NONEXPERIMENTAL VERSUS EXPERIMENTAL ESTIMATES OF EARNINGS IMPACTS

May 2003

Steven Glazerman
Dan Levy
David Myers

Draft submitted to the *Annals of the Academy of Political and Social Sciences*

NOT FOR CITATION OR QUOTATION

ABSTRACT

Nonexperimental or "quasi-experimental" evaluation methods, in which researchers use treatment and comparison groups without randomly assigning subjects to the groups, are often proposed as substitutes for randomized trials. Yet, nonexperimental (NX) methods rely on untestable assumptions. To assess these methods in the context of welfare, job training, and employment services programs, we synthesized the results of 12 design replication studies, case studies that try to replicate experimental impact estimates using NX methods. We interpret the difference between experimental and NX estimates of the impacts on participants' annual earnings as an estimate of bias in the NX estimator.

We found that NX methods sometimes came close to replicating experiments, but were often substantially off, in some cases by several thousand dollars. The wide variation in bias estimates has three sources. It reflects variation in the bias of NX methods as well as sampling variability in both the experimental and NX estimators.

We identified several factors associated with smaller bias; for example, comparison groups being drawn from the same labor market as the treatment population and pre-program earnings being used to adjust for individual differences. We found that matching methods, such as those using propensity scores, were not uniformly better than more traditional regression modeling. We found that specification tests were successful at eliminating some of the worst performing NX impact estimates. These findings suggest ways to improve a given NX research design, but do not provide strong assurance that such research designs would reliably replicate any particular well-run experiment.

If a single NX estimator cannot reliably replicate an experimental one, perhaps several estimators pertaining to different study sites, time periods, or methods might do so on average. We therefore examined the extent to which positive and negative bias estimates cancel out. We found that this did happen for the training and welfare programs we examined, but only when we looked across a wide range of studies, sites, and interventions. When we looked at individual interventions, the bias estimates did not always cancel out. We failed to identify an aggregation strategy that consistently removed bias while answering a focused question about earnings impacts of a program.

The lessons of this exercise suggest that the empirical evidence from the design replication literature can be used, in the context of training and welfare programs, to improve NX research designs, but on its own cannot justify their use. More design replication would be necessary to determine whether aggregation of NX evidence is a valid approach to research synthesis.

NONEXPERIMENTAL VERSUS EXPERIMENTAL ESTIMATES OF EARNINGS IMPACTS [1]

I. ASSESSING ALTERNATIVES TO SOCIAL EXPERIMENTS

Controlled experiments, where subjects are randomly assigned to receive interventions, are desirable but often thought to be infeasible or overly burdensome, especially in social settings. Therefore, researchers often substitute nonexperimental or "quasi-experimental" methods, in which researchers use treatment and comparison groups, but do not randomly assign subjects to the groups.[2] Nonexperimental (NX) methods are less intrusive and sometimes less costly than controlled experiments, but their validity rests on untestable assumptions about the differences between treatment and comparison groups.

Recently, a growing number of case studies have tried to use randomized experiments to validate NX methods. To date, this growing literature has not been integrated in a systematic review or meta-analysis. The most comprehensive summary (Bloom et al. 2002) addresses the portion of this literature dealing with mandatory welfare programs. However, efforts to put the

[1] This research was supported by grants from the William and Flora Hewlett Foundation and the Smith Richardson Foundation; however, the conclusions do not necessarily represent the official position or policies of the Hewlett Foundation or the Smith Richardson Foundation. The authors thank Harris Cooper, Phoebe Cottingham, Allen Schirm, Jeff Valentine, and participants of workshops held by the Campbell Collaboration, Child Trends, Mathematica, and the Smith Richardson Foundation. Also, we are grateful to the authors of the studies that we included in this review, many of whom spent time answering our questions and providing additional data.

[2] This paper uses the term "nonexperimental" as a synonym for "quasi-experimental," although "quasi-experimental" is used in places to connote a more purposeful attempt by the researcher to mimic randomized trials. In general, any approach that does not use random assignment is labeled nonexperimental.

quantitative bias estimates from these studies in a common metric and combine them to draw general lessons have been lacking.

This paper reports on a systematic review of such replication studies to assess the ability of NX designs to produce valid impacts of social programs on participants' earnings.[3]

Specifically, this paper addresses the following questions:

- Can NX methods approximate the results from a well-designed and well-executed experiment?
- Which NX methods are more likely to replicate impact estimates from a well-designed and well-executed experiment and under what conditions are they likely to perform better?
- Can averaging multiple NX impact estimates approximate the results from a well-designed and well-executed experiment?

The answers to these questions will help consumers of evaluation research, including those who conduct literature reviews and meta-analyses, decide whether and how to consider NX evidence. They will also help research designers decide, when random assignment is not feasible, whether there are conditions that justify a NX research design.

A. BETWEEN AND WITHIN STUDY COMPARISONS

Researchers use two types of empirical evidence to assess NX methods: between-study comparisons and within-study comparisons (Shadish 2000). This paper synthesizes evidence from within-study comparisons, but we describe between-study evidence as background.

Between-study comparisons. Between-study comparisons look at multiple studies that use different research designs and study samples to estimate the impact of the same type of program. By comparing results from experimental studies with those of NX ones, researchers try to derive

[3] Findings reported here are drawn from a research synthesis prepared under the guidelines of the Campbell Collaboration. The published protocol is available at http://www.campbellcollaboration.org/doc-pdf/qedprot.pdf.

the relationship between the design and the estimates of impact. Examples include Reynolds and Temple (1995), who compared three studies; and Cooper et al. (2000; Table 2), the National Research Council (2000; Chapter I, Tables 6–7), and Shadish and Ragsdale (1996)**,** who all compared dozens or hundreds of studies by including research design variables as moderators in their meta-analyses. These analyses produced mixed evidence on whether quasi-experiments produced higher or lower impact estimates than experiments.

An even more comprehensive between-study analysis by Lipsey and Wilson (1993) found mixed evidence as well. For many types of interventions, the average of the NX studies gives a slightly different answer from the average of the experimental studies, while, for some, it gives a markedly different answer. The authors found 74 meta-analyses that distinguished between randomized and nonrandomized treatment assignment and showed that the average effect sizes for the two groups were similar, 0.46 of a standard deviation from the experimental designs and 0.41 from the NX designs. But such findings were based on averages over a wide range of content domains, spanning nearly the entire applied psychology literature. Graphing the distribution of differences between random and nonrandom treatment assignment within each meta-analysis (where each one pertains to a single content domain), they showed that the average difference between findings based on experimental versus NX designs was close to zero, implying no bias. But 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. Thus, the between-study evidence does not resolve whether differences in impact estimates are due to design or to some other factor.

Within-study comparisons. In a within-study comparison, researchers estimate a program's impact by using a randomized control group, and then re-estimate the impact by using one or more nonrandomized comparison groups. We refer to these comparisons, described formally below, as “design replication” studies. The nonrandomized comparison groups are formed and their outcomes adjusted by using statistical or econometric techniques aimed at estimating or eliminating selection bias. Design replication studies can use multiple comparison groups or the same comparison group with multiple sample restrictions to examine the effect of different comparison group strategies. The NX estimate is meant to mimic what would have been estimated if a randomized experiment had not been conducted. If the NX estimate is close to the experimental estimate, then the NX technique is assumed to be “successful” at replicating an unbiased research design.

Within-study comparisons make it clear that the difference in findings between methods is attributable to the methods themselves, rather than to investigator bias, differences in how the intervention was implemented, or differences in treatment setting. For this reason, within-study comparisons can yield relatively clean estimates of selection bias. On the other hand, it is more difficult to rule out the effects of chance for a given set of within-study comparisons. Therefore, general conclusions require, as in this paper, several within-study comparisons in a variety of contexts.

B. DESIGN REPLICATION TO ESTIMATE BIAS

The current review differs from standard meta-analysis because the “effect size” of interest is not the impact of some intervention on a given outcome, but the discrepancy between experimental and NX impact estimates. This, we argue, is itself an estimate of the bias. Bias can never be directly observed, because the true impact, θ, is not known. This review includes two equivalent types of studies that allow us to estimate the bias empirically. The first type presents up to *K* NX estimators, $\hat{\theta}_k$, of the impact, where *k*=1,…*K*, and one experimental estimate, $\hat{\theta}_0$, such that E[$\hat{\theta}_0$]=θ. The second type compares average outcome for a control group, $\overline{Y}_0$, with the (adjusted) average outcome, $\overline{Y}_k$, for some comparison group based on NX method *k*. The relationship among these variables is shown in equations (1) and (2), where $\overline{Y}_T$ represents the average outcome for the treated group and $B(\hat{\theta}_k)$ is the bias.

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

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

Using these estimates, we can estimate the bias associated with each of the k estimators, defined as $B(\hat{\theta}_k) = E[\hat{\theta}_k - \theta]$. Since the true parameter is not observable, we estimate the bias as the difference between the NX and experimental estimators. Subtracting equation (2) from equation (1) yields two forms of the bias estimate, $\hat{B}(\hat{\theta}_k)$, corresponding to the two types of reporting formats discussed above:

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

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

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

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

The goal of the analysis in this review is to model $\hat{B}(\hat{\theta}_k)$ as a function of the characteristics and context of the study, the estimator, and the intervention whose impact is being estimated. We recognize an important practical limitation in estimating such models, which is that the observed bias estimates vary not only because of the performance of the NX method (in reducing selection bias) and other contextual variables already noted, but because of random sampling error in both the experimental and NX estimators. This sampling variance makes it difficult to judge when bias estimates are large enough or spread out enough to be evidence that a NX method has failed. Therefore, we have refrained throughout the analysis from making general statements that go beyond the collection of case studies we reviewed.

II. DATA AND METHODS

In recent years the number of design replication studies has been growing to the point where it is now possible to begin synthesizing the results to look for patterns. This paper draws on such a synthesis of the design replication literature, focusing on studies that used earnings as an outcome.[4] The rest of this section describes the methods we used to assemble the design replication studies, construct the dataset, and conduct the analysis.

A. INCLUSION CRITERIA AND SEARCH STRATEGY

To be included in the review, a 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 NX estimate of the same impact.*** Because some studies estimated bias directly by comparing comparison and control groups, the presence of a treatment group is not required.
- ***The experimental-NX comparison was based on estimates from the same experiment.*** This criterion excludes the between-study comparisons described in Section I.
- ***The experimental and NX estimates pertained to the same intervention in the same sites.*** This criterion excludes, for example, a study of education programs in Bolivia (Newman et al. 2002), which compared findings from an experimental design in one

[4] A broader review we have undertaken for the Campbell Collaboration includes design replication studies that estimate bias for other outcomes, such as student achievement, school dropout, and receipt of pubic assistance benefits. Forthcoming results from that study examine binary indicators for whether the experimental and NX estimators support the same statistical inference and whether they support the same policy conclusion.

region with findings from a NX design in another. Such a study confounds regional differences with differences in study design.

- ***The intervention's purpose was to raise participants' earnings.*** This criterion restricts our focus to programs that provide job training and employment services.[5]

The search process produced dozens of candidate studies. We narrowed them down to 33 for closer examination, and determined that 12, listed in Table II.1, met the search criteria. The 12 studies correspond to 9 interventions; four of these studies addressed the same intervention, the National Supported Work Demonstration (NSW). All of the interventions involved job training or employment services, such as job search assistance or vocational rehabilitation, and participation was mandatory in about half of them. In terms of location, three interventions were single-site programs (in San Diego, CA, Riverside, CA, and Bergen, Norway); one was a multisite program in a single state (Florida); and the remaining five were multistate in the U.S. Half of the interventions were studied in the 1990s; only one (NSW) was studied before 1980. Seven of the studies appeared in peer-reviewed journals or in books; three are final reports of government contractors; and two are working papers or unpublished manuscripts.

The quality of the evidence in these studies—in particular, the quality of the experiment—is critical to our analysis. The use of design replication as a validation exercise assumes that the experimental estimators in the studies are themselves unbiased.[6] Common threats to the validity

[5] An important area excluded by this criterion was health-related interventions (for example, MacKay et al. 1995 and 1998). 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. We initially applied a broader criterion that included school-related outcomes such as school dropout and test scores, but ultimately focused on interventions with earnings as the main outcome to limit the number of confounding factors. A forthcoming Campbell review will draw on the wider literature.

[6] 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

of the experimental estimator include: differential attrition or nonresponse, randomization bias, spillover effects, substitution bias, John Henry effects, and Hawthorne effects.[7] Bias could also arise from non-uniform collection of data from treatment and control groups and from assignments that were not truly random. Noncompliance with treatment assignment, even if monitored and documented, can threaten an experiment's ability to answer interesting policy questions.

To evaluate the validity of the experimental estimates, we assessed the nine experiments in our review and found them to be of generally high quality. Most were well-funded and were carried out by research organizations with established track records in random assignment and data collection. The Manpower Demonstration Research Corporation (MDRC) oversaw random assignment in four of the experiments; Abt Associates oversaw two; and Mathematica Policy Research (MPR), two. The remaining experiment was overseen by university-based researchers. Because details of the experimental designs and their implementation were not reported in all the replication studies, we retrieved background reports and methodological appendixes, examined nonresponse analyses, and corresponded with researchers. We concluded that most of the experiments had relatively low crossover and attrition rates and that the attrition and nonresponse

(continued)
treated or the local average treatment effect), as long as the NX estimator purports to measure the same thing.

[7] 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 that 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 of their inclusion in an experiment.

did not appear to be related to treatment status in a way that would threaten the conclusions' validity.

B. PROTOCOL AND CODING

Once the studies were assembled, we followed a procedure laid out in a formal protocol (Glazerman et al. 2002) to extract data from the source studies and code them for analysis. For example, the coding form had questions about the source of the comparison group used for each NX estimator in each study. Two coders read the 12 studies and extracted all the information needed for the analysis. They coded two studies together to ensure a consistent understanding of the coding instrument and process. Then, each one coded a subset of the rest, with ample consultation built into the process to increase coding accuracy (see Glazerman et al. 2002 for details on this and other aspects of the research synthesis methods). We also contacted authors of nearly every source study to obtain clarification and, sometimes, additional data. Further details of the variables that were coded are mentioned below.

C. ANALYSIS METHODS

The goal of our analysis is to determine how selection bias varies with the type of estimator employed, the setting, and the interaction between the setting and the type of estimator. To answer this, we model $B(\hat{\theta}_{jk})$, the bias associated with estimator k, as a function of the characteristics and context of the study (indexed by *j*) and its intervention, captured in a vector labeled *Z*, and the characteristics of the estimator itself, captured in a vector labeled *W*.

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

We use the absolute value of $B(\hat{\theta}_{jk})$ on the left-hand side of the equation, because a researcher or research synthesist wants to choose designs to minimize the bias, whatever its direction. An interaction between study-level and estimator-level variables is included to capture the interplay between method and context.

One might expect that some types of NX design perform better than others, and that some designs are more appropriate under certain study conditions. To test this, it is important to describe each type of NX estimator. Of the many classification schemes, the most commonly used are those given by Campbell and Stanley (1966) and Campbell and Cook (1979). Alternative formulations by economists such as Heckman and Hotz (1989) and Heckman et al. (1998) are also useful for categorizing methods in a general way; however, we prefer to avoid forcing the methods into mutually exclusive categories, because many of the estimators used multiple approaches. Instead, we describe each estimator by a vector of characteristics that pertain to: (1) the source of the comparison group and (2) the analytic techniques used to adjust for differences between the comparison group and the treatment population.

Because there is a limited range of NX designs assessed in the design replication literature, we must use very gross indicators to categorize NX designs. For the source of the comparison group, we coded three binary indicator variables: one for whether the comparison group is drawn from a national dataset, such as Survey of Income and Program Participation (SIPP); one for whether the comparison group is based on sample members from the same geographic area as the treated population; and one for whether the comparison group is formed by using the randomized control group from a different experiment. For the type of statistical adjustment, we used four variables to broadly indicate: (1) whether background variables were used as covariates in a regression model; (2) whether matching methods, such as stratification on an estimated propensity score, were used; (3) whether the estimator used pre-intervention measures of the outcome—examples include difference-in-differences models, fixed effect models, or even regression or matching models using baseline earnings; and (4) whether the estimator was based on an econometric sample selection model. The selection modeling indicator (4) would be set to one, for example, if the estimator used the inverse Mills' ratio or instrumental variables with a set of variables that were included in a model of program participation, but excluded from the model of earnings determination.

We constructed other indicators to identify conditions under which quasi-experiments were potentially more likely to replicate experiments. One set of indicators measured whether a specification test was conducted, and if so, whether the test would have led the researcher to avoid the estimator a priori. Another set of indicators measured whether the background variables used in the regression or in the matching procedure were detailed, as with a survey, or sparse, as is typically the case with administrative data. Variables included in the *W* vector include the experiment's sample size, grouped into categories for small, medium, and large; and the program's estimated effectiveness—effective, ineffective, or indeterminate.

To estimate the average bias reduction associated with these design and context variables, we used both bivariate analyses (tabulations) and multivariate analyses (regression). Because such a small collection of studies limits the degrees of freedom, we expect to find the data consistent with several competing explanations for why the estimated bias is high or low. The bivariate analyses use sample weights to account for the unequal sample sizes of the source studies, although we found that weighting made little difference to the qualitative findings. Similarly, for the multivariate analyses, we tried alternative aggregation procedures to deal with lack of statistical independence among bias estimates from a single study. To minimize artificial replication, the regression results in the next chapter use the average of the absolute value of the bias estimates associated with each unique combination of design variables. For example, if one study produced eight quarterly bias estimates corresponding to impacts after random assignment, we aggregated them into a single estimate for the two-year period, as long as the policy interpretation for the two-year period made sense.

A constraint on more detailed analyses than those just described was dictated by having just 12 replication studies. While many of these studies assessed multiple NX estimators, resulting in more than 1,000 bias estimates, the overall diversity of designs was not as comprehensive a catalogue of quasi-experimental methods as those described by Cook and Campbell (1979) and others. Among those methods that were assessed, not every method was assessed in every setting. As more empirical work comes to light, more sophisticated analysis may be possible.

III. RESULTS

Our synthesis of design replication studies takes an important first step toward answering the research questions of this paper. However, interpretation of the evidence remains a challenge. Even among authors of the studies we reviewed there was no consensus on how to judge the size of differences between experimental and NX impact estimates. The authors differed in the extent to which they probed the statistical and policy significance of their results. Some focused narrowly on their own case studies; others made broader statements praising or condemning a NX method. Four studies concluded that nonexperimental methods performed well; four found evidence that some nonexperimental methods performed well while others did not; and four found that nonexperimental methods did not perform well or that there was insufficient evidence that they did perform well. A summary of their conclusions is given in Appendix A. In this section, univariate analyses describe the range of bias estimates in the literature. Bivariate analyses then relate the absolute size of the bias to several explanatory factors. The multivariate analysis that follows uses regression to determine whether the different explanations of bias overlap and whether one predominates. Finally, we examine the distribution of the bias estimates to consider whether they cancel out across studies and whether their variation is due to true variation in the performance of NX methods or some other explanation.

A. UNIVARIATE ANALYSIS

From the 12 studies, we extracted 1,150 separate estimates of the bias, about 96 estimates per study. While some of the bias estimates were close to zero, some were very large, over- or under-estimating annual earnings impacts by as much as $10,000 or more. Table III.1 shows the bias estimates by study.

The definition of a "large" bias depends on the program and the policy decision at stake. However, 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 about $1,200 was used to justify the program's expenditure levels, one of the highest per trainee (about $16,500) for any federal training program. A difference of $800 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 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, a difference of $1,000 or more, can make a dramatic difference in the policy recommendation.

Another benchmark is the average earnings of control group members. In many of the studies we reviewed, the inflation adjusted annual earnings of control group members was about $10,000, which includes zero earnings for non-workers. Thus a $1,000 bias would represent 10 percent of earnings, a substantial amount.

As mentioned earlier, one should interpret the statistics in Table III.1 with caution. The average of the bias estimates can be substantially influenced by outliers reflecting small samples or unrealistic estimators. However, the average does indicate whether the estimates are centered on zero and whether they tend to over- or underestimate impacts relative to the experimental benchmark. Eight of the 12 studies in our analysis showed that NX methods tended to understate impacts; four showed the opposite. All the studies included bias estimates that were both negative and positive, except for the one by Bratberg et al., in which all the econometric and matching techniques had negative bias estimate. As one would expect, the study with the greatest *number* of estimates (Bloom et al. 2000) found the broadest *range* of estimates, with very large positive and negative values.[8]

The absolute value of the bias provides a more direct measure of the performance of the NX estimator, where a smaller value always represents better performance. With that measure, the typical NX estimate of impact on annual earnings deviates from the corresponding experimental estimate by about $2,000. The average absolute value in any one study ranged from twice that amount—as in the attempts by Dehejia and Wahba and by Smith and Todd to replicate the findings of the NSW experiment using national datasets—to less than $600 per year, as in the two studies by Hotz and colleagues.

B. BIVARIATE ANALYSIS

To begin to explain the range of NX bias, we conducted simple bivariate analyses, examining the relationship between several possible explanatory variables and the size of the bias. The candidate variables are those that describe the quasi-experimental approach and the

[8] It is important to recall that a wide range of bias *estimates* does not necessarily imply a wide range of biases, because of sampling error in both the experimental and NX impact estimates.

study in which it was implemented, including the source of the comparison group, the statistical method, and the quality of the data.

For each value of an explanatory variable, we computed the average of the absolute value of the bias for all NX estimators with that value (see Table III.2). For the entire sample of studies, the unweighted average of the absolute value of the bias associated with using NX methods was about $1,500. However, this was based on all 1,150 bias estimates without aggregating to account for non-independence of the estimates or unequal sample size. Therefore, we constructed two sets of weights. The first (weight 1) gives more emphasis to estimates based on studies that had larger samples as measured by the number of control group members in the randomized experiment; the other (weight 2) multiplies the sample-size weight by a factor inversely proportional to the number of estimates for a given sample. For example, if a researcher used 10 different methods to estimate the same impact for one site or subgroup, then the corresponding bias estimates received a weight of 1/10 times the sample size. Although the results vary somewhat by type of weight used, the qualitative conclusions drawn from them do not, so we focus on the results in the last column, which account for sample size and frequency of sample. Both weights reduce the average absolute value of the bias to about $1,100.

Table III.2 shows that some factors are indeed associated with higher and lower bias. As one would expect, the source of the comparison group has a role. The average bias was lower (under $900) when the comparison group came from the same labor market as the treated population or was composed of randomized control group members from a separate experiment, and higher (over $2,000) when the comparison group was drawn from a national dataset. This finding suggests that, while convenient, publicly available datasets at the national level are not the best for evaluating training or welfare programs.

Aspects of the statistical method also were associated with the size of the bias. There was little difference between regression and matching methods overall, but some matching methods performed better than others. In particular, one-to-one propensity score matching had lower bias than other propensity score methods or non-propensity score matching. Five of the studies (Lalonde 1986; Heckman et al. 1998; Gritz and Johnson 2001; Bratberg et al. 2002; and Bloom et al. 2002) included some form of econometric selection correction procedure such as the Heckman two-step estimator or instrumental variables estimator, but these methods performed poorly on average, about as poorly as using no method at all.

Rather than examine all quasi-experimental estimators, it may be more productive to focus on the performance of those one would expect (in the absence of a randomized experiment) to be the best ones. To make such *a priori* predictions, researchers use specification tests, as illustrated by Heckman and Hotz (1989) in their re-analysis of Lalonde's replications of the NSW experiment. The typical specification test applies the NX estimator to outcome data from before the intervention. If the estimated impacts, which should be zero since nobody has been exposed to the intervention, are larger than would be expected by chance, then the estimator is rejected, and its use, not recommended. Many of the design replication studies that we reviewed did not conduct specification tests. Among those that did, the average absolute bias of rejected estimators was nearly $2,900, almost three times the bias of recommended ones. This suggests, consistent with the findings of Heckman and Hotz, that specification testing, where feasible, can help eliminate poor-performing estimators. The estimated bias of the recommended estimators, however, was still large in absolute value, over $1,000.

Some authors (Heckman et al. 1998; Smith and Todd 2001) have suggested that data quality may be as important as the research design. By categorizing estimators by the richness of the background variables —used as covariates in a regression or as matching variables— to explain the size of the bias, we found some support for this claim.[9] The results in Table III.2 suggest that the estimators based on a more extensive set of variables in a regression or matching method had lower bias. The most important variable to include in the variable set was prior earnings. Studies without it had a bias of about $1,600; those with it, $1,000.[10]

Finally, we found that the performance of NX methods was related to the sample size and direction of impacts for the experiment. Specifically, the NX methods more closely replicated the experiments when the randomized control groups were large and when the experiments did not show the program was effective. One possible explanation for the large average bias (over $2,700) in small studies is that the experimental impacts were not precisely estimated, so the estimate of bias is also not precisely estimated. Another possible explanation is the size of the nonrandomized comparison group, which tends to be small when the control group is small, so the larger estimated bias may reflect random noise in the NX estimate. Because the sample sizes of control and comparison groups are correlated, it is difficult to distinguish between these two stories. The relationship between the direction of the experimental impact and the size of bias suggests that a false positive finding—concluding from the NX evidence that a program works when it does not—may be more common than a false negative.

[9] The coding of the variables representing quality of background data (for regression or matching) necessarily involves some subjectivity. To be systematic we applied the following criteria: If the specification included several quarters of baseline earnings and a large number of relevant background variables, we coded the quality of the data as “very extensive.” If the specification contained some baseline measure of earnings and a set of individual background variables that captures the key elements that are likely to affect outcomes, then it was coded as “extensive.” Otherwise, it was coded as “poor.”

[10] Some researchers such as Bloom et al. (2002) and Smith and Todd (2002) tried to determine the number of quarters of prior earnings needed to reduce bias to acceptable levels, but there were not enough other examples to draw any general lessons.

A limitation of this bivariate analysis is that the design elements listed in Table III.2 are not independent. For example, a study that uses a national data set to select a comparison group is likely also to use a relatively poor set of controls; this means that the large average bias for studies that use a national data set could be reflecting a poor set of controls. It is difficult to distinguish these explanations. We therefore proceed with multivariate regression analysis to try to disentangle the factors associated with lower bias.

C. MULTIVARIATE ANALYSIS

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 as explanatory variables (see Table III.3). As suggested earlier, other types of explanatory variables could also explain bias. However, we have limited degrees of freedom, so we use a parsimonious model that includes indicator variables for each design replication study to proxy for all the measured and unmeasured characteristics that vary only at the study level. Because the regression models focus on NX design configurations, we did not weight the individual bias estimates. We averaged them within design types so that each design type would be represented no more than once for each study. This aggregation resulted in an analysis dataset of 69 bias estimates. The regression results are meant to be illustrative, because some of the design attributes are highly correlated with each other, the data set is very small, and the results depend on the regression specification used.

Keeping these limitations in mind, we found the regressions largely confirm what one would expect. 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 models shown in odd-numbered columns represents the bias associated with raw mean differences, estimated to be in the range of $4,400 to $5,800 in annual earnings (see row

1). This coefficient is the expected bias, if one did not make any adjustments to the "average" comparison group in our sample. In the regression models shown in the even-numbered columns we include a separate intercept for each study, a study-level fixed effect describe above.

The entries in the next two rows suggest that using background data as either covariates or matching variables is about equally effective at reducing bias. These techniques reduce the bias by about $3,100 to $3,600, once we account for the studies' fixed effects (column 6). The sensitivity of this result to the inclusion of fixed effects suggests that the relative performance of regression-based designs versus matching designs is confounded with contextual factors.

Combining methods is better than applying them individually. Models (5) and (6) include an interaction term with a positive coefficient, which suggests that the bias reduction from these two methods is not fully additive, although there is likely some increased benefit from their combination. In model (5), for example, the bias from raw differences in means, represented by the intercept, is $5,775. This value is reduced to $2,550, if only regression is used, and to $3,312, if only matching is used (holding comparison group variables fixed at the value of the omitted categories). If matching and regression are both used, they reinforce each other to reduce the bias to $1,038.

Baseline measures of the outcome are important. This is suggested by the negative coefficients on the difference-in-difference indicator, which equals one if the estimator uses pre-intervention earnings in any way, show that using baseline measures of the outcome is important, as reported in the literature. For the simpler models in (2) and (4), difference-in-difference estimators reduce the bias by about $1,600 in annual earnings, a reduction slightly larger than that achieved with 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.

The one estimator that did not reduce bias at all, in fact increased it, was the selection correction estimator, but this should be interpreted cautiously. Few estimates in our data were based on econometric methods such as the two-step estimator. Of these, one study (Bratberg et al. 2000) rejected the specification based on a hypothesis test, but still reported the bias estimate, which was particularly large.[11] Of the others, none produced a compelling justification for the exclusion restrictions that typically justify such an approach. An exclusion restriction is an assumption that some variable is known to influence participation in the program (selection into treatment) but not the outcome of interest (earnings).

The use of a comparison group that is matched to the same labor market or geographic area reduced bias by about $600. Funders of evaluation research probably prefer to use large national data sets to evaluate programs, because secondary analyses are far less costly than new data collection. Our findings suggest that such a strategy comes with a penalty, an *increase* of average bias by about $1,700 (column 6).

We coded another comparison group strategy that determined whether the source was a control group from another study or another site. Several studies—for example, those by Hotz et al. (1999 and 2000) and Bloom et al. (2002)—compared the control group from one site to the control group from another site and labeled one as the nonrandomized comparison group. We included the "control group from another site" indicator variable in the regression primarily to distinguish between those studies from others that used comparison groups that are more readily available to researchers, such as eligible nonapplicants (for example, Heckman et al. 1998) or

[11] 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 U.S. welfare participants. Some of this effect is measured by the study fixed effect (see the even-numbered columns in Table III.3).

individuals who applied to the program but were screened out (for example, Bell et al. 1995). One might argue that control groups are not available to most evaluators, so the more relevant bias estimates are the larger ones found when the “other control group” indicator equals zero.

The regression analysis described above is robust to the definition of the dependent variable. We conducted the same analysis using the signed value of the bias and found very similar results. Those results, available from the authors, show that, overall, the unadjusted bias is large and negative. The regressors representing design features increase the bias (toward zero) in much the same that that they decreased the absolute value of the bias as shown in Table III.3. Other dependent variables can be used to further analyze the bias estimates. For example, we created two indicator variables, one for whether the NX impact estimate led to the same statistical inference and another for whether it led to the same policy conclusion. Constructing these variables required some additional information, such as the threshold value that would change the policy conclusion, but they allow us to include in a meta-analysis the results from a wider range of design replication studies, including those that focus on education interventions and those whose outcomes are not measured in earnings. The analyses based on these binary outcome variables are beyond the scope of the current paper and will be presented in future work.

D. AGGREGATION OF NONEXPERIMENTAL EVIDENCE

Now we turn to the third research question, whether averaging multiple NX impact estimates can approximate the results from a well-designed and well-executed experiment. The above discussion suggests that, while some factors are associated with lower bias estimates, a single NX estimator cannot reliably replicate an experimental one. The inability to achieve reliable replication may be to due bias or it may also be due to sampling error in either the experiment or the quasi-experiment. However, a possibility exists that a large enough group of estimates pertaining to different study sites, time periods, or interventions, might do so on average. We therefore examined the extent to which positive and negative bias estimates cancel out. If they do, it would provide motivation for those who conduct literature reviews to be able to accumulate a large body of NX evidence, when experiments are scarce, to draw valid conclusions. A useful way to make this assessment is by examining the full distribution of bias estimates for various groupings, such as by intervention, by method, or for a collection of interventions, and looking for a pattern of estimates that tends toward zero.

The distribution of the 1,150 bias estimates from the 12 studies reviewed in this paper provides a case where the bias estimates do appear to cancel out (Figure III.1). The distribution is nearly centered on zero with a slight skew. The average bias was about -$600. Applying the weights described above brings the overall bias closer to zero, -$217; and removing the outliers and applying weights makes it even smaller, about -$97.[12] This is a crude indicator, but suggests, consistent with the work of Lipsey and Wilson (1993), that, if enough NX studies are combined, the average effect will be close to what the experimental evidence would predict.[13]

Rarely, however, is the NX research used to answer such broad questions as whether all programs are effective. Instead, we would like to identify dimensions along which the bias begins to cancel out for more focused questions such as "What is the average impact of Program X?" For the studies reviewed in this paper, the average bias was sometimes close to zero (see

[12] Removing the outliers in this case is probably reasonable because the outlying bias estimates correspond to NX impact estimates that were implausible on their face (given the collection of other impact estimates). One cannot count on being able to identify this type of outlier in general, when an experimental benchmark is not available.

[13] We also examined the distribution across studies for a given method—matching and regression—and found a similar result. This suggests that aggregation need not be done across methods if a large collection of studies and interventions is used.

Table III.1), but often was still substantial. Each of the studies in the review addressed a single intervention, although some assessed more NX estimators, analyzed more subgroups, had larger samples, or included more sites. The distribution of bias estimates within studies—particularly studies that use multiple sub-groups, study sites, or time periods, in addition to multiple estimators—makes this clearer. Figures III.2 and III.3 display the distribution of bias estimates for two of the studies that examined the largest number of estimators. For the first study (Bloom et al. 2002), the bias estimates are centered roughly on zero, with an average of -$151; but for the second study (Smith and Todd, 2002), they clearly are not, with an average of -$2,563. It is possible to remove outliers from the estimates reported by Smith and Todd to achieve an average bias that is closer to zero, but identifying outliers without the benefit of a randomized experiment as a benchmark may be difficult. The within-study evidence from the other studies (Table III.1), suggests that the average bias across all methods, subgroups, and time periods is sometimes positive, sometimes negative, and often still in the hundreds of dollars. This suggests that a mechanistic application of a large number of NX estimators might improve the inference one could draw from such evidence, but not in a predictable way. Whether the average bias, properly weighted within and between studies, is really close enough to zero for policy makers, and whether the bias cancels out within a narrower domain of research, are questions that we plan to address as more design replication studies are completed.

IV. WHAT HAVE WE LEARNED ABOUT NONEXPERIMENTAL METHODS?

Our preliminary review of the evidence suggests that the 12 design replication case studies we identified, even taken together, will not resolve any of the longstanding debates about NX methods. From the case studies we uncovered some factors that might reduce bias, but we have not identified a reliable strategy for eliminating it either in a single study or in a collection of studies. The findings can be summarized in terms of the three empirical questions posed in Section I.

1. ***Question: Can NX methods approximate the results from a well-designed and well-executed experiment?***

   Answer: Occasionally, but many NX estimators produced results dramatically different from the experimental benchmark.

   - Quantitative analysis of the bias estimates underscored the potential for very large bias. Some NX impact estimates fell within a few hundred dollars of the experimental estimate, but others were off by several thousand dollars.
   - The size and direction of the "average" bias depends on how the average is computed and what weighting assumptions are applied.
   - The average of the absolute bias over all studies was over $1,000, which is about ten percent of annual earnings for a typical population of disadvantaged workers.

2. ***Question: Which NX methods are more likely to replicate impact estimates from a well-designed and well-executed experiment and under what conditions are they likely to perform better?***

   Answer: We identified some factors associated with lower estimated bias. However, even with these factors present, the estimated bias was often large.

   - The source of the comparison group made a difference in the average bias estimate. For example, bias was lower when the comparison group was: drawn from within the evaluation itself rather than from a national dataset; locally matched to the treatment population; or drawn as a control group in an evaluation of a similar program or the same program at a different study site.
   - Statistical adjustments, in general, reduced bias, but the bias reduction associated with the most common methods— regression, propensity score matching, or other forms of matching—did not differ substantially. Estimators that combined methods had the lowest bias. Classical econometric estimators that used an instrumental variable or a separate predictor of program participation performed poorly.
   - Bias was lower when researchers used measures of pre-program earnings and other detailed background measures to control for individual differences.
   - Specification tests were useful in eliminating the poorest performing NX estimators.
   - Experiments with larger samples were more likely to be closely replicated than those with smaller samples.

- "No impact" or indeterminate impact findings from an experiment were more nearly replicated than were positive experimental impact findings.

3. ***Question: Can averaging multiple NX impact estimates approximate the results from a well-designed and well-executed experiment?***

   Answer: Maybe, but we have not identified an aggregation strategy that consistently removed bias while answering a focused question about earnings impacts.

   - Estimated biases were both positive and negative, and their distribution across all the studies reviewed was centered roughly on zero. This was true both for the full set of estimators and for groups of estimates across all studies that used a single method, such as regression or matching.
   - For a given intervention, the distribution of bias estimates was sometimes centered near zero, and sometimes was not.

We caution that this summary of findings gives only part of the picture and it does so for a specific area of program evaluation research: the impacts of job training and welfare programs on participant earnings. A somewhat more complete story can be developed in the short term as additional design replication studies, including some that are now in progress, come to light.

In the meantime, those who plan and design new studies to evaluate the impacts of training or welfare programs on participants' earnings can use the empirical evidence to improve NX evaluation designs, but not to justify their use. Similarly, those who wish to summarize a group of NX studies or average over a set of different NX estimates to reach a conclusion about the impact of a single program can draw on the design replication literature to identify stronger or weaker estimates, but not to justify the validity of such a summary.

REFERENCES

General References

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].

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.

Glazerman, Steven M., Dan M. Levy, and David Myers. “NX Replications of Social Experiments in Education, Training, and Employment Services” Revised Protocol. Philadelphia, PA: The Campbell Collaboration. December, 2002. http://www.campbellcollaboration.org/doc-pdf/qedprot.pdf

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

Heckman, James J., and V. Joseph Hotz. “Choosing Among Alternative NX 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.

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

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.

Studies in the Systematic Review

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 NX Methods." Scandinavian Journal of Economics, vol. 104, no. 1., 2002, pp. 147-171.

Dehejia, Rajeev, and Sadek Wahba. "Causal Effects in NX 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.

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.

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 NX Estimators?" *Journal of Econometrics*, forthcoming 2002.

TABLE III.1

DESCRIPTIVE STATISTICS OF BIAS ESTIMATES BY STUDY

	Bias Estimates (Annual Earnings in 1996 Dollars)				
Study	Range of Estimates	Average of Estimates	Average of Absolute Value of Types of Estimates[a]	Number of Estimates	Number of Types of Estimates
Bell et al. 1995	[-$723, +$5,008]	$661	$813	54	3
Bloom et al. 2002	[-21,251, +12,215]	498	1,114	564	8
Bratberg et al. 2002	[-18,702, -654]	-4,826	2,907	13	5
Dehejia and Wahba 1999	[-1,939, +1,212]	173	4,163	40	4
Fraker and Maynard 1987	[-3,673, +871]	-751	1,103	48	3
Gritz and Johnson 2002	[-1,091, +3,189]	497	780	48	2
Heckman et al. 1998	[-7,669, +8,154]	-423	3,273	45	17
Hotz, et al. 2000	[-1,682, +2,192]	-128	585	36	2
Hotz, et al. 1999	[-1,248, +438]	-174	371	64	6
Lalonde 1986	[-5,853, +4,143]	-636	2,849	112	8
Olsen and Decker 2001	[-1,548, +1,107]	-363	1,397	10	5
Smith and Todd 2002	[-11,743, +4,829]	-1,655	4,019	116	6
Total	**[-$21,251, +$12,215]**	**-$637**	**$2,325**	**1,150**	**69**

NOTES:

[a] The average of the absolute value of the bias is used to compare different research designs. Therefore we calculate it by first averaging bias within design type and then averaging the results across the 69 design types.

TABLE III.2

AVERAGE BIAS BY CHARACTERISTICS OF ESTIMATOR

	Average of Absolute Value of Bias Estimate (Annual Earnings in 1996 Dollars)		
Explanatory Variable Categories	Unweighted	Weight 1 (sample size)	Weight 2 (sample size, frequency)
Entire Sample	$1,477	$1,101	$1,110
Source of Comparison Group[a]			
Same labor market	932	821	885
Control group from another site	843	902	814
National data set	2,817	2,409	2,131
Statistical Method: General[a]			
Regression	1,101	1,010	958
Matching	1,143	828	924
Selection correction or instrumental variables	2,251	2,071	1,412
None, simple mean differences	2,791	1,323	1,515
Statistical Method: Type of Matching			
Propensity score matching: one to one	1,047	739	744
Propensity score matching: one to many	1,181	852	929
Other matching technique	1,231	1,037	1,297
Did not use matching	1,750	1,288	1,311
Statistical Method: Specification Test Result			
Specification not recommended	4,027	3,165	2,870
Specification recommended	1,155	857	1,103
No test conducted	1,247	1,047	988
Quality of Background Data: Regression			
Poor set of controls	2,336	1,438	1,590
Extensive set of controls	1,228	1,030	1,036
Very extensive set of controls	1,026	1,008	1,016
Did not use regression	2,431	1,412	1,589
Quality of Background Data: Matching			
Poor set of covariates	1,752	1,290	1,313
Extensive set of covariates	1,392	951	1,330
Very extensive set of covariates	1,113	802	920
Did not use matching	1,750	1,288	1,311
Quality of Background Data: Overall			
Used prior earnings	1,224	1,040	1,003
Did not use prior earnings	2,662	1,379	1,591
Experimental Sample Size			
Small (<500 controls)	2,533	2,378	2,728
Medium (500 to 1,500 controls)	1,080	1,001	960
Large (>1,500 controls)	819	819	800
Experimental Impact Finding			
Program is effective	2,089	1,288	1,276
Program is ineffective	1,197	920	1,105
Indeterminate	924	1,021	911
Number of observations	1,150	1,150	1,150

[a] Categories are not mutually exclusive or exhaustive.

TABLE III.3

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

	Model Specification					
Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	4,467 ***		4,687 ***		5,775 ***	
	(657)		(1,231)		(1,120)	
Statistical Method						
Regression	-1,583 **	-1,516 **	-1,476 **	-1,416 **	-3,225 ***	-3,572 ***
	(729)	(705)	(675)	(706)	(1,195)	(1,284)
Matching	-478 *	-1,268 **	-807 **	-1,427 ***	-2,463 *	-3,178 **
	(715)	(794)	(692)	(799)	(1,375)	(1,508)
(Regression) x (Matching)					951	1,320
					(1,422)	(1,484)
Difference-in-differences	-1,874	-1,596	-1,859	-1,568	-3,532 ***	-3,231 **
	(763)	(816)	(718)	(813)	(1,253)	(1,336)
(Regression) x (Diff-in-diffs)					2,325	2,676 *
					(1,455)	(1,600)
(Matching) x (Diff-in-diffs)					1,889	1,774
					(1,477)	(1,547)
Selection correction	2,508 *	2,376	4,619	2,441	3,291 ***	3,072 **
	(1,248)	(1,305)	(1,048)	(1,299)	(1,163)	(1,284)
Comparison Group Strategy						
Geographic match			-387	-646	-673	-581
			(973)	(1,182)	(957)	(1,160)
National dataset			1,145	1,695	915	1,668
			(1,062)	(1,536)	(1,043)	(1,479)
Control group from another site			-1,762	N/A	-2,124 **	-1,346
			(1,011)	N/A	(995)	(2,863)
Study dummies included	No	Yes	No	Yes	No	Yes
Number of studies	12	12	12	12	12	12
Number of bias estimate types (cells)	69	69	69	69	69	69

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

FIGURE III.1

DISTRIBUTION OF BIAS ESTIMATES FOR ALL 12 STUDIES

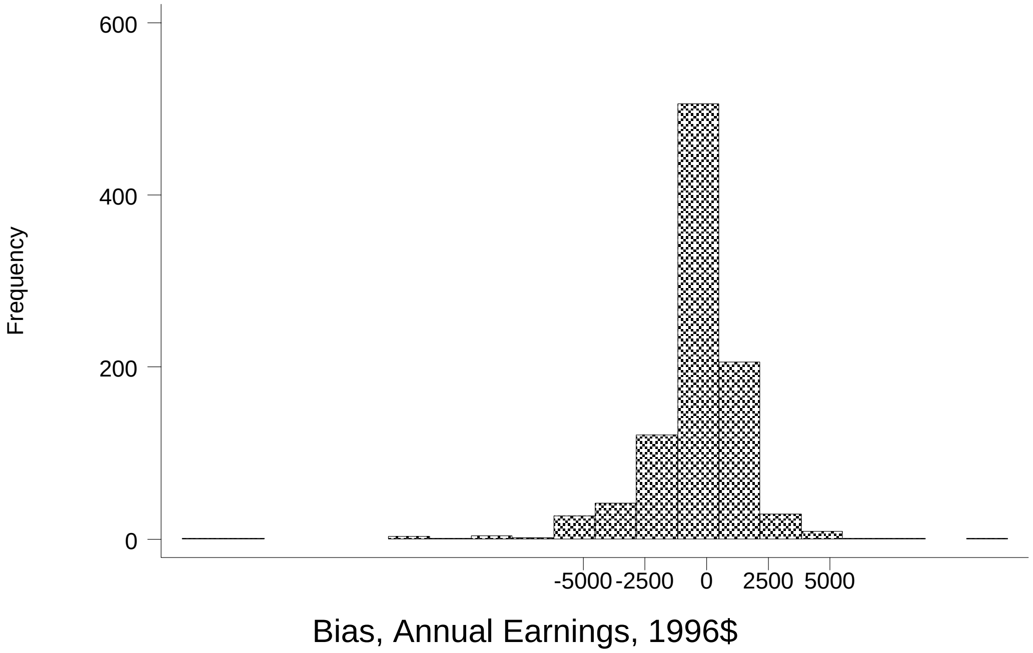


FIGURE III.2

DISTRIBUTION OF BIAS ESTIMATES FROM NEWWS (Bloom 2002)

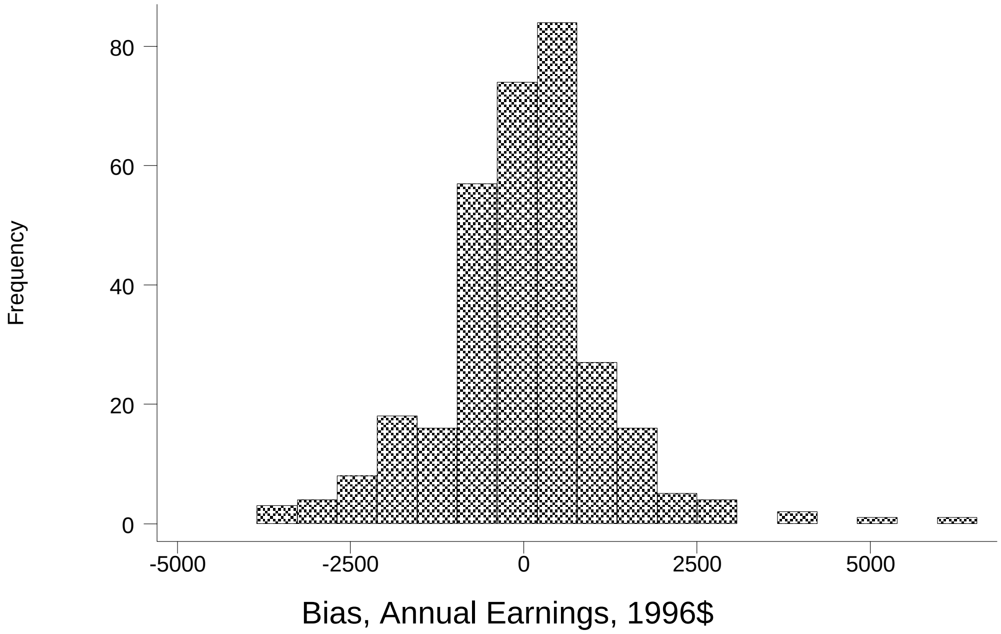
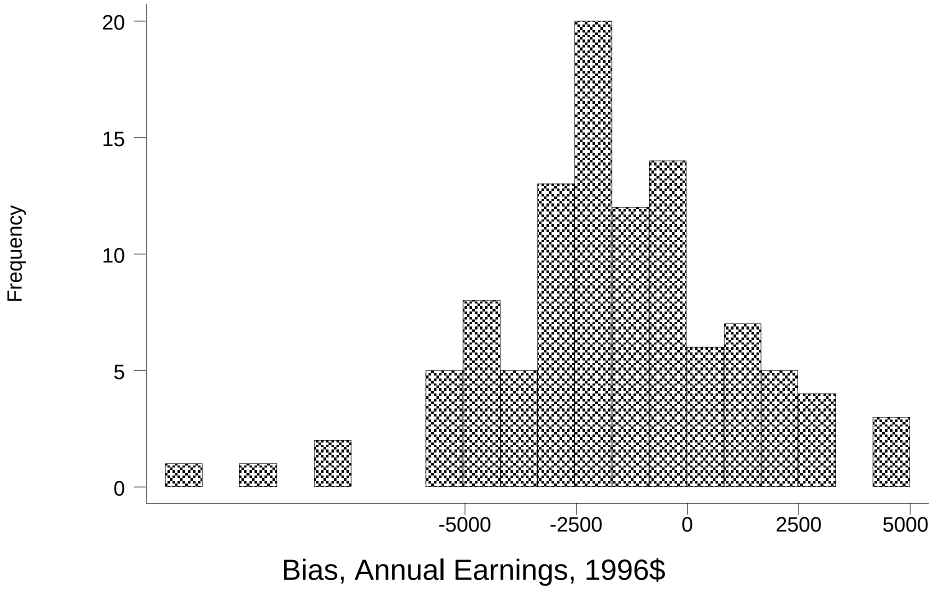


FIGURE III.3

DISTRIBUTION OF BIAS ESTIMATES FROM SUPPORTED WORK (Smith and Todd)



APPENDIX A

WHAT DID THE STUDIES CONCLUDE?

An alternative way to review the literature is to summarize what the authors concluded in their own words. The 12 design replication studies divided into three equal groups about the value of nonexperimental methods and the degree of similarity between nonexperimental findings and those of randomized experiments: four studies concluded that nonexperimental methods performed well; four found evidence that some nonexperimental methods performed well while others did not; and four found that nonexperimental methods did not perform well or that there was insufficient evidence that they did perform well (see Table A.1).

The four studies that found positive results (evidence of small bias) qualified their conclusions by indicating that a researcher needs detailed background data (particularly prior earnings), overlap in background characteristics, or intake workers' subjective ratings of the applicants they screened.

It is important to probe the authors' conclusions further than the present discussion allows. The various study authors used different standards to assess the size of the bias and, in some cases, reached different conclusions with the same data. Furthermore, the studies are not of equal value. Some more realistically replicated what would have been done in the absence of random assignment than 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. A recent summary by Bloom et al. (2002; Chapter 2) describes many of these studies individually. Section III of this paper presents a quantitative analysis of all the studies combined.

TABLE A.1

AUTHORS CONCLUSION FROM THE STUDIES REVIEWED

Authors and Year of Publication	Type of Intervention[a]	Methods Examined[b]	Comparison Samples Used[c]	Study Conclusion (Verbatim)
Nonexperimental Estimators Performed Well				
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	Occupational 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.”
Gritz and Johnson 2001	Job training	PSM, SC	No-shows	“Caution should be used in interpreting the results from the application of either matching method…The specification checks for the econometric models suggest that the application of either approach to gauge the impact of participation in Job Corps will yield biased estimates.”
Heckman et al. 1998	Job training	OLS, PSM, DD, SC	Eligible nonparticipants; 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 A.1 *(continued)*

Authors and Year of Publication	Type of Intervention[a]	Methods Examined[b]	Comparison Samples Used[c]	Study Conclusion (Verbatim)
Nonexperimental Estimators Performed Poorly				
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.”
Lalonde 1986	Supported work	OLS, DD, SC	CPS, PSID	“Many of the econometric procedures do not replicate the experimentally determined results.”
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.”

[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; CPS = Current Population Survey; PSID = Panel Study of Income Dynamics