a study of welfare participants in the United States. Some of this effect is measured by the study fixed effect (see the even-numbered columns in Table IV.4).

probably prefer to use large national data sets to evaluate programs because secondary analyses are far less costly than new data collection, but our findings suggest that such a strategy comes with a penalty. The estimators that used national data *increased* the bias by about $1,800 on average.

We coded another comparison group strategy that determined whether the source was a control group from another study or another site. Several of the studies under review, such as Friedlander and Robins (1995), Hotz et al. (1999 and 2000), Lee (2001), Wilde and Hollister (2002), and Bloom et al. (2002), used the technique of comparing the control group from one site to the control group from another site and labeling one as the nonrandomized comparison group. The authors then tried to apply the techniques mentioned above to equate the two groups. We mainly include this dummy variable in the regression to distinguish between those studies and the ones that test the effectiveness of nonexperimental methods by using naturally occurring comparison groups, such as eligible nonapplicants (for example, Heckman et al. 1998) or individuals who applied to the program but were screened out (for example, Bell et al. 1995). The bias was greater on average for the natural comparison groups than for the other-site control groups, as indicated by the coefficient estimates on the "Control group from another site" variable from Table IV.4 of about -$1,000. This effect is somewhat collinear with the set of study dummies such that it is difficult to determine where the true parameter lies within that range. More important, however, is that the bias found in raw mean outcome comparisons with the most *realistic* comparison groups is even higher than suggested above.

All of the estimates in Table IV.4 should be considered illustrative, not conclusive. We plan to refine and conduct additional analyses of these data, described below, which may reveal a different or more complex story.

We excluded five studies from the preceding analysis because they did not involve earnings as an outcome. Three of those five were studies of education interventions (class size, remedial writing, and dropout prevention), and two were welfare reform demonstrations (measuring employment and welfare receipt, respectively). For four of the five (all except the employment study), we were able to convert the bias estimates into "effect size" units. But even after we standardized the units in which the dependent variable is measured, we identified several reasons that made it particularly challenging to conduct statistical analyses on the sample of studies.

First, the sample size is very small. Our data set (presented in Table A.2 in Appendix A) contains 10 unique design-attribute combinations from four studies, leaving very few degrees of freedom for analysis of variance. Second, the different units in which impacts are expressed can make it difficult to assess policy relevance.

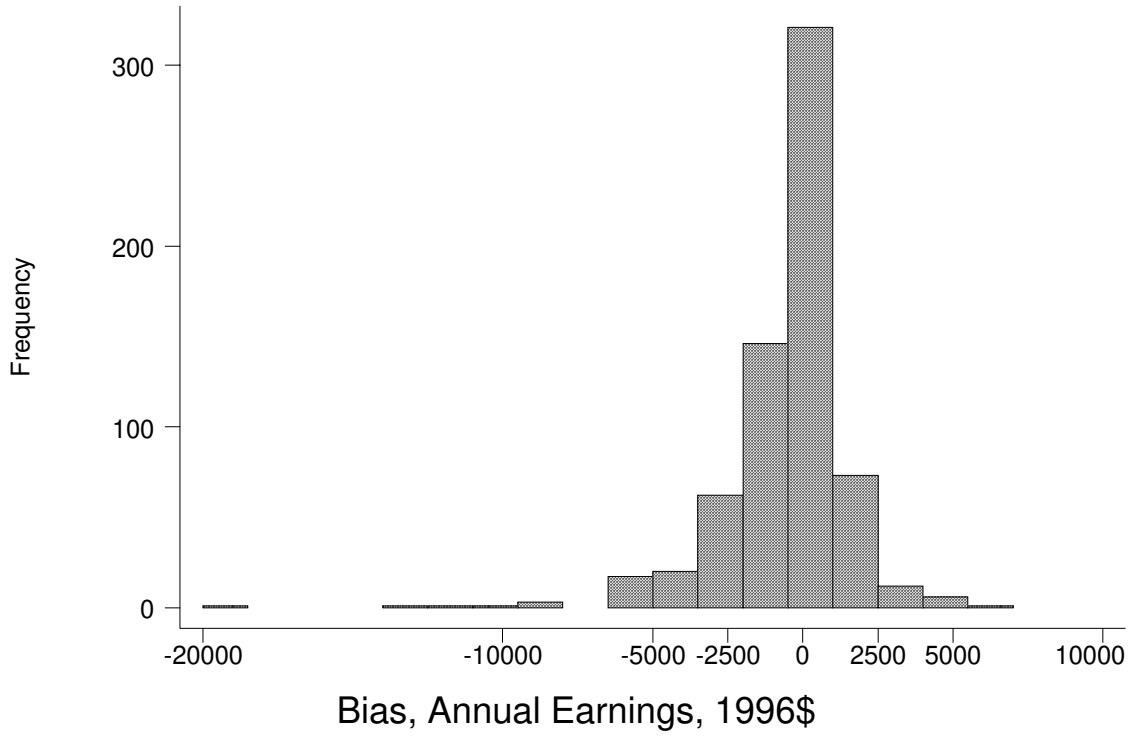
The small number of education-related studies makes statistical modeling difficult, but it also provides an opportunity to look more closely at the individual studies that were excluded from the analysis above. A detailed discussion of these studies is presented in Appendix B.

D. VARIABILITY OF THE BIAS

One hypothesis states that nonexperimental methods introduce variance, not bias, to the impact estimates. Our finding that some bias estimates were very large in absolute value is not necessarily inconsistent with this hypothesis. One way to address the issue directly is to examine the distribution of the bias estimates, paying attention to the sign, and seeing if that distribution is centered around zero. Figure IV.1 shows a histogram with the distribution of all the bias estimates extracted from the 11 studies that used earnings as an outcome and Figure IV.2 shows the same thing for four of the remaining studies whose outcome could be expressed in terms of standardized effect size. In both graphs it is apparent that the distribution is nearly centered around zero with a slight skew. This is a crude way to judge, but suggests, consistent with the

FIGURE IV.1

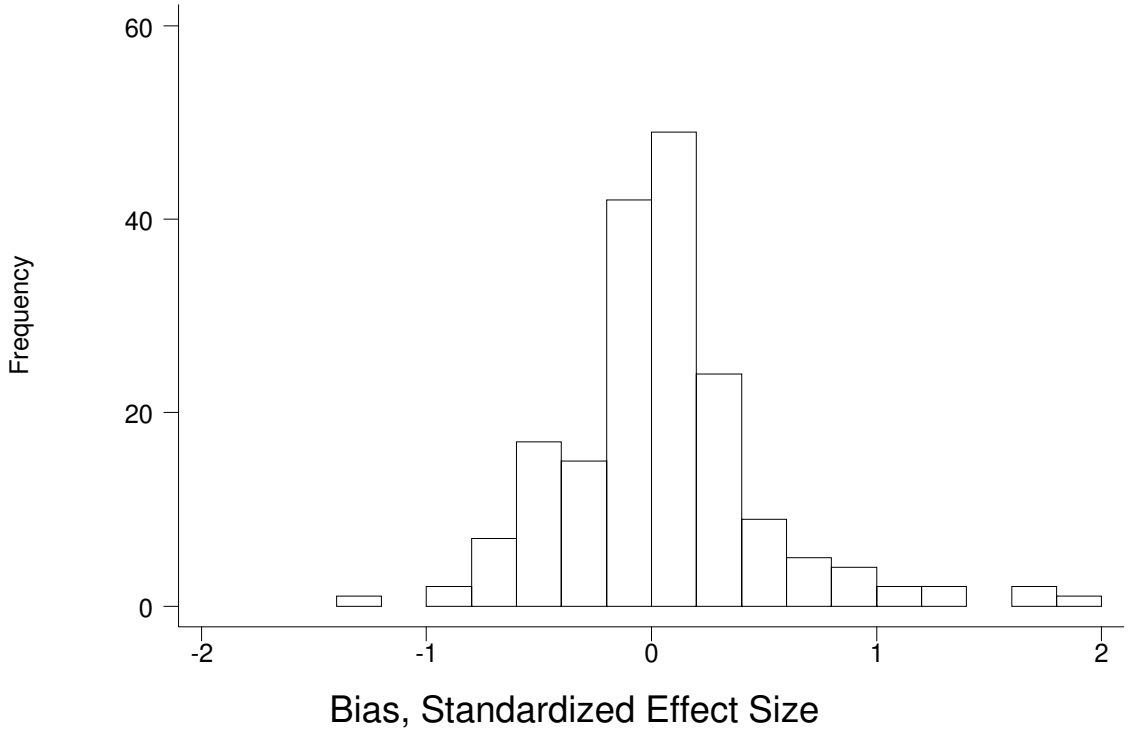
DISTRIBUTION OF BIAS ESTIMATES FOR STUDIES WITH EARNINGS AS AN OUTCOME



work of Lipsey and Wilson (1993) described in Chapter I, that if enough nonexperimental studies are done on the topic, the average effect will be close to what the experimental evidence would predict. It is clear that we need to identify dimensions along which the bias begins to cancel out. Whether or not the average bias, properly weighted within and between studies, is really close enough to zero for policy makers, or whether the bias cancels out within a narrower domain of research, is a question we plan to address in the near future and address more definitively as more design replication studies are completed.

FIGURE IV.2

DISTRIBUTION OF BIAS ESTIMATES FOR STUDIES WITH OUTCOMES OTHER THAN EARNINGS



THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

V. DISCUSSION AND NEXT STEPS

When it comes to social program evaluation, many of us acknowledge the superiority of controlled randomized trials but often substitute nonrandomized, quasi-experimental methods when proposing, conducting, or synthesizing evaluation research. There are several reasons for this. Many types of nonexperimental studies are easy to complete. For example, if random assignment is not a requirement, then retrospective data can be used, trimming several years off the turnaround time of an evaluation. Nonexperimental studies can often be completed using existing data collected by others, so even researchers with limited resources can conduct them. Even when investigators conduct experiments, policy makers often ask hard questions that the randomized trial was not designed to answer, tempting the researcher to cross over into nonexperimental analysis.[15]

These phenomena in turn lead to so much nonexperimental evidence for research synthesists that it is hard to ignore, tempting them to at least try to reconcile it with the experimental evidence. Consider the case of doing a systematic review of research on the effectiveness of some intervention of which there are 100 studies. If only 5 of the studies are experiments, do we throw out the other 95 studies? Would that be a complete review of the state of knowledge? What if three of the experiments were poorly implemented?

[15] For example, researchers are often asked to comment on dosage effects, such as "How effective is the training program for those who did not drop out early?" when in fact, we do not know who in the control group would have dropped out early.

In all these situations, even the strongest believer in randomized trials finds him or herself at least considering nonexperimental evidence. This systematic review of the literature on design replication studies says whether the best available empirical evidence can provide any guidance on how to treat nonexperimental evidence, both in weighing research evidence and in designing options for new evaluation research.

A. SUMMARY OF FINDINGS

A preliminary review of the evidence suggests that the 16 design replication case studies we identified, even taken together, will not resolve any enduring debates about the use of nonexperimental methods. We did, however, examine the distribution of empirical bias estimates and identify some factors associated with lower bias. The interim findings can be summarized in terms of the three empirical questions laid out in Chapter I.

Question: Do quasi-experimental methods replicate the findings from experimental impact evaluations?

Answer: Occasionally, but not in a way that can be easily predicted.

- A simple vote-count of the authors' conclusions gives a mixed picture. Half of the design replication studies we identified found no support for the validity of the nonexperimental methods they tried. In the other half of the studies, the authors concluded that one or more of the nonexperimental approaches was promising, but the conditions under which a randomized control group can be successfully mimicked have not yet been identified.
- An examination of the bias estimates suggests that the methods we examined – mostly nonequivalent comparison group designs with either linear regression or propensity score matching – were effective in isolated cases. However, more often they were very inaccurate in approximating the experimental result.

Question: Do we know the conditions under which nonexperimental impact estimates are likely to replicate experimental impact estimates?

Answer: We identified some factors associated with lower bias –generally, an initially similar comparison group and better data. Sophisticated methods sometimes help, but cannot make up for a lack of good data.

- Bias was lower in absolute value when the comparison group was drawn in a way that would make it more likely to be initially comparable. For example, bias was lower when the comparison group was drawn from within the evaluation itself rather than from a national dataset, when it was locally matched to the treatment population, and when it was itself drawn as a control group in an evaluation of a similar program or the same program in a different study site.

- Bias was lower when pre-intervention measures of the outcome were used to adjust for treatment/comparison groups' initial differences.

- Statistical adjustments in general reduced bias, but matching methods (most of which used propensity scores) were not necessarily more effective than simple regression in reducing bias.

- The size of the bias estimates did not appear to depend on whether the intervention was related to education, training, employment services, or welfare.

Question: Can the biases inherent in a single nonexperimental estimator be offset or cancelled out by averaging over many studies, time periods, or settings?

Answer: The data suggest that it may be possible, but we have not identified a more narrow dimension along which biases consistently cancel out.

- The estimated biases went in both the positive and negative direction, and their distribution across all the studies reviewed was centered roughly around zero.

- For any one intervention or context, however, there was not enough data to conclude that the overall bias would average out to zero.

We caution that this summary of findings gives only part of the picture. A somewhat more complete story can be developed in the short term as we update our database with the design replication studies that are now in progress and conduct additional analysis described below.

B. NEXT STEPS

This report has identified many of the challenges in synthesizing an impressive, but still limited body of design replication studies. Several concrete tasks lie before us that will allow for more opportunities to understand the evidence on how nonexperimental estimators perform. Here we list those tasks, discuss their feasibility, and where possible, foreshadow the findings we expect to report in a future draft of this report.

The first task is to incorporate more detailed codes to describe the nonexperimental estimators. Some of the proposed new variables are as follows:

- ***Specification Test Result.*** Would the estimator have been selected using a specification test? This allows us to know whether researchers have the capability of identifying promising nonexperimental estimators *a priori*, without the benefit of an experimental benchmark. A problem is that very few studies that we reviewed reported specification tests. Those that did attempt to identify more promising estimators this way (Friedlander et al. 1995; Heckman et al. 1998; Bratberg et al. 2002; Lee 2002; see also Heckman and Hotz 1989 where the approach was first demonstrated), had modest to mixed success in narrowing down the bias estimates to the best ones. Using Lalonde's data, Heckman and Hotz found that the most inaccurate nonexperimental impact estimates could have been eliminated a priori through specification tests. Friedlander and Robins showed that the estimators that did not fail the specification test performed somewhat better than those that did fail, but not enough to make them promising tools for evaluation. Heckman et al. also found a correspondence between the hypothesis testing based on pre-intervention measures and the bias estimated from post-training outcomes. Bratberg et al. found that none of their proposed estimators were rejected by a specification test, even though the econometric selection correction estimators produced wildly inaccurate impact estimates. They commented that "a prudent evaluator… would retreat to unadjusted OLS" because the selection terms were insignificant. However, that fact would be difficult to prove, since we cannot observe the counterfactual behavior of the econometrician. Regardless, evaluators should exercise caution when the specification test has insufficient power.

- ***Realism of the Estimator.*** Would the estimator have been used to evaluate the program, as opposed to one that is used for a sensitivity analysis or simply to describe the comparison group? After coding this variable we can re-do the analysis, dropping from the sample any estimators such as the raw mean difference, which are rarely suggested as a serious alternative to random assignment. The resulting analysis could make the overall findings more relevant, but requires us to make subjective judgments about whether a given estimator is realistic. Some estimates are clearly labeled as illustrative, descriptive, or part of a sensitivity analysis. For others, it is less clear. Some initial attempts to remove these unrealistic bias estimates indicate that doing so mainly affects the answer to the overall question, "How far off are nonexperimental estimates in general?" We will use multiple definitions of "realism" to enhance the interpretation of results in the final report.

- ***Rich regressors for OLS.*** How richly descriptive is the set of background variables used as covariates to adjust for differences? How important is it to have pre-intervention measures of the outcome as control variables? We will use codes that distinguish between different classes of data quality. For example, a study relying on administrative records might have basic demographic information like age, sex, and race/ethnicity. A detailed survey, however, might contain background information that is known to be more directly related to the outcomes of interest, such as education and experience (for labor market outcomes) or parents' education (for

educational outcomes). Initial analyses on incompletely coded data suggest that this variable could be very important, consistent with the hypothesis listed in Chapter III.

- ***Rich variables for matching.*** How richly descriptive is the set of background variables used in the matching procedure? Here we will use the same categories measured for the data used in OLS regression. Again, we believe this will be an important explanatory variable in the planned analysis.

- ***Type of matching method.*** What particular method was used to perform the matching? Here we will distinguish among exact cell-matching and each of the many ways to implement propensity score matching (using subclassification on the propensity score, kernel density estimation, matching with or without replacement etc.) Examining the point estimates within the several studies in our database that implemented propensity score matching in several ways, we found many cases where the particular method had little effect on the estimator's performance.

Once these variables are defined and coded consistently, it will be possible to re-examine not only the size of the bias, as we did above, but the effect of research design on the direction and variance of the bias. This leads to another task, which is to further analyze the data, focusing on alternative coding schemes for the dependent variable, as the three examples below suggest.

1. Signed Value of the Bias

The regression analysis described in this report used the absolute value of the bias, aggregated within studies up to the point of unique estimators. The next step is to re-do the analysis on the full set of over 1,000 bias estimates without taking the absolute value. This analysis would use a multi-level model like the one discussed in Chapter II to account for the non-independence of multiple bias estimates within studies. This analysis will provide quantitative estimates of the overall mean and variance of the bias for all the studies and for selected subgroups of the studies.

2. Dummy Indicator for Same Statistical Inference

One way to determine whether differences between experimental and nonexperimental estimates are due to chance or systematic bias is to construct a hypothesis test that would

compare the estimated bias against its sampling distribution and use the result of that test as a dependent variable. Another approach is to code a dummy variable indicating whether one would reach the same statistical inference (sign and significance) using either the experimental or nonexperimental impact estimates and use the dummy as an outcome variable. For those design replication studies that report only the bias and not the separate impact estimates (e.g. Hotz et al. 1999; Heckman et al 1998; Smith and Todd 2002; Bloom et al. 2002), we will define the variable to indicate whether the bias is statistically significant (different from zero).

3. Dummy Indicator for Same Policy Conclusion

Even if the bias estimates are converted to a common metric for all studies, it may be difficult to interpret the magnitudes of the bias estimates or the reductions in bias associated with different research designs. We discussed the difficulties with expressing the impacts (and therefore bias) as a percentage of the impact or a percentage of the control group mean. Another possibility is to measure the bias as a dummy variable indicating whether it is considered large enough to influence the policy decision. The dependent variable would then be a measure of whether the experimental and nonexperimental estimates led to the same policy decision or conclusion. Judging policy significance will be challenging because the policy decision is never fully known and the threshold of evidence is often subjective. However, we can use multiple coding schemes to assess policy-relevant magnitudes in different ways and use sensitivity analysis to determine whether the subjectively is a problem.

Conducting further statistical analysis will give us a better understanding of what can be learned from the current body of 16 design replication studies. By adding new studies to this list and updating the analysis, we expect to advance our ability to assess nonexperimental research methods in the evaluation of social interventions.

REFERENCES

General References

Abadie, Alberto, and Guido W. Imbens. "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects." Unpublished Manuscript. September 2001.

Ashenfelter, Orley, and David Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *The Review of Economics and Statistics*, vol. 67, no. 4, November 1985, pp. 648-660.

Bloom, Howard. "Using Non-Experimental Methods to Estimate Program Impacts: Statistical Models, Matches and Muddles." University of California at Berkeley Seminar Series "Evaluating Welfare Reform: Non-Experimental Approaches." Fall 2000. Last located at [http://ucdata.berkeley.edu/new_web/welseminar/fall2000/bloomabstract.html].

Burtless, Gary. "The Case for Randomized Field Trials in Economic and Policy Research." *Journal of Economic Perspectives,* vol. 9, no. 2, Spring 1995, pp. 63-84.

Chalmers, T.C., H. Smith Jr., B. Blackburn, B. Silverman, B. Schroeder, D. Reitman, and A. Ambroz. "A Method for Assessing Study Quality of a Randomized Experiment." *Controlled Clinical Trials,* vol. 2, 1981, pp. 31-49.

Cook, Thomas, and Donald Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Hopewell, NJ: Houghton Mifflin, 1979.

Cooper, H., Charlton, K., Valentine, J. C., and Muhlenbruck, L. "Making the Most of Summer School: A Meta-Analytic and Narrative Review." *Monographs of the Society for Research in Child Development.* Malden, MA: Blackwell, 2000.

DiPrete, Thomas A., and Henriette Engelhardt. "Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation." MPIDR Working Paper, WP 2000-013. Rostock, Germany: Max-Planck-Institute for Demographic Research, December 2000.

Glazerman, Steven M. "Assessing Study Quality in Systematic Reviews." Washington, DC: Mathematica Policy Research, Inc., June, 2002.

Grasdal, Astrid. "The Performance of Sample Selection Estimators to Control for Attrition Bias." *Health Economics,* no. 10., 2001, pp. 385-398.

Heckman, James J., and V. Joseph Hotz. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association,* vol. 84, no. 408, December 1989, pp. 862-874.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies,* vol. 64, 1997, pp. 605-654.

Heckman, James J., and Jeffrey A. Smith. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives,* vol. 9, no. 2, Spring 1995, pp. 85-110.

Lechner, Michael. "A Note on the Common Support Problem in Applied Evaluation Studies." Working Paper. St. Gallen, Switzerland: University of St. Gallen, December 2000.

McConnell, Sheena, and Steven M. Glazerman. "National Job Corps Study: The Benefits and Costs of Job Corps." Washington, DC: Mathematica Policy Research, Inc., 2001.

McKay, James R., Arthur I. Alterman, Thomas McLellan, Chris R. Boardman, Frank D. Mulvaney, and Charles P. O'Brien. "Random Versus Nonrandom Assignment in the Evaluation of Treatment for Cocaine Abusers." *Journal of Consulting and Clinical Psychology*, vol. 66, no. 4, 1998, pp. 697-701.

McKay, James R., Arthur I. Alterman, Thomas McLellan, Edward C. Snider, and Charles P. O'Brien. "Effect of Random Versus Nonrandom Assignment in a Comparison of Inpatient and Day Hospital Rehabilitation for Male Alcoholics." *Journal of Consulting and Clinical Psychology*, vol. 63, no. 1, 1995, pp. 70-78.

Michalopoulos, Charles, Howard Bloom, and Carolyn J. Hill. "Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" Paper presented at the fall research conference of the Association for Public Policy Analysis and Management, November, 2001.

Newman, John, Menno Pradhan, Laura Rawlings, Geert Ridder, Ramiro Coa, and Jose Luis Evia. "An Impact Evaluation of Education, Health and Water Supply Investments of the Bolivian Social Investment Fund." *World Bank Economic Review*. June 2001.

Pradhan, Menno, Laura Rawlings, and Geert Ridder. "The Bolivian Social Investment Fund: An Analysis of Baseline Data for Impact Evaluation." *World Bank Economic Review*, vol. 12, no. 3, June 2001, pp. 457-482

Reynolds, Arthur J., and Judy A. Temple. "Quasi-Experimental Estimates of the Effects of a Preschool Intervention: Psychometric and Econometric Comparisons." *Evaluation Review*, vol. 19, no. 4, August 1995, pp. 347-373.

Shadish, William R. "The Empirical Program of Quasi-Experimentation." In L. Bickman, ed., *Research Design: Donald Campbell's Legacy*. Thousand Oaks, CA: Sage, 2000.

Shadish, William R., and Kevin Ragsdale. "Random Versus Nonrandom Assignment in Psychotherapy Experiments: Do You Get the Same Answer?" *Journal of Consulting and Clinical Psychology*, vol. 64, 1996, pp. 1290-1305.

Wortman, Paul. "Judging Research Quality." In H. Cooper and L. Hedges, eds., *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994, pp. 97-107.

Studies in the Systematic Review

Agodini, Roberto, and Mark Dynarski. "Are Experiments the Only Option?" Princeton, NJ: Mathematica Policy Research, Inc., August 2001.

Aiken, Leona S., Stephen G. West, David E. Schwalm, James Carroll, and Shenghwa Hsuing. "Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remedial Writing Program." *Evaluation Review*, vol. 22, no. 4, April 1998, pp. 207-244.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen C. Cain. *Program Applicants as a Comparison Group in Evaluating Training Programs.* Kalamazoo, MI: Upjohn Institute for Employment Research, 1995.

Bloom, Howard, Charles Michalopoulos, Carolyn Hill, and Ying Lei. "Can Non-Experimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" New York, NY: Manpower Demonstration Research Corporation, June 2002.

Bratberg, Espen, Astrid Grasdal, and Alf Erling Risa. "Evaluating Social Policy by Experimental and Nonexperimental Methods." *Scandinavian Journal of Economics,* vol. 104, no. 1., 2002, pp. 147-171.

Dehejia, Rajeev, and Sadek Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, vol. 94, no. 448, December 1999, pp. 1053-1062.

Fraker, Thomas, and Rebecca Maynard. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, vol. 22, no. 2, Spring 1987, pp. 194-227.

Friedlander, Daniel, and Philip Robins. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review,* vol. 85, no. 4, September 1995, pp. 923-937.

Gritz, R. Mark, and Terry Johnson. "National Job Corps Study: Assessing Program Effects on Earnings for Students Achieving Key Program Milestones." Washington, DC: Battelle Memorial Institute, June 2001.

Heckman, James J., Hidehiko Ichimura, Jeffrey C. Smith, and Petra Todd. "Characterizing Selection Bias." *Econometrica,* vol. 66, no. 5, September 1998, pp. 1017-1098.

Hotz, V. Joseph, Guido W. Imbens, and Jacob Klerman. "The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program." NBER Working Paper 8007. Cambridge, MA: National Bureau of Economic Research, November, 2000.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. "Predicting the Efficacy of Future Training Programs Using Past Experiences." NBER Technical Working Paper 238. Cambridge, MA: National Bureau of Economic Research, May 1999.

Lalonde, Robert. "Evaluating the Econometric Evaluations of Training with Experimental Data." *The American Economic Review*, vol. 76 no. 4, 1986, pp. 604-620.

Lee, Wang S. "Propensity Score Matching on Commonly Available Nonexperimental Comparison Groups." Working Paper. Bethesda, MD: Abt Associates, November 2001.

Olsen, Robert, and Paul Decker, "Testing Different Methods of Estimating the Impacts of Worker Profiling and Reemployment Services Systems." Washington, DC: Mathematica Policy Research, Inc., 2001.

Smith, Jeffrey C., and Petra Todd. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, forthcoming 2002.

Wilde, Elizabeth Ty, and Robinson Hollister. "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact Against Experimental Estimates of Impact with Education Test Scores as Outcomes." Discussion Paper No. 1242-02. Madison, WI: Institute for Research on Poverty, January 2002.

APPENDIX A

TABLES

THIS PAGE IS INTENTIONALLY LEFT BLANK FOR DOUBLE-SIDED COPYING

TABLE A.1

CONTENTS OF DATASET USED IN STATISTICAL ANALYSES

Study	Number of Bias Estimates	Mean Absolute Value of Bias (Earnings, 1999 Dollars)	Statistical Method Codes				Comparison Group Strategy Codes		
			OLS	Matching	Diff in Diff	Selection Correction	Geographic Matching	National Dataset	Other Control Groups
Bell et al. 1995	21	551	0	0	0	0	1	0	0
	42	676	1	0	0	0	1	0	0
Bloom et al. 2002	40	747	0	0	0	0	0	0	1
	26	352	0	0	0	0	1	0	1
	48	779	0	1	0	0	0	0	1
	52	407	0	1	0	0	1	0	1
	40	917	1	0	0	0	0	0	1
	26	425	1	0	0	0	1	0	1
	80	1,342	1	0	1	0	0	0	1
	36	546	1	0	1	0	1	0	1
	24	736	1	1	0	0	0	0	1
	26	392	1	1	0	0	1	0	1
	48	775	1	1	1	0	0	0	1
	52	431	1	1	1	0	1	0	1

TABLE A.1 (*continued)*

Study	Number of Bias Estimates	Mean Absolute Value of Bias (Earnings, 1999 Dollars)	Statistical Method Codes				Comparison Group Strategy Codes		
			OLS	Matching	Diff in Diff	Selection Correction	Geographic Matching	National Dataset	Other Control Groups
Bratberg et al. 2002	4	1,079	0	1	0	0	1	0	0
	1	785	1	0	0	0	1	0	0
	4	10,790	1	0	0	1	1	0	0
	1	785	1	0	1	0	1	0	0
	3	828	1	1	0	0	1	0	0
Dehejia & Wahba 1999	8	16,686	0	0	0	0	0	1	0
	38	506	0	1	0	0	0	1	0
	2	995	1	0	0	0	0	1	0
	10	297	1	1	0	0	0	1	0
Fraker and Maynard 1987	6	1,272	0	0	1	0	0	1	0
	12	1,147	1	0	0	0	0	1	0
	6	621	1	0	1	0	0	1	0
	24	1,068	1	1	0	0	0	1	0

TABLE A.1 (*continued)*

Study	Number of Bias Estimates	Mean Absolute Value of Bias (Earnings, 1999 Dollars)	Statistical Method Codes				Comparison Group Strategy Codes		
			OLS	Matching	Diff in Diff	Selection Correction	Geographic Matching	National Dataset	Other Control Groups
Heckman et al. 1998	1	7,142	0	0	0	0	0	0	0
	1	4,674	0	0	0	0	1	0	0
	3	13,419	0	0	0	0	0	1	0
	1	2,995	0	1	0	0	0	0	0
	1	652	0	1	0	0	1	0	0
	3	2,565	0	1	0	0	0	1	0
	1	2,177	0	1	1	0	0	0	0
	1	929	0	1	1	0	1	0	0
	3	2,649	0	1	1	0	0	1	0
	1	860	1	0	0	0	1	0	0
	3	5,131	1	0	0	1	1	0	0
	2	631	1	0	1	0	1	0	0
	1	2,427	1	1	0	0	0	0	0
	12	1,130	1	1	0	0	1	0	0
	3	1,475	1	1	0	0	0	1	0
	1	957	1	1	1	0	0	0	0
	4	940	1	1	1	0	1	0	0
	3	2,769	1	1	1	0	0	1	0

TABLE A.1 (*continued)*

Study	Number of Bias Estimates	Mean Absolute Value of Bias (Earnings, 1999 Dollars)	Statistical Method Codes				Comparison Group Strategy Codes		
			OLS	Matching	Diff in Diff	Selection Correction	Geographic Matching	National Dataset	Other Control Groups
Hotz, Imbens, and Klerman 2000	18	430	0	0	0	0	0	0	1
	18	739	1	0	0	0	0	0	1
	8	569	0	0	0	0	0	0	1
	8	193	0	0	1	0	0	0	1
	8	415	0	1	0	0	0	0	1
	8	310	1	0	1	0	0	0	1
Lalonde 1986	8	15,275	0	0	0	0	0	1	0
	2	2,785	0	0	1	0	0	1	0
	38	3,432	0	1	0	0	0	1	0
	2	3,206	1	0	0	0	0	1	0
	12	1,361	1	0	0	1	0	1	0
	10	4,095	1	1	0	0	0	1	0
Olsen and Decker 2001	1	1,252	0	0	0	0	1	0	0
	4	2,041	0	1	0	0	1	0	0
	1	276	1	0	0	0	1	0	0
	4	1,076	1	1	0	0	1	0	0

TABLE A.1 (*continued)*

Study	Number of Bias Estimates	Mean Absolute Value of Bias (Earnings, 1999 Dollars)	Statistical Method Codes				Comparison Group Strategy Codes		
			OLS	Matching	Diff in Diff	Selection Correction	Geographic Matching	National Dataset	Other Control Groups
Smith and Todd 2002	13	13,089	0	0	0	0	0	1	0
	16	2,039	0	0	1	0	0	1	0
	79	3,684	0	1	0	0	0	1	0
	20	1,904	1	0	0	0	0	1	0
	6	3,195	1	0	1	0	0	1	0
	18	3,154	1	1	0	0	0	1	0

TABLE A.2

CONTENTS OF DATASET, STUDIES EXCLUDED FROM STATISTICAL ANALYSIS

			Statistical Method Codes				Comparison Group Strategy Codes		
Study	Number of Bias Estimates	Mean Absolute Value of Bias (Effect Size)	OLS	Matching	Diff in Diff	Selection Correction	Geographic Matching	National Dataset	Other Control Groups
Agodini and Dynarski 2001	16	0.31	0	1	0	0	1	0	0
	16	0.31	0	1	0	0	0	1	0
	16	0.35	1	0	0	0	1	0	0
	16	0.32	1	0	0	0	0	1	0
Aiken et al. 1998	4	0.14	1	0	0	0	1	0	0
Lee 2001	10	0.28	0	0	0	0	1	0	1
	50	0.24	0	1	0	0	1	0	1
	10	0.25	1	0	0	0	1	0	1
Wilde and Hollister 2002	11	0.76	0	1	0	0	0	0	1
	11	0.43	1	0	0	0	0	0	1

NOTE: Studies excluded because outcome could not be expressed in terms of annual earnings.

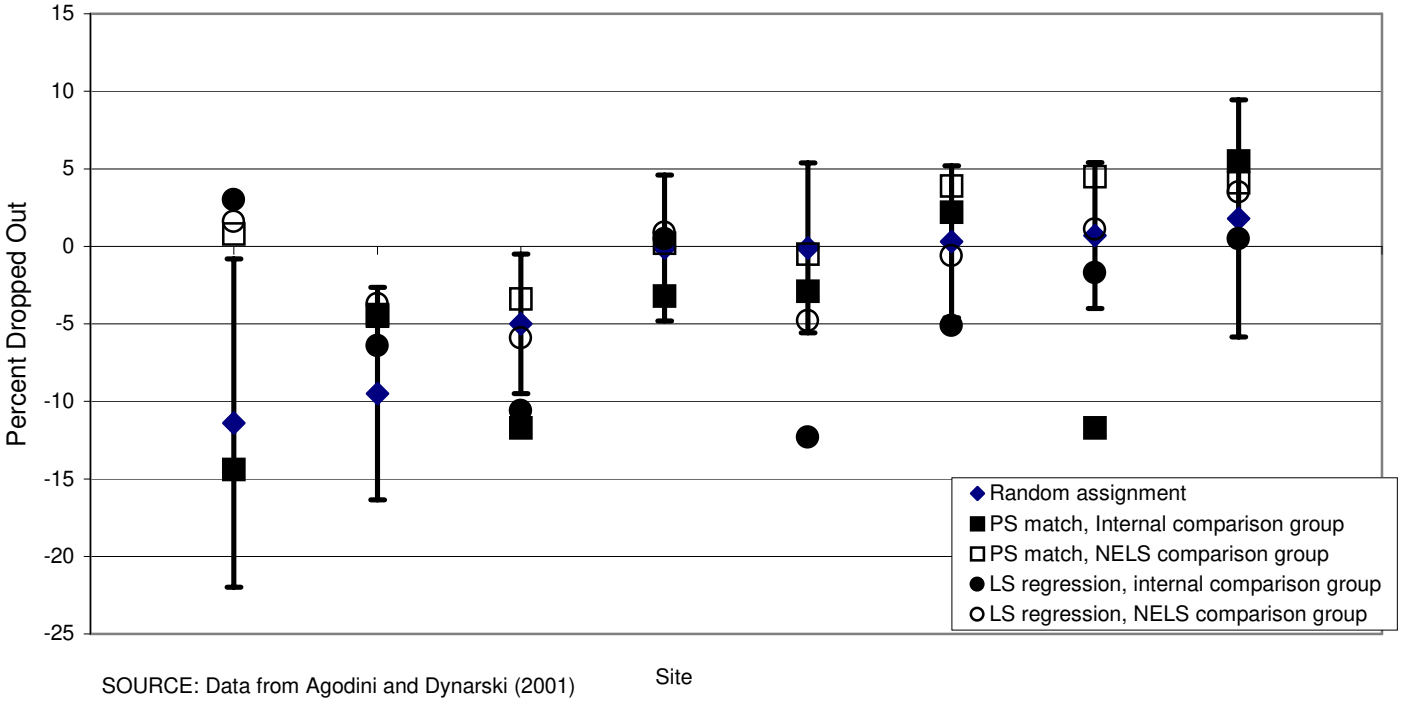
APPENDIX B

Figures B.1 through B.3 show the design replication results from the three studies that examined education interventions: school dropout prevention (Agodini and Dynarski 2001), class size reduction (Wilde and Hollister 2002), and remedial writing (Aiken et al. 1998). In each graph we present both the experimental and corresponding quasi-experimental impact estimates for the given study site or outcome. For each experimental impact estimate there are multiple quasi-experimental impact estimates, one for each method/comparison group combination, whose closeness to the benchmark can be directly compared.

Unfortunately, there is no definitive way to explain the experimental-nonexperimental differences found in Figures B.1 through B.3, much less to interpret the size of those differences. Some of the possible explanations are listed below.

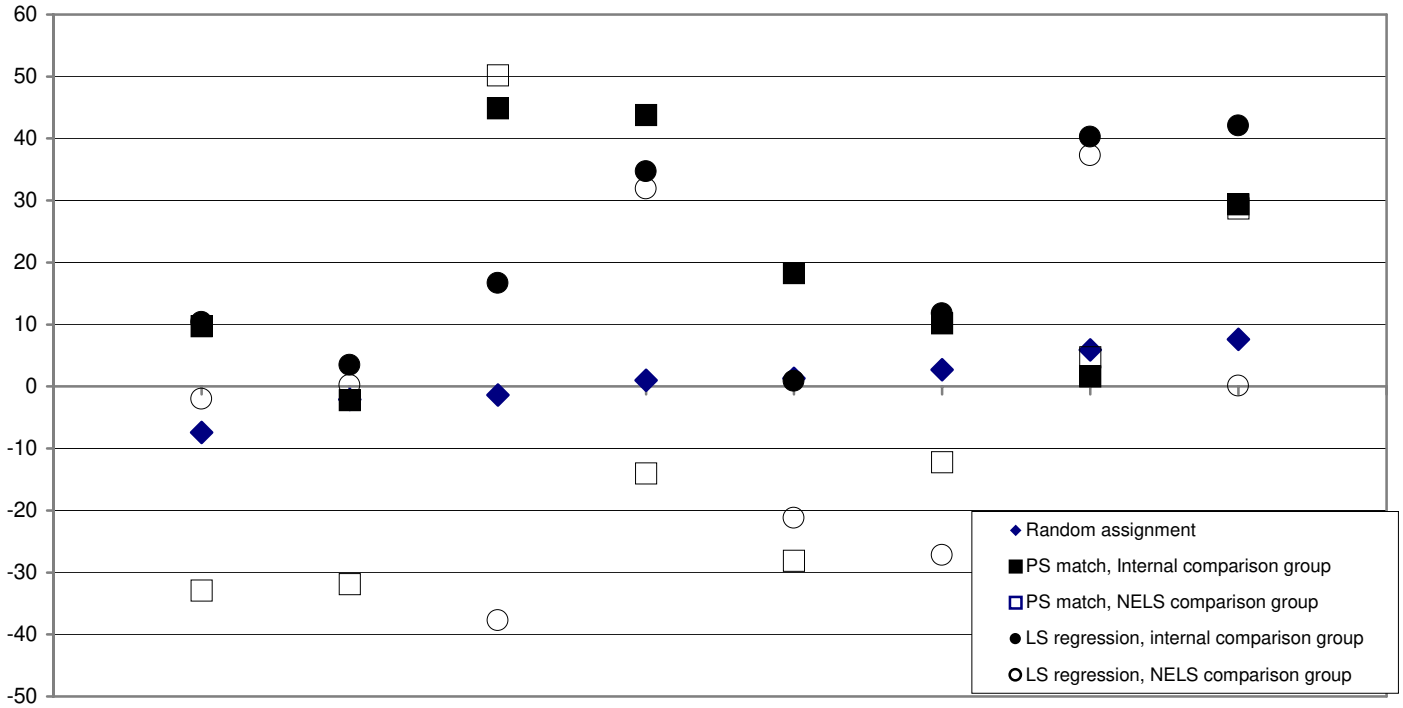
Source of the comparison group. In Figure B.1 the different shadings of the symbols (solid or hollow) refer to comparison groups that were drawn either internally to the study, using a convenience sample of students from similar school districts, or externally to the study, using data from the National Education Longitudinal Study (NELS). In Figure B.2 the nonexperimental results for any given school are based on comparison samples from other schools that also happen to be in the Tennessee class size experiment (but which were not part of random assignment for that particular school). In Figure B.3, the comparison groups are either students who were eligible for treatment but arrived too late to be recruited into the study (nonequivalent comparison) or who were ineligible because of their score on a placement test (regression discontinuity). None of the comparison group types appears to have a distinct advantage over the others, except perhaps the late arrivals and ineligibles used in the remedial writing study, where the nonexperimental impact estimates are clustered well within the confidence interval of the experimental estimates.

Figure B-1a Middle School



SOURCE: Data from Agodini and Dynarski (2001)

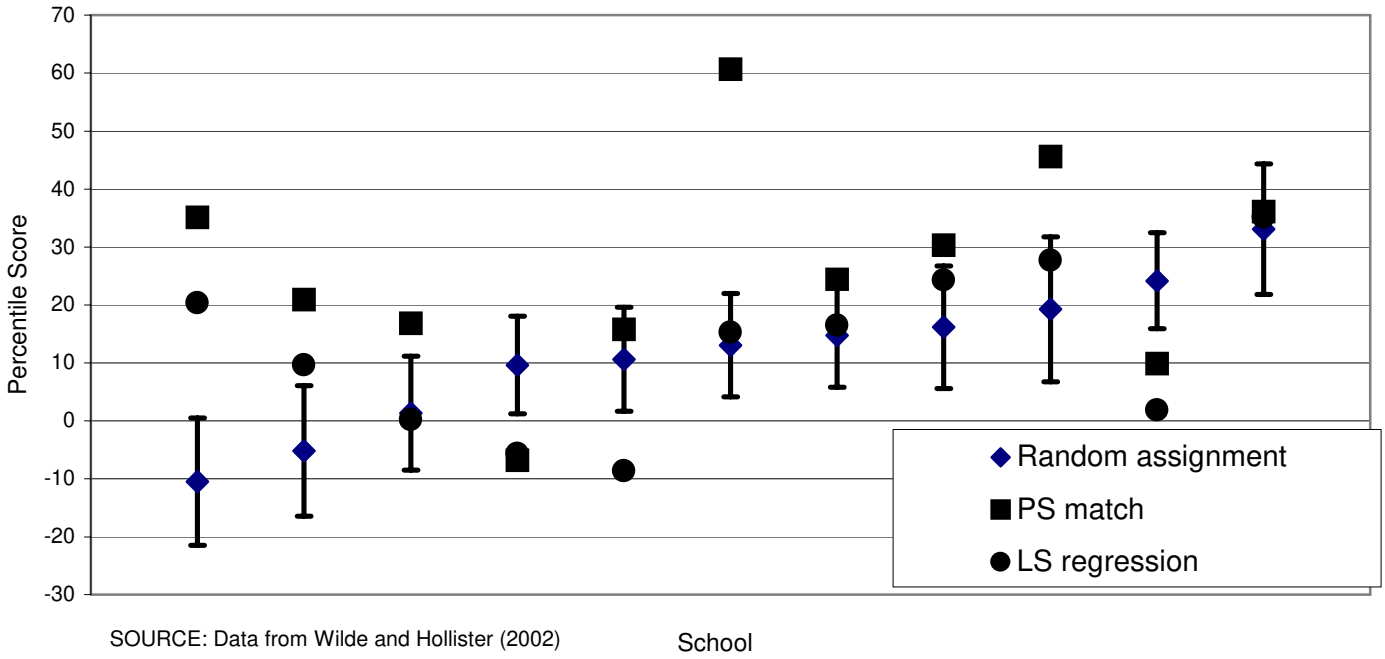
Figure B-1b High School



SOURCE: Data from Agodini and Dynarski (2001)

FIGURE B-2

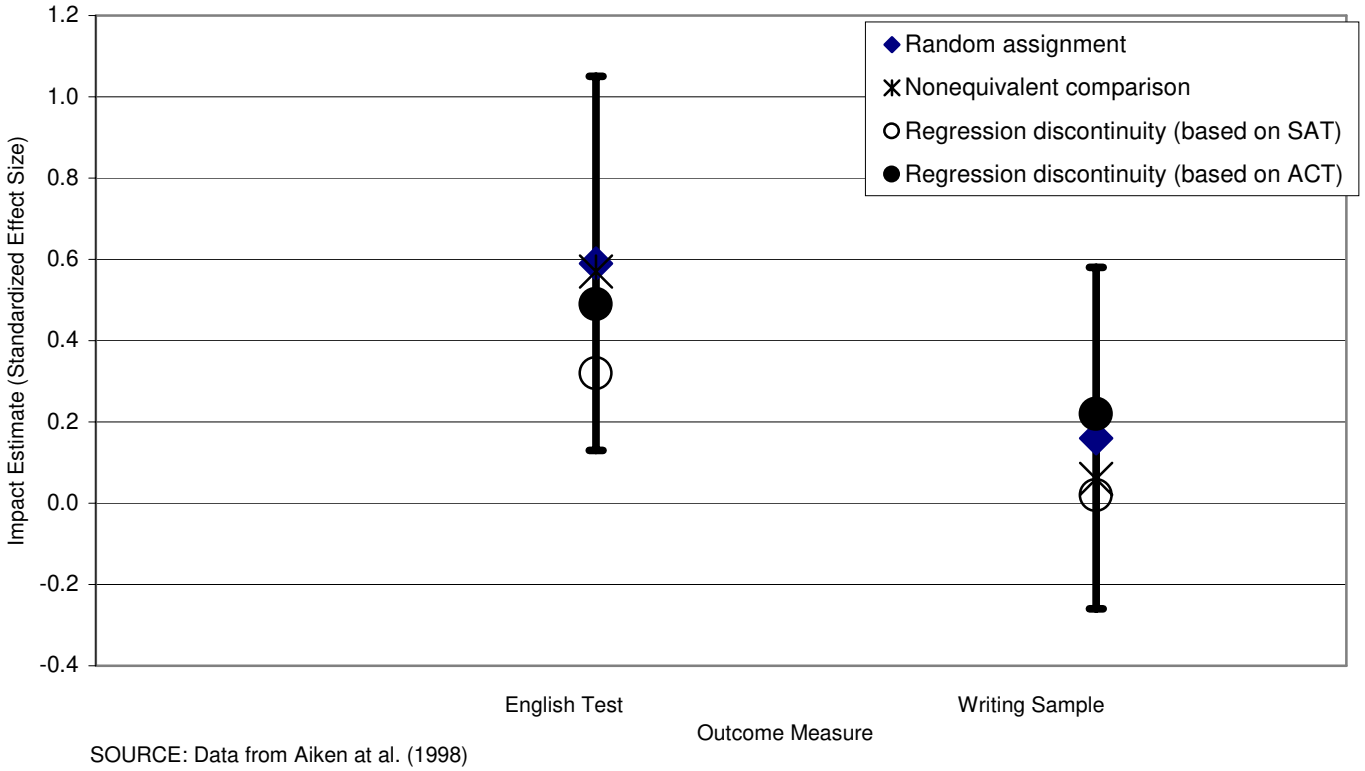
EXPERIMENTAL AND QUASI-EXPERIMENTAL IMPACT ESTIMATES
OF REDUCED CLASS SIZE, BY SCHOOL



SOURCE: Data from Wilde and Hollister (2002)

FIGURE B-3

EXPERIMENTAL AND QUASI-EXPERIMENTAL IMPACT ESTIMATES
OF REMEDIAL WRITING BY OUTCOME



SOURCE: Data from Aiken at al. (1998)

Statistical Method. All of the nonexperimental estimators in these three studies used either propensity score (PS) matching or least squares (LS) regression to adjust for differences between the nonrandomized comparison group and the treated population. These are shown in Figures B.1 and B.2 using different symbols. The bias, estimated by the vertical gaps between experimental and nonexperimental estimates, does not appear to be smaller for any one method. Whether background data are used as covariates in a linear regression or as predictors in a propensity score matching estimator, the impacts appear about equally inaccurate.

Educational Setting. One of these studies took place in elementary schools, while one took place in middle and high schools, and one in college. The size of the bias estimated in each of these settings will be numerically different, although the different units across studies and the confounding of setting with all the other study and intervention characteristics makes it difficult to infer that the methods generally work better in one setting than another.

Type of outcome. One of these studies used the dropout rate and the other two used student test scores as the primary outcome of interest. As with educational setting, the type of outcome is largely confounded with the study itself. With a sample size of three and not much variation in the accuracy, we are not able to identify any relationships from the available evidence.

Size of the impact. The sites are sorted in Figures B.1 and B.2 by the size of the experimental impact estimate. One might hypothesize that the accuracy of the nonexperimental estimator might be related to the presence or absence of a true program effect. In other words, the methods may have different rates of false positive and false negative findings. There does not, however, appear to be any relationship in the three figures between the size of the impact and the accuracy of the nonexperimental methods.

Chance. It may be that in fact there is no systematic bias in the nonexperimental estimators, but rather an element of additional variance or unpredictable "noise." Taking the Figure IV.1A as an example, the average difference between experimental and nonexperimental impacts across all eight middle school dropout prevention sites is less than two points for the regression adjusted estimates (-1.1 and 1.9 for the internal and NELS comparison groups, respectively) and less than four points for the propensity score matching more estimates (3.6 and –2.2 for the internal and NELS comparison groups). These averages are much smaller than the 5 to 10 point discrepancies found at any given site. Depending on the policy decision to be made, and how one interprets the variation in impacts across sites, this observation that aggregation reduces bias suggests that adding more program sites to the sample would result in an average prediction error that is even closer to zero, even though we have little ability to predict the performance in any one school.