Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons

Thomas D. Cook
William R. Shadish
Vivian C. Wong

Abstract

This paper analyzes 12 recent within-study comparisons contrasting causal estimates from a randomized experiment with those from an observational study sharing the same treatment group. The aim is to test whether different causal estimates result when a counterfactual group is formed, either with or without random assignment, and when statistical adjustments for selection are made in the group from which random assignment is absent. We identify three studies comparing experiments and regression-discontinuity (RD) studies. They produce quite comparable causal estimates at points around the RD cutoff. We identify three other studies where the quasi-experiment involves careful intact group matching on the pretest. Despite the logical possibility of hidden bias in this instance, all three cases also reproduce their experimental estimates, especially if the match is geographically local. We then identify two studies where the treatment and nonrandomized comparison groups manifestly differ at pretest but where the selection process into treatment is completely or very plausibly known. Here too, experimental results are recreated. Two of the remaining studies result in correspondent experimental and nonexperimental results under some circumstances but not others, while two others produce different experimental and nonexperimental estimates, though in each case the observational study was poorly designed and analyzed. Such evidence is more promising than what was achieved in past within-study comparisons, most involving job training. Reasons for this difference are discussed. © 2008 by the Association for Public Policy Analysis and Management.

INTRODUCTION

Comprehensive program evaluation depends on validly determining a program's causal impacts. Debate has been vigorous about the role experiments and observational studies should play in identifying such impacts. The main reason for preferring experiments is that, when perfectly implemented, they create intervention and control groups that do not initially differ in expectation and so do not differ on any measured *or unmeasured* variables. However, the regression-discontinuity design (RD) and instrumental variables (IV) also provide unbiased causal inference in theory. So additional technical justification for preferring experiments is required. It comes from experimental estimates being more precise than RD and IV estimates (Goldberger, 1972) and also from the experiment's assumptions being more transparent in research practice. IV's main assumption is that the instrument is only correlated with outcome through treatment. This assumption is well warranted when

Journal of Policy Analysis and Management, Vol. 27, No. 4, 724–750 (2008)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20375

the treatment is allocated via random assignment (Angrist, Imbens, & Rubin, 1996) or RD (Hahn, Todd, & van der Klauuw, 2001), but otherwise it is usually unclear whether the assumption is met in practice. RD's main assumption is that the functional form relating the assignment variable to outcome is fully known. This can usually be checked quite well in the data, but perhaps not quite as well as with the major threats to experiments—poor initial randomization, differential attrition, and treatment crossovers. Experiments have other advantages too. Their findings are more general since both RD and IV estimate local average treatment effects (LATE), while the experiment estimates an average treatment effect (ATE) across all units in the treatment group. And the results of experiments are considered more credible in most high-level policy settings.

The final justification for experiments seems to come from within-study comparisons. These take the effect size from an experiment and contrast it with the effect size from an observational study sharing the same treatment group. The observational data are then analyzed to take account of selection differences, mostly using OLS, propensity scores, or some form of IV. The purpose is to see if the adjusted observational study generates the same causal estimate as the experiment, the latter functioning as the benchmark. The most recent reviews conclude that the experiments and adjusted observational studies have nearly always differed in their causal estimates (Glazerman, Levy, & Myers, 2003; Bloom, Michalopoulos, & Hill, 2005). Since the theoretical rationale for the experiment is so compelling, the implication is that individual observational studies are often biased in the causal estimates they produce.

The debate between experiments and observational studies is now closed in several disciplines. The randomized experiment reigns supreme, institutionally supported through its privileged role in graduate training, research funding, and academic publishing. However, the debate is not closed in all areas of economics, sociology, and political science or in interdisciplinary fields that look to them for methodological advice, such as public policy. In these fields, concern is real about the difficulty of mounting an experiment on many topics of interest (Heckman & Smith, 1995; Heckman, LaLonde, & Smith, 1999), about the limited generality of experimental versus survey data (Zellner & Rossi, 1986), and about the many technical problems that have beset recent large social experiments such as the National JTPA or Moving to Opportunity studies (Orr et al., 1996; Ludwig, Duncan, & Ladd, 2003). Alternatives to the experiment will always be needed, and a key issue is to identify which kinds of observational studies are most likely to generate unbiased results. We use the within-study comparison literature for that purpose.

The earliest within-study comparisons (LaLonde, 1986; Fraker & Maynard, 1987) used data from job training evaluations. The randomized experiment involved a treatment implemented quite locally, with the nonequivalent comparison cases being chosen from extant national datasets such as the Panel Study on Income Dynamics or the Current Population Survey. OLS or Heckman-type (IV) selection models were then used to adjust for selection. Since quite different effect sizes were found across the set of experiments and observational studies examined, the conclusion was drawn that statistical adjustment procedures do not eliminate selection bias in observational studies. So experiments are not just preferred; they are considered to be uniquely relevant as instruments for warranting causal conclusions.

But the procedure used in these early studies contrasts the causal estimate from a locally conducted experiment with the causal estimate from an observational study whose comparison data come from national datasets. Thus, the two counterfactual groups differ in more than whether they were formed at random or not; they also differ in where respondents lived, when and how they were tested, and even in the actual outcome measures. Later within-study comparisons have moved toward identical measurement in the treatment, randomized control, and nonequivalent comparison group as well as toward treatment and comparison units that are as

local as possible, though the same exact location is rarely feasible (Heckman, Ichimura, & Todd, 1997, 1998; Heckman et al., 1998; Smith & Todd, 2005). The aspiration is to create an experiment and an observational study that are identical in everything except for how the control and comparison groups were formed.

This is not the only way in which the design of within-study comparisons has evolved over time. Another is to omit the treatment group data altogether since they are redundant in a comparison of whether a counterfactual group was formed at random or not (for example, Bloom, Michalopoulos, & Hill, 2005). Yet another is to use newer analytic tools to adjust the observational data for selection, especially propensity scores in later years. The substantive scope of within-study comparisons has also expanded. Examples from domains other than job training are now available. In addition, formal reviews of the within-study comparison literature now exist, albeit limited to job training. Glazerman, Levy, and Myers (2003) synthesized 12 within-study comparisons meta-analytically while Bloom, Michalopoulos, and Hill (2005) and Smith and Todd (2005) have provided qualitative syntheses. These reviews have led to three major claims.

First, in all within-study comparisons except one, the experimental and adjusted observational study estimates were judged different. And the one exception (Dehejia & Wahba, 1999) was undermined when Smith and Todd (2005) showed its results to be sensitive to minor alternative specifications of the population. The finding that causal estimates were so different has had considerable resonance because it seems to undercut the validity of the analytic tools used in some social sciences to adjust for selection.

Second, past reviews have identified some conditions associated with greater (but not adequate) similarity in the experimental and observational study estimates. These are when the adjustments include pretest values of the outcome; when nonequivalent comparison cases are local; when measurement is similar in the randomly and nonrandomly formed counterfactual groups; when OLS or propensity score methods are used but not Heckman-type selection models; and when data quality is higher for constructing multivariate matches.

Third, when Glazerman, Levy, and Myers (2003) summed the within-study comparison results in job training, they concluded that the mean experimental estimate did not differ from the mean observational estimate. Bloom, Michalopoulos, and Hill (2005) concluded the same, as did Lipsey and Wilson (1993) across multiple substantive areas. There is no compelling theoretical reason why study-specific biases should exactly cancel out across studies; indeed, it is easy to imagine situations where selection bias operates systematically—more in one direction than another. So we place little weight on the average effect size from a set of experiments tending to agree with the average effect size from a set of observational studies on the same topic.

The present paper has three main and three incidental purposes. Incidentally, it seeks to update the literature since Glazerman, Levy, and Myers (2003) and Bloom, Michalopoulos, and Hill (2005), thus including examples from domains other than job training. We have closely examined the job training studies before 2005. While they are generally of only modest technical merit as within-study comparisons, our own reading agrees with the outline of their substantive conclusions. So instead of re-reviewing the studies here, we concentrate on more recent within-study comparisons that allow us to broach three research questions that are more finely differentiated than whether experiments and observational studies produce similar causal results. Specifically, we decompose the notion of observational study to ask three central questions:

(1) Do experiments and RD studies produce comparable effect sizes? We would expect this, since the selection process into treatment is completely known for each method (Thistlethwaite & Campbell, 1960; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). But many vagaries are associated with how multi-year

experiments and RD studies are actually implemented and with the many discretionary decisions every data analyst has to make. So similar causal estimates cannot be taken for granted. Moreover, experiments are more efficient than RD (Goldberger, 1972). For identical sample sizes, the two methods should therefore differ in standard errors and statistical significance patterns even if they are similar in causal estimates.

(2) Are estimates similar when an experiment is compared to an observational study whose sampling design uses intact group matching to minimize any initial differences between the intervention and comparison populations on pretest means and/or slopes? Most discussions of matching assume two or more different populations (for example, smokers and nonsmokers; attending job training or not) and individuals from these populations are then matched in some way or another. But sampling plans can also seek to minimize population differences before allocating treatments. Such matching can be of intact groups rather than individuals, and should involve matching on the highest correlates of outcome, usually pretest measures of the outcome itself. The matching can also be local, from the same geographical area as the treatment cases come. To exemplify such matching, imagine being asked to evaluate a school reform about to enter selected schools. Comparison schools can be selected from within the same school district or subdistrict, matching them to intervention schools on multiple prior years of achievement and on their race or class composition. The comparison schools so selected would still differ from the intervention schools on some unmeasured correlates of treatment, and so hidden bias is a possibility. But the total bias will nearly always be less than when students from intervention schools are matched with individual students from schools in the same or other districts. We will test whether experiments provide similar causal answers to observational studies whose prospective sampling plan calls for maximizing treatment and comparison group overlap and so minimizes population differences instead of, or prior to, any individual case matching.

(3) The dominant question in within-study comparisons assumes that the treatment and comparison populations obviously differ and asks whether statistical adjustments for this are effective. In within-study comparisons of job training, clear group differences are the norm and the failure of selection adjustments there inclines us not to expect much correspondence between experimental and observational study results. At most, we hope to identify some conditions under which experiments and observational studies produce comparable answers despite obvious population differences. A likely condition for this is when the selection process into treatment is perfectly known, this being the key feature of both experiments and RD. The problem for research practice, though, is that we are rarely sure that the selection process is completely known.

A second incidental purpose follows from the fact that "observational study" conflates "quasi-experiment" and "nonexperiment." In *quasi-experiments*, original data are expressly collected to test a specific causal hypothesis in a specific setting. To rule out alternative causal interpretations, heavy reliance is placed on both design and researcher control. Of great importance, therefore, are implementing the intervention at a known point in time, collecting pretest data on the outcome and at several pre-intervention times if possible, selecting comparison groups to minimize initial differences without necessarily eliminating them, deliberately collecting original data on other variables correlated with both outcome and selection into treatment, and even replicating the intervention at different known times. *Nonexperiments* involve less researcher control over the treatment and choice of comparison cases, less prospectively designed and original data collection, greater reliance on survey and other archival sources for generating comparison data, and a greater emphasis on statistical modeling than design control to rule out selection bias. As a result, the effectiveness of statistical adjustments for selection tends to be more central in nonexperimental than in quasi-experimental work. This characterization

is admittedly fuzzy at the margin. For instance, Cook and Campbell (1979) represent the design tradition but they include interrupted time series (ITS) in their pantheon of better quasi-experimental designs even though many of them involve archival data and no researcher control over the intervention. The key with ITS is the presumption of treatment exogeneity, the same assumption that dominates in work on natural experiments, whether with or without multiple data collection waves. Though the distinction between nonexperiments and quasi-experiments is fuzzy, an incidental purpose of this paper is to explore whether studies emphasizing the structural features of quasi-experimentation do a better job of bias reduction than the statistical modeling features that characterize research better classified as nonexperimental than quasi-experimental.

The final incidental purpose of this study is to explicate what constitutes a better-designed within-study comparison. This task is necessary because, if experimental and observational study results did not coincide, this could be because of real differences between the methods, or flaws in the way these method differences were tested. So we explicate seven criteria for warranting a within-study comparison as being of high technical merit.

CRITERIA FOR EVALUATING WITHIN-STUDY COMPARISONS

Enumerating the Criteria

1. There must be two or more counterfactual groups varying in whether treatment assignment was random or not. We use "control group" for a randomly formed counterfactual and "comparison group" for a nonrandom one.
2. The experiment and observational study should estimate the same causal quantity. One of them should not involve, say, an intent to treat (ITT) causal estimator and the other a treatment on treated one (TOT). Similarly, since the local average treatment effect (LATE) in RD is assessed at the cutoff, the experiment's average treatment effect (ATE) should be estimated at that same point. We should not confound how comparison groups are formed with differences in estimators.
3. The contrast between control and comparison groups should not be correlated with other variables related to study outcome. So the contrast groups should be measured in the same ways at the same times; they should undergo the same experiences between treatment assignment and posttest measurement; and to the extent possible, they should come from the same geographic location. Otherwise, the method contrast of interest is confounded with extraneous details like these.
4. Analysts of the experimental and observational data should be blind to each other's results. This is to prevent the publication of analyses that might inadvertently exaggerate or minimize method differences in causal results.
5. The experiment should meet all of the usual criteria of technical adequacy. So the treatment and control group should be properly randomized and there should be no differential attrition or treatment crossovers. Otherwise, the experiment cannot serve as a valid causal benchmark,
6. The observational study should meet all the usual technical criteria as a good example of its type. To contrast a good experiment with a poor exemplar of a particular observational study is to confound the assignment mechanism of interest with irrelevant quality differences between design types. Sometimes the criteria for a high quality observational study are easy to specify. With RD, for example, they are an assignment variable, a cutoff score, and an outcome, plus analytic sensitivity to specified functional form, fuzzy allocation, reduced statistical power, and limited causal generalization (Cook, 2008). However, quality standards are not all as clear in other observational study

types. When selection processes are not well known, two or more nonequivalent groups are desirable, as are pretest and posttest measure of the same outcome. But it is less clear how nonequivalent groups should be selected to minimize population differences or how covariates should be selected to adjust for selection, including hidden bias. Technical quality standards vary by observational study type, and some are more opaque than one might like.

7. Within-study comparisons require deciding how comparable the experimental and observational study estimates are. Even in a well-conducted experiment, sampling and measurement error render causal estimates partially unreliable. Assuming power of 0.80, two exact replicates will result in similar statistical significance patterns in only 68 percent of comparisons—the probability of both studies producing a reliable finding is 0.80×0.80 and of both generating a nonsignificant one is 0.20×0.20. Large sample sizes reduce this problem, but even better would be to directly test the obtained difference between the two estimates. But such direct tests are rare in the literature except for Black, Galdo, and Smith (2005), Wilde and Hollister (2007), and McKenzie, Gibson, and Stillman (2007). Analysts can also compare the different estimates to an external criterion; but justifying such a criterion is not easy. Glazerman, Levy, and Myers (2003) suggested that job training results are similar if the annual earnings difference is $1,000 or less. But this standard has not been adopted. One could also ask whether a different policy decision would ensue because of the difference in experimental and nonexperimental estimates (Wilde & Hollister, 2007). However, this requires many (and often heroic) assumptions about how policy decisions might vary because effect sizes vary, assumptions that could also vary with the substance of a within-study comparison. Some fuzziness in determining how close the experimental and observational study results should be is inevitable, and we present our results in terms of method differences in (a) statistical significance patterns, (b) causal estimates in the original metric or a standardized one, and (c) the percentage difference between causal estimates when they are not close to zero.

Discussing These Seven Criteria

The seven criteria above have evolved over time, and few were available to the pioneers of within-study comparisons, whose main motivation seems to have been to ascertain whether comparison groups could be inexpensively created from extant data sets. The goal for within-study comparisons has now subsequently changed, coming to emphasize the comparison of experimental and observational study estimates with everything else held constant. This required ever-tighter control over correlates of treatment assignment and outcome, and has made later studies technically superior to earlier ones. This review is based on within-study comparisons published in the last decade.

The seven criteria above reflect a traditional theory-testing goal. The independent variable is the mode of treatment assignment, the outcome is the difference between causal estimates in the experiment and the adjusted observational study, and everything else that is correlated with the assignment difference is held constant. The goal is to learn which observational study methods are better empirically warranted than others. But within-study comparisons have also been used to probe the quality of the causal methods used in a given field and hence also to probe the validity of any substantive claims these methods have warranted. This purpose requires comparing experiments to the most common types of observational study in a discipline, even if their technical quality is below the current state of the art. In this rationale, if substantive researchers often create comparison groups from national data sets, then this is the practice to be examined even if it fails to test stronger observational study methods. The distinction

here is like that between efficacy and effectiveness experiments (Flay, 1986). Efficacy studies evaluate the effects of best practice, while effectiveness studies examine the effects of modal or common practice. Our purpose in this paper is efficacy-related, designed more to identify better causal practice in observational studies than to pass judgment on current practice in any discipline or field of application.

CASE 1: CONTRASTING EXPERIMENTAL AND REGRESSION-DISCONTINUITY DESIGN RESULTS

Aiken, West, Schwalm, Carroll, and Hsiung (1998)

Aiken et al. (1998) evaluated whether enrolling in a remedial writing class at a large state university would improve writing skills. Assignment to the class was based on ACT or SAT cutoff scores, depending on which test a student took for admission to the university. All students scoring below the relevant cutoff were treated and all those scoring above it were not, thus enabling the RD design. The randomized experiment was restricted to a sample of students whose "admission test scores . . . fell within a fixed range just below the normal cutoff scores" (p. 212). They were asked to volunteer for an experiment randomly assigning them to the remedial course or going straight into regular English writing classes. Thus, the ATE for the experiment is estimated at a point just below the LATE for the RD estimate. Also, the RD sample had to take the course to which their ACT or SAT scores assigned them, whereas students could refuse the randomization invitation and still stay at the university. The two study outcomes were performance on a scored essay and also on a multiple choice Test of Standard Written English (TSWE). Sample sizes were modest: 108 in the experiment and 240 in the RD study—118 for the ACT and 122 for the SAT assignment variables.

The experiment produced a statistically significant standardized effect size of 0.59 on the TSWE, and the RD produced a reliable effect of 0.49 for the ACT assignment and a nonreliable 0.32 for the SAT. On the writing task, all effect sizes were nonsignificant—0.16 in the experiment, 0.22 for the ACT variable in RD, and 0.02 for the SAT. Thus, the three largest effects were for TSWE, irrespective of design type, and the three smallest effects were for the writing test, again irrespective of design type. The pattern of statistical significance between the experiment and its RD comparisons was similar in three of four cases. In magnitude terms, the RD and experimental estimates differed by 0.10 standard deviation units with ACT and by 0.27 with SAT. The corresponding writing test differences were 0.06 and 0.14. Given this pattern of results, Aiken et al. (1998) concluded that their experiment and RD study produced comparable results. In our judgment, this claim is better warranted for the ACT assignment variable, where the RD and experimental estimates differed by less than 20 percent whatever the outcome. But when SAT was the assignment variable, effect size differences were larger and statistical significance patterns less concordant. It is difficult to know why this was the case. But sample sizes were modest to begin with and were reduced further by the partition into ACT and SAT scores. Since RD estimates are also less efficient than experimental ones, we might not be surprised that RD estimates based on small and then partitioned samples are not very stable.

Buddelmeyer and Skoufias (2003)

The authors reanalyzed data from PROGRESA in Mexico. Villages were randomly assigned to PROGRESA, and the experiment compared eligible families in villages randomly assigned to PROGRESA with eligible households in control villages. Eligibility depended on the score on a family material resources scale, thus enabling RD. Sample sizes were large, and so the authors compared experimental

and RD results across different gender groups, types of outcome, and rounds of data collection. From fear that the assignment and outcome variables were not linearly related, the data were analyzed in several nonparametric ways, all predicated on weighting observations closer to the cutoff more heavily. This cutoff is not where the experimental ATE is estimated since the experimental cases were distributed across all the assignment range below the cutoff.

The authors' results corroborate their claim of comparable RD and experimental estimates. Whatever the design, PROGRESA had little to no impact on boys' or girls' work activities in either round of follow-up. But each of the designs affected boys' and girls' school attendance in the second round but not the first. Examining all the outcomes across gender, the RD and experimental estimates showed similar patterns of statistical significance in over 80 percent of the tests. As to magnitude estimates, the nonreliable experimental and RD effects are very small and obviously convergent. The school attendance results were not close to their experimental benchmarks in round one, but in round two they were almost identical for girls, and within 60–70 percent of each other for boys, depending on the kernel estimator used. As with Aiken et al. (1998), the agreement between causal estimates is not total but is close despite the experimental and RD estimates being made at different points on the assignment variable.

Black, Galdo, and Smith (2005)

Black, Galdo, and Smith (2005) conducted the most rigorous within-study comparison of RD. They reanalyzed data from a job training experiment in Kentucky to take advantage of the fact that individuals were assigned to training based on a single score from a 140-item test predicting the likelihood of long-term unemployment—the feature making an RD study possible. The experiment was designed so that individuals scoring in a symmetrical interval around the RD cutoff were randomly assigned to job training or control status, thus entailing that the same causal entity was estimated in both the experiment and the RD study. The dependent variables were weeks of unemployment insurance (UI) benefits, amount of such benefits, and annual earnings from work. Since the assignment and outcome variables were not linearly related, nonparametric analyses were used. Also examined was how the correspondence between experimental and RD results varied with proximity to the cutoff.

Black, Galdo, and Smith (2005) discovered a close correspondence between the experimental and RD results near the cutoff. They analyzed results across the 24 cells formed by crossing three outcomes, two analysis directions (from above and below the cutoff), and four estimators, of which one was parametric and three nonparametic. They found that none of the six experimental and parametric RD estimates reliably differed. Nonparametrically, only one of the six differed either for the smoothed RD estimate or the one-sided kernel estimator that Hahn, Todd, and van der Klaauw (2001) recommend as best current analytic practice. Two of the six comparisons differed for the simple Wald estimator, but the analysts expressed least confidence in this estimator. So the experimental and RD estimates rarely differed close to the cutoff.

But this correspondence might be a product of large standard errors aggravated by the need to use bootstrapping to compute them. However, the numerical estimates for the number of weeks receiving UI seem close, whether bias is from above or below. The estimates seem more discrepant for annual earnings, but this is mostly because estimates with bias from above and below are analyzed separately. Averaging them results in an experimental estimate of about $1,350, a parametric RD estimate of about $715, a smoothed RD estimate of about $845, a Wald estimator of about $950, and a one-sided kernel estimator of about $1,350. For the UI benefit outcome, the estimate is −$15 in the experiment, and −$123, $18, $85, and $300

in the RD estimates. None is reliably different from zero. All these estimates come from the narrow band around the cutoff where cause is most validly estimated in RD. Expanding the band reduces the correspondence between experimental and RD estimates, perhaps because of nonlinearity in how the assignment and outcome variables are related. In any event, statistical theory reveals one condition where the experiment and RD generate a similar causal conclusion—namely, at the same LATE—and another where the correspondence should be less—namely, at other points on the assignment variable when functional form is nonlinear.

It is not theoretically surprising that all three studies yoking an RD study to an experiment have produced quite similar causal estimates. More surprising are two things. First, that this has now occurred in three very complex situations where implementation of both the experiment and RD study might be expected to be very imperfect—in Mexican villages, freshman English classes at the largest U.S. state university, and a statewide Kentucky job training program. Also important is that the correspondence was sometimes achieved in imperfect analyses—as with the different causal estimators in the PROGRESA experiment and RD study. The implication is that RD is a robustly effective tool that mere mortals can use if they have considerable, but not necessarily perfect, sensitivity to its assumptions.

Second, the experiment and RD study should produce different standard errors and statistical significance patterns. But that was generally not the case here. This is probably because, in two cases, the experiments had fewer units. Without stratification by ACT and SAT scores, Aiken et al. (1998) had 108 experimental cases and 240 RD ones; Black, Galdo, and Smith (2005) had 1,964 experimental cases and almost 55,000 RD ones. In Buddelmeyer and Skoufias (2003), the sample sizes were similar but so high that each design had very high power anyway—9,575 experimental and 9,183 RD cases. Of course, it is not inevitable that RD studies will have more cases or so many cases that statistical significance hardly matters. But larger sample sizes may occur more often with RD since the range on the assignment variables can be broader than is found (or needed) with an experiment. It is not a contradiction to claim that theory predicts smaller standard errors in experiments and then to discover in the examples considered here that statistical significance patterns hardly differed between the experiment and RD study.

CASE II: CONTRASTING EXPERIMENTS AND OBSERVATIONAL STUDIES WITH MATCHED INTACT COMPARISON GROUPS

Most observational studies compare nonequivalent populations that differ in treatment exposure and for which pre- and post-intervention data are available. The pre-intervention data involve a single pretest observation on the outcome measure, several such waves data, or proxy measures if the same instrument is not available at pretest and posttest. Studies with these design features constitute the vast majority of within-study comparisons and are responsible for the claim that observational studies usually fail to reproduce experimental results.

However, researchers often have some discretion over how comparison groups are selected, and they can choose intact group matches instead of—or prior to—individual unit matches. The issue is: Do similar causal estimates result from experiments and observational studies when the intact comparison groups are purposively selected to maximize overlap with the treatment group on at least pretest values of the outcome measure? Of course, nonequivalent comparison groups can be selected on other attributes too, especially for being geographically local. But the emphasis here is on matching an entire treatment group with one or more intact comparison groups on pretest means and/or slopes. If such matching is successfully achieved, the major internal validity concern is then with hidden bias due to omitted variables. We now describe the three studies we could find where an experiment is explicitly compared to intact group matching.

Aiken, West, Schwalm, Carroll, and Hsiung (1998)

The authors' experiment on the efficacy of a college remedial English course was compared to a matched group design as well as the RD study discussed earlier. The comparison group was composed of volunteers who were not contacted the summer before their first quarter or who applied after pretest information for the experiment and RD study had been collected. The experimental and comparison students were all volunteers, but the volunteering dynamic was not necessarily identical in each group, adding to the other sources of nonequivalence.

Admission to the university required having a SAT or ACT score, and the observational study sample was restricted to those hard-to-reach and late applicants whose scores fell within the same narrow bandwidth that was used for selection into the randomized experiment. This is the key for reducing population nonequivalence on factors correlated with the writing outcomes. Indeed, Aiken et al. (1998) discovered that entry-level ACT and SAT scores did not differ between the control and comparison groups (see their Table 2). Nor did pretest essay writing scores or pretest multiple choice scores assessing knowledge of English (see their Table 3), though these last tests may have been taken somewhat later by the students applying late or who could not be contacted over summer. The fact that the control and comparison groups did not differ on pre-intervention observables highly correlated with writing skill implies that the study sampling design sufficed to eliminate all bias on these observables, limiting bias concerns to unobserved variables. ANCOVA was used to analyze the observational study, with each pretest outcome measure serving as a control for its own outcome value. This would ordinarily not be a strong analysis, but the close pretest correspondence meant that the covariate served more to increase precision than to reduce bias. The authors presented their results in standard deviation units. The reliable experimental and observational study effect sizes were 0.57 and 0.59 for the test of knowledge of English, while for essay writing the nonreliable effect sizes were 0.16 and 0.06. Thus, by criteria of both effect size magnitude and statistical significance, the experiment and observational study produced comparable results. This does not necessarily mean that hidden bias was ruled out, of course, since countervailing bias from other sources could have reduced the total bias to zero.

Aiken et al. (1998) was a particularly good within-study comparison. The randomized experiment was carefully managed, demonstrating pretest means did not differ and differential attrition did not occur between the experiment and observational study. The two groups also underwent similar experiences and measurement schedules, except perhaps for posttest timing, thus reducing the threat of bias from third variables. The same causal estimator was used in the experiment and observational study. The achieved effect sizes were so close that it is easy to conclude that the experimental and observational study results were not different. The experimental and observational study populations were clearly different in some ways since the comparison group included students who were harder to reach during summer or were late applicants. But matching on a narrow bandwidth on the application scores limited this group nonequivalence, reducing selection bias through design details rather than analysis. As a within-study comparison, the major negative with Aiken et al. (1998) is that the analysts were not blind to the results of each type of study.

Bloom, Michalopoulos, and Hill (2005)

Bloom, Michalopoulos, and Hill (2005) evaluated the National Evaluation of Welfare-to-Work Strategies (NEWWS)—see also Bloom et al. (2002) and Michalopoulos, Bloom, & Hill (2004). The experiment took place in 11 sites. But one within-study comparison focused on five sites where a nonequivalent comparison group

was constructed from welfare enrollees who served as the control group in a randomized experiment conducted at a different job training center *in the same state at the same time*. Actually, in four of these cases the comparison units provided an even closer match, coming from the same city and not just the same state. The matches so achieved are highly focal in the sense that comparison group members are also enrolled as control cases at another job training study site. And the matches are intact since the analyses we report are based on all the original control group enrollees now serving as the comparison group for a different site in the same city. We make no attempt to match individuals here, though the authors did in some of their work. The study outcome was earnings from on-the-books employment scaled as two years of prior quarterly earnings plus five years of post-intervention quarterly earnings. Thus, two short interrupted time series were created, one for a counterfactual created at random and the other created systematically.

Figure 1 displays the means over time for the within-city contrasts, the intervention point being 0 on the time scale. Powerful statistical analyses reported in Table 6 of Michalopoulos, Bloom, and Hill (2004, p. 166) showed no reliable mean or slope differences in Oklahoma City (N controls = 831; N comparisons = 3,184), Detroit (N controls = 955; N comparisons = 1,187), or Riverside (N controls = 1,459; N comparisons = 1,501). Only in the smallest site, Portland (N controls = 328; N comparisons = 1,019), were the control and comparison groups reliably different. Pooling the four within-city or the five within-state data led to pretest means and slopes that did not differ between the control and comparison groups, implying once again that the careful selection of aggregated, nonequivalent comparisons sufficed to eliminate selection bias on very important observables. Since the post-intervention time series did not differ between the control and comparison groups in the three largest sites or the pooled analyses, it seems that any hidden bias due to unobserved initial differences would have had to be almost perfectly offset by bias from observables operating in the opposite direction. We surmise that this is more possible than plausible.

When the intervention and comparison sites were from different states or different cities in the same state (Grand Rapids and Detroit), selection bias was readily observed and never successfully adjusted away. Ten different kinds of analysis were used to see if the observed selection bias was reduced to zero after adjustments based on OLS, propensity scores, random growth models, and Heckman-type selection models. But no analysis recreated the experimental results, supporting the authors' general conclusion about the ineffectiveness of observational studies. However, this overlooks the analysis presented here, in which selecting a matched, within-city intact job training comparison group did reduce most of the bias, creating a salient exception to the authors' preferred conclusion.

Diaz and Handa (2006)

The authors compared experimental estimates from PROGRESA with observational study estimates from a nationally representative data set of Mexican households (ENIGH). Villages were eligible for PROGRESA if their average standing on a measure of material welfare was in the bottom quintile nationally; and within experimental villages, families were assigned to treatment based on their score on essentially the same material welfare scale. Because of fiscal constraints and logistical challenges, PROGRESA was launched in 31 states over 11 phases from 1997 to 2000. Thus some eligible villages were not part of the original study, and other villages were too affluent to be included, though some individual families in them might otherwise have been eligible. This sampling plan created two kinds of nonequivalent comparison villages. Sample 1 consists of all non-PROGRESA villages in the ENIGH survey, including those too affluent to qualify for PROGRESA; Sample 2 consists of villages that qualified for PROGRESA but had not yet received it. Sample 2 interests us here because of its higher probability of a closer match at the village level since

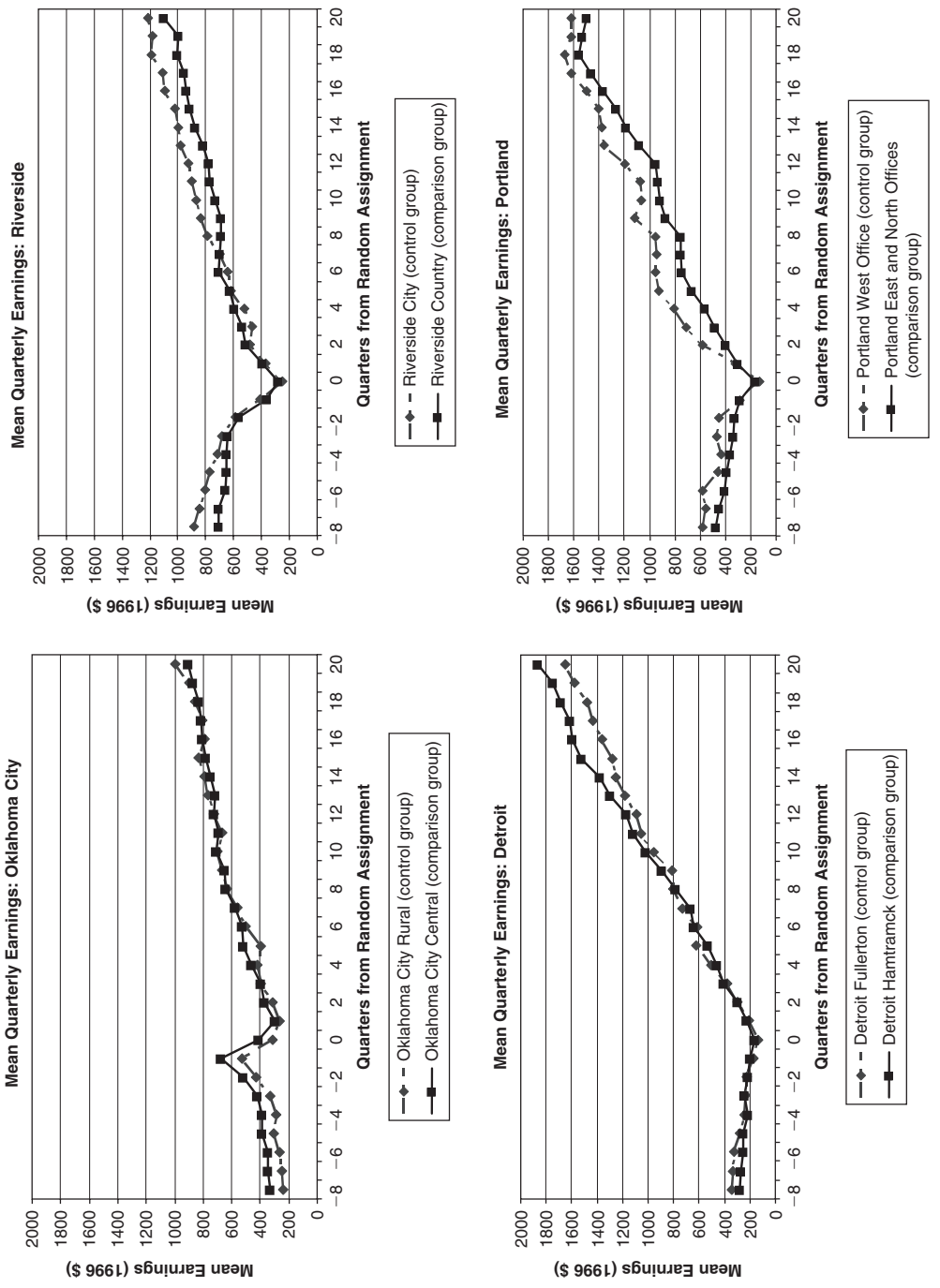

Figure 1. Quarterly Earnings for Sites Using Within-Site Controls (Michalopoulos, Bloom, & Hill, 2005).

all the treatment, control, and comparison villages were eligible for PROGRESA. So we ask: How comparable are the control and comparison families in Sample 2, given the prior matching for village level program eligibility?

Three outcomes were examined in Diaz and Handa, but we omit food expenditures here because it was measured differently in the PROGRESA study and the ENIGH comparison samples, thus confounding mode of treatment assignment and type of measure. But teen school dropout was measured the same way in both kinds of study and showed a reliable treatment effect in the experiment. Child labor was also assessed the same way but showed no effect in the experiment. Diaz and Handa (2006) tested how well OLS and several propensity score methods (caliper, kernel, and local linear) performed relative to the experiment.

The Sample 2 posttest means showed few differences between the control and comparison groups from the intact matched villages. For school dropout, 48 percent of the control children aged 13–16 were enrolled in school versus 51 percent of the comparisons, a nonreliable difference that persisted when various combinations of covariates were used. For child labor, both the control and comparison groups had 12 percent of the 12–16 year-olds working for pay when no covariates were used. Thus, the selection of intact matched villages produced similar profiles of the households in the experimental control and the quasi-experimental comparison groups on the variables observed. This is all the more striking because the two groups of villages in Sample 2 were not totally comparable. The authors' Table 1 shows that, relative to experimental controls, the comparison village households were on average smaller, older, better educated, and more materially advantaged. But none of these population differences operated strongly enough to provide different causal answers in the experiment and observational study either before or after they were used as covariates.

Nonetheless, these initial family differences that emerged despite matching the experimental and comparison villages remind us that matching intact groups does not guarantee total comparability. At best, it reduces population differences and diminishes the magnitude of the task that statistics are forced to play in adjusting for these differences. Even so, Diaz and Handa's (2006) sampling design for Sample 2 was sufficient to eliminate most of the bias from both observed and unobserved variables; and it did so without the extensive local matching in Aiken et al. (1998) and Bloom, Michalopoulos, and Hill (2005).

To summarize, we found three cases where researchers took pains to select intact comparison groups likely to overlap with the treatment group on pretest means and even slopes. In each case, comparable experimental and observational study effect estimates resulted, suggesting that any sources of bias remaining were quite minor in their influence. It may seem obvious that selection bias will be substantially reduced when a study's sampling design eliminates all or most group differences on the most important pretreatment observables. However, this point is not central in current scholarly discourse about observational studies. It focuses more on statistical adjustments for group nonequivalence than on sampling to minimize nonequivalence prior to, or instead of, individual case matching. The three studies also suggest that, though hidden bias could have played a major role in these exemplars, it did not in any of them. If it had, the control and comparison groups would have been consistently different at posttest; but they were not. Hidden bias is not logically ruled out, of course. But if it did operate, it had to have been obscured by an equal amount of unadjusted selection bias operating in the opposite direction. We leave readers to assess the plausibility of this across the three within-study comparisons examined here. To discover three concordant cases out of three is noteworthy but still too few to warrant a firm conclusion about the general absence of hidden bias after initial group matching. The point we want to emphasize, though, is the need to ask not just "How can we statistically control for population differences?" but also to ask, "How can we choose nonequivalent

Table 1. Bias reduction achieved by outcome and mode of analysis (Shadish, Clark, & Steiner, in press).

	Mean Difference (Standard Error)	Absolute Bias	Percent Bias Reduction (PBR)	R^2
Mathematics Outcome				
Covariate-adjusted randomized experiment	4.01 (.35)	0.00		0.58
Unadjusted quasi-experiment	5.01 (.55)	1.00		0.28
Propensity score (PS) stratification	3.72 (.57)	0.29	71%	0.29
Plus covariates	3.74 (.42)	0.27	73%	0.66
PS linear ANCOVA	3.64 (.46)	0.37	63%	0.34
Plus covariates	3.65 (.42)	0.36	64%	0.64
PS nonlinear ANCOVA	3.60 (.44)	0.41	59%	0.34
Plus covariates	3.67 (.42)	0.34	66%	0.63
PS weighting	3.67 (.71)	0.34	66%	0.16
Plus covariates	3.71 (.40)	0.30	70%	0.66
PS strat. w/predictors of convenience	4.84 (.51)	0.83	17%	0.28
Plus covariates	5.06 (.51)	1.05	−5%[a]	0.35
ANCOVA using observed covariates	3.85 (.44)	0.16	84%	0.63
Vocabulary Outcome				
Covariate-adjusted randomized experiment	8.25 (.37)			0.71
Unadjusted quasi-experiment	9.00 (.51)	0.75		0.60
PS stratification	8.15 (.62)	0.11	86%	0.55
Plus covariates	8.11 (.52)	0.15	80%	0.76
PS linear ANCOVA	8.07 (.49)	0.18	76%	0.62
Plus covariates	8.07 (.47)	0.18	76%	0.76
PS nonlinear ANCOVA	8.03 (.50)	0.21	72%	0.63
Plus covariates	8.03 (.48)	0.22	70%	0.77
PS weighting	8.22 (.66)	0.03	96%	0.54
Plus covariates	8.19 (.51)	0.07	91%	0.76
PS strat. w/predictors of convenience	8.77 (.48)	0.52	30%	0.62
Plus covariates	8.68 (.47)	0.43	43%	0.65
ANCOVA using observed covariates	8.21 (.43)	0.05	94%	0.76

Note: All estimates are based on regression analyses. For propensity score stratification stratum weights by propensity score quintiles were used. Standard errors for propensity score methods are based on 1,000 bootstrap samples (separate samples for each group), with refitted propensity scores and quintiles for each sample (predictors remained unchanged). Each model is presented with only the propensity scores used in the adjustment, and then with the same propensity score adjustment plus the addition of covariates based on backward stepwise inclusion (with main effects only).

[a] This adjustment increased bias by 5%.

populations with minimal initial differences, as intact group matching on high correlates of the outcome achieves?"

CASE III: CONTRASTING EXPERIMENTS AND OBSERVATIONAL STUDIES WITH MANIFESTLY DIFFERENT POPULATIONS BUT WHERE THE SELECTION PROCESS INTO TREATMENT IS KNOWN

Reliable pre-intervention differences often occur in observational studies, and we now describe six within-study comparisons contrasting an experiment with a nonexperiment where the treatment and comparison groups initially differ.

We begin by examining two of these studies where the selection into treatment was well known.

Diaz and Handa (2006)

The authors' Sample 1 includes villages too affluent to be eligible for PROGRESA. This sample is of interest because the comparison children clearly differed from the randomized control group. Thus, the percentage of poor Mexican children between 13 and 16 still in school was 58 percent versus 48 percent for the experimental controls. To test whether OLS and propensity score adjustments could reduce this bias, the authors were particularly fortunate because the same covariates were available for constructing the propensity scores that were also used to determine individual household eligibility for PROGRESA. So the selection process into treatment was known and could be well modeled using the same variables as in the experiment, though some would likely not be included in the propensity score model or would be weighted differently in it than in the experiment. In any event, a well-balanced propensity score model was possible that came close to mirroring the selection process into treatment. In this circumstance, the selection bias was reduced to essentially zero, the percentage leaving school being 48 percent in both the control and adjusted comparison groups. Obviously, the same causal estimate would ensue once the intervention group was added to the analysis.

To explore the limits of propensity score analysis, Diaz and Handa (2006) reduced their larger set of covariates to those they termed "off the shelf," those that a careless analyst might use because they are routinely available in surveys—respondent age and sex, schooling of household head, household composition, and whether the household receives social security. These predictors of convenience reduced little of the bias. Diaz and Handa's Sample 1 illustrates that extensive knowledge and careful measurement of the selection process into treatment can significantly reduce bias. The multivariate eligibility criterion for entry into PROGRESA were well known and the authors had access to the same variables for crafting bias adjustments to deal with the reality that households in the comparison villages were on average more affluent than in the treatment ones. However, the study also illustrates how poor selection models are when they include only demographic information.

Shadish, Clark, and Steiner (In Press)

Shadish, Clark, and Steiner (in press) studied how group coaching in either math or vocabulary affected subsequent performance in the domain coached. Their within-study comparison randomly assigned college students to being in a randomized experiment or an observational study on one of these two topics. All students then underwent the same experiences at the same time in the same setting, including responding to the same battery of measures at the same times in the same ways. For students exposed to math coaching, their post-intervention math scores served as the intervention-relevant outcome, and the math scores of those exposed to vocabulary training served as controls. The converse was true for those experiencing the vocabulary intervention. Thus, a replicated test was possible for describing the extent of any initial bias in the observational study and for indexing the extent of bias reduction due to covariates. The design of the within-study comparison in Shadish Clark, and Steiner (in press) has four arms instead of the usual three, and is depicted in Figure 2.

In the randomized experiment, the number of students did not differ by condition, nor did the pretest means. But 38 percent of the observational study students chose math and 62 percent vocabulary, and those selecting math had reliably higher pretest scores on the math test than did those choosing vocabulary. Conversely,

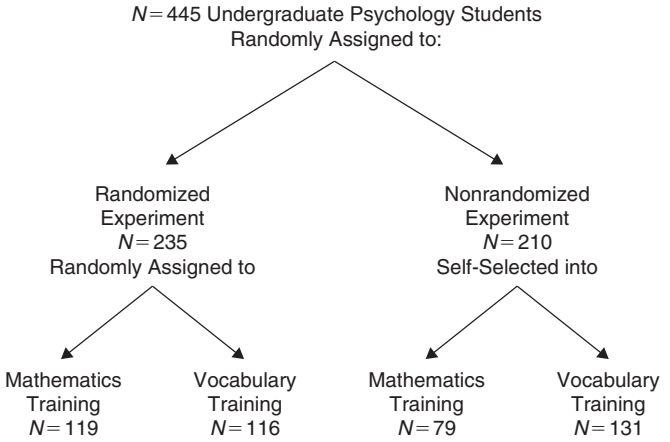

Figure 2. The Design of Shadish, Clark, and Steiner (in press).

students choosing vocabulary had reliably higher literacy pretest means than those choosing math. So the observational study involved demonstrably nonequivalent groups at pretest. The size of this bias is assessed as the difference between the unadjusted posttest difference between the experiment and the observational study. In the math condition, the mean was about 25 percent higher for the self-selected than the randomly selected students, while for vocabulary students the corresponding difference was 9 percent. The issue is: Will statistical adjustments eliminate such selection bias? Many pre-intervention measures were used to assess selection: (1) motivation to learn more about, or avoid, instruction in math and language arts; (2) general math and language arts proficiency as assessed from achievement tests; (3) past test scores and grades in math and language arts; (4) demographic details; and (5) personality assessment on a well-known national test (Goldberg, 1997). The study outcomes were content-valid scales specifically aligned with the math and vocabulary curriculum covered in the coaching treatments.

Table 1 shows the results from the OLS and propensity score analyses. Propensity scores were used as covariates, strata, or weights, once with and once without adding the individual covariates to the model. But the mode of propensity score analysis made little systematic difference, and OLS analyses did at least as well as propensity scores. Perhaps other forms of analysis would too. However, it is striking that, as in Diaz and Handa (2006), bias was not significantly reduced when just the demographic variables were used. These so-called predictors of convenience achieved excellent balance (namely, matching on the observables), but little bias reduction. Later analyses (Steiner et al., under review) showed that motivation to expose oneself to math or language arts was the most important single covariate, particularly for the math effect. Next most important were the pretest measures of general math and language arts achievement. The implication of these and other data is that the self-selection process was largely driven by individuals choosing to be coached in the subject matter of greatest proficiency, with the result that measuring this process led to significant bias reduction. Therefore, the quality of the covariates was more important in reducing bias than data analysis, however it was conducted.

Shadish, Clark, and Steiner (in press) was an unusually stringent within-study comparison test. Because individuals were randomly assigned to the experiment or observational study and underwent the same testing procedures, the control and comparison groups did not differ on third variables correlated with outcome.

The short-term laboratory setting allowed researchers to control the random assignment process and ensure that no attrition or treatment crossovers occurred, creating an experiment that could serve as a valid causal benchmark. A quality observational study was developed because individuals were contrasted who did or did not self-select into a particular kind of coaching and who otherwise overlapped in attending the same local university, taking the same psychology course, and volunteering for the same experiment. Most important, the authors theorized that their study selection process depended on an individual's interests and cognitive strengths and used psychometrically sound measures to assess them. The observational study analysis required computing adjusted observational study effect sizes via OLS and propensity scores that met the balance criteria in Rubin (2001), and were carried by an author blind to how the propensity scores he computed affected the bias reduction achieved in a particular outcome model. The study's main limitations follow from its laboratory setting and short-lasting intervention. They reduce the ability to extrapolate the study results to research practice in public policy that is characterized by longer treatments and field settings.

Shadish, Clark, and Steiner (in press) is different from traditional within-study comparisons in two salient ways. One is the fourth arm in their design—the treatment group in the quasi-experiment. This means that the experimental and observational study groups are similarly composed except for the selection process in the observational study, the focus of theoretical interest. So selection is not confounded with irrelevant population differences, as it usually is in within-study comparisons with only three treatment arms. Second, in the traditional three-arm design, cause in the quasi-experiment is usually estimated at the mean of the randomized treatment group. The estimator is thus a conditional average treatment effect for the treated. The four-group design offers more flexibility, and Shadish, Clark, and Steiner estimated their quasi-experimental effect at the mean of the combined treatment and control groups in the quasi-experiment. Given random assignment to the experiment or observational study, this target population is identical to the mean in the randomized experiment. Thus, their estimator was the unconditional average treatment effect for the study population—a population more general than the treated one. This made little substantive difference, though, for when the data were reanalyzed using the traditional target group instead of the broader one, the experimental and quasi-experimental estimates were again similar.

To summarize: Diaz and Handa represents a clear case of the selection model in the treatment group being essentially recreated for the nonequivalent comparison group and resulting in the reduction of nearly all selection bias. In Shadish, Clark, and Steiner, knowledge of the selection process is somewhat less clear, being dependent on general cultural "wisdom" and some empirical evidence presented in Steiner et al. (under review), which shows that students self-selected themselves into mathematics or vocabulary instruction based largely on their fear of mathematics or their positive attraction to it. Given high-quality measurement of these motivational forces, it became possible to model this central component of the selection process and thereby to reduce the bias to close to zero. Knowledge of the selection process can significantly reduce selection bias provided the selection process is valid and reliably measured.

MORE AMBIGUOUS WITHIN-STUDY COMPARISON CASES

Hill, Reiter, and Zanutto (2004)

Hill, Reiter, and Zanutto (2004) used the Infant Health Department Program (IHDP) as their experiment. IHDP began in 1985 and provided low birth weight infants with weekly visits from specialists, daily child care services, and transportation to and from care centers. In the experiment, IHDP increased children's

scores on the Peabody Picture Vocabulary Test (PPVT) by 6.39 units (s.e. = 1.17) at ages 3 and 4. The observational study was constructed from births in the National Longitudinal Survey of Youth (NLSY) between 1981 and 1989 (N = 4,511). The NLSY and IHDP populations differed markedly in terms of birth weight, days in hospital, and weeks preterm. OLS and propensity scores were used to see if this difference could be eliminated. Since the selection of covariates is important, four models were used to estimate the observational study effects. Model 1 controlled for sociodemographic variables only; Model 2 added the unemployment rate of the county where the infant resides; Model 3 added controls for the state the infant was born in; and Model 4 controlled for sociodemographic factors and county-level unemployment rates, with treated infants being matched only to NLSY children living in the same state.

After adjustment, the sample means were much closer for all four models. By conventional statistical significance criteria, all of the OLS and propensity score tests would have correctly concluded that IHDP was helpful. But only the third model brought the estimate close to the experimental one (6.20 versus 6.39), and then only in one form of propensity score analysis. Other models and modes of analysis overestimated the treatment effect—by about 5 points for OLS, by about 4 points for propensity scores based only on demographics, and 2 points when matching took place within states. So Hill, Reiter, and Zanutto (2004) is a glass half full or half empty. The correct causal direction would always have been identified, but an obviously close correspondence with the experiment depended on state matching variables being in the analysis and on a single type of regression run—a tenuous set of contingencies that cries out for replication. However, Hill, Reiter, and Zanutto (2004) suffers as a within-study comparison. It tests the efficacy of individual case-matching using a national dataset to select the matching cases. But we cannot be sure that PPVT was administered in the same way in the experiment and NLSY survey. The timing of the testing varies also, though Hill, Reiter, and Zanutto (2004) did sample NLSY children symmetrically around the 1985 IHDP date. Another confound is that experimental, OLS, and propensity score methods estimated different causal quantities, as the authors themselves pointed out. All in all, we are not inclined to assign much weight to the contingent and not yet replicated findings of Hill, Reiter, and Zanutto (2004).

McKenzie, Gibson, and Stillman (2007)

McKenzie, Gibson, and Stillman (2007) took advantage of a lottery that allowed winners to emigrate from Tonga to New Zealand. Pretest means did not differ between the 120 winners and 78 losers. But 55 winners did not take up the offer to leave, and so a TOT estimate was also computed using the lottery as an IV (Angrist, Imbens, & Rubin, 1996). The main causal benchmark used was the TOT estimate of the effect of actual migration rather than of being offered the chance to migrate. In the experiment, the ITT effect was of NZ$91 and the reliable TOT estimate was NZ$274.

Two comparison groups were used. One drew from 60 non-applicant households living in the same village as lottery applicants, and the sole time-varying covariate collected on them was a retrospective self-report of income, though data on employment, income, and demography were also available and used. The second comparison group was drawn from the 2003 Tongan Labor Force Survey (TLFS) of 8,299 individuals, of whom 2,979 were in the target age range of 18–45. It contains information on income, age, sex, marital status, education, and birthplace but not about past income. The authors modeled the selection process for lottery applicants by comparing the pre-migration retrospective earnings of lottery winners and local comparison group members, discovering that lottery applicants earned NZ$56 more per week prior to migration. Five approaches were made to reduce this bias:

(1) a simple difference between migrants' pre- and post-migration incomes; (2) an OLS analysis with various covariates; (3) a difference-in-difference regression which subtracted out prior income; (4) several types of propensity score matches that used all the covariates, including retrospective income; and (5) an IV analysis using log distance to an office of the New Zealand Immigration Service as the instrument.

This distance IV performed best, and the difference-in-difference estimator next best. Once again the worst estimate came from an OLS model that used individual background variables and no prior earnings. The TLFS sample was not used in the analyses that produced comparable estimates because past income records were not available on these survey respondents. Also, there were only 60 cases in the local comparison group. Moreover, extensive use was made of a retrospective income measure that might depend on whether a Tongan had chosen to emigrate and so was already earning at the higher New Zealand rate, thus affecting how he calibrated past earnings. For all these reasons we assign little weight to McKenzie, Gibson, and Stillman (2007) as a within-study comparison, while acknowledging that most selection bias seems to have been removed with a set of covariates that seems weaker than in other studies.

Agodini and Dynarski (2004)

The authors examined whether a prevention program for students at risk for dropping out (called SDDAP) would affect dropouts, absenteeism, self-esteem, and motivation to graduate from high school. In the experiment, random assignment took place within 16 middle and high schools. In the observational study, students in these schools were individually matched via propensity scores using two data sources. One was the National Educational Longitudinal Study (NELS); the other was composed of students attending four different schools simultaneously participating in a quasi-experimental study of school restructuring that used the same measures as in SDDAP. However, the four comparison schools came from different states than the 16 treatment and control schools, and, while the experimental students were from every middle and high school grade, the comparison school students were just 7th graders in two middle schools, 9th graders in one high school, and 10th graders in another high school. The analytic plan was to generate 128 different propensity score analyses by combining two comparison data sets, four outcomes, and 16 school-specific contrasts of experimental and observational study results.

The number of covariates for computing propensity scores varied—potentially 20 for the SDDAP comparison and 13 for the NELS one. Both sets of covariates included demographic measures, but only the SDDAP data had pretest information on all four outcomes, though pretest reading and math scores were only spottily available. Acceptable balance was attained in only 29 of the 128 planned analyses, perhaps because the treated and matched students came from different states or were distributed so differently across middle and high school grades. However, where successful balance was achieved, Agodini and Dynarski (2004) concluded that the experimental and observational study findings did not agree.

But this was not a good within-study comparison. The authors' Table 5 shows reliable pre-intervention differences between the intervention and control groups at two experimental sites, casting doubt on the validity of the experimental benchmark. Also, third variable confounds are rampant because local experimental data are compared with national NELS data, resulting in different population characteristics, local history, setting, and measurement. Matches are better in the SDDAP comparison, but still not good since they come from different states with their own policies about dropping out and absenteeism. Sample sizes are also an

issue. Propensity score analysis is a large-sample technique (Luellen, 2007). But for the better SDDAP comparisons, the combined treatment and comparison samples are 390, 200, 162, 140, and 113. This is especially problematic in balance tests. The needed within-strata tests of individual covariates will entail small samples and the need to accept the null hypothesis in order to move on in the analysis. The covariates themselves also seem limited. Prior reading and math achievement tests are generally not available; nor are pretest measures of outcomes for the NELS sample. The quality of covariates would be less of a problem if the assignment process were simple, but as the authors describe it on page 182, it is based on a number of student and school staff factors that varied by site and was overall quite complex.

The main problem is that Agodini and Dynarski (2004) compare an experiment that was conducted only intermittently well across sites to a poor-quality observational study that was based on data sources of convenience. No effort was made to implement a well-designed quasi-experiment from scratch. Yet this could have been done by, for example, matching each treatment school with one or more comparison schools from the same school district over several prior years of actual dropout in the case of high schools or on a composite of demographic and past predictors of dropout in the case of middle schools. Many other options are available, of course. But to design a nonexperiment that uses only data conveniently on hand is likely to entail comparing an experiment to a weak observational study. While such a strategy might be appropriate for testing the effectiveness of a commonly used but suboptimal analytic practice, it constitutes a weak test of the potential of better quasi-experimental practice.

Wilde and Hollister (2007)

The Tennessee Project Star experiment randomly assigned kindergarten classes to regular or reduced class sizes in order to see how class size affects achievement. Wilde and Hollister's (2007) reanalysis of the data was limited to 11 of the 80 study schools, those where the total kindergarten enrollment exceeded 100. This size restriction allowed 11 separate experimental and nonexperimental estimates to be contrasted. Sample sizes in the treatment classes varied from 24 to 56 students per school with a median of 29—two classes. The number of randomly selected control classrooms is similar, though inevitably each has about 30 percent more students. Reliable benefits of smaller classes were observed in 7 of the 11 experiments and in a pooled analysis across all 11.

To construct the observational study, individual students in the treatment classes were matched to students from comparison classes in all the other study schools across the state—not just to students in the 10 other schools with large kindergartens or to students attending schools in the same district as intervention students. Propensity scores were calculated from data on (1) the student's sex, race, birth year and quarter, and free lunch status; (2) teacher's education, rung on the career ladder, and years of experience teaching; and (3) community type (urban, rural, inner city, or suburban). Strikingly absent are pretest achievement measures assessed at the student, classroom, or school level. Moreover, the authors' Appendix Table A reveals many small sample sizes within the various treatment group strata. Of the 46 strata outlined, 16 have 4 cases or fewer, 7 have samples 5 to 10, and only 13 have samples exceeding 10. Of these 46, 24 show reliable within-strata treatment/comparison differences between the 0.05 and 0.10 level. This is far more than chance and suggests that balance may not have been achieved in the propensity score analysis. Moreover, the outcome tests were also of lower power, based on an average of only 29 treatment students in each of 11 experiments. This means that in a propensity score analysis of 5 strata there will be only about 6 students per stratum;

and these 6 will not be evenly distributed across the strata, given that the treatment and comparison populations are different.

Wilde and Hollister (2007) discovered that 8 of their 11 differences in experimental and observational study effect sizes were reliable, and so they concluded that the observational studies failed to reduce the selection bias. We are less sure because of the small samples in each experiment, the general paucity of covariates, the absence of theory about presumed selection processes, the absence of pretest achievement measures at the child, classroom, or school levels, and the failure to match intact schools early in the analysis. But Wilde and Hollister (2007) also pooled their data, thus reducing the sample size problem due to analyzing single-school data. They discovered a reliable effect of 12.72 percentile points for the average experiment and 17.79 for the average nonexperiment. Each of these estimates reliably differed from zero, indicating that lower class sizes increased achievement. But they also reliably differed from each other, indicating that they disagree as to the size of the advantage. The authors then forcefully argue that, for public policy, the size of the difference is more important than its reliability.

Though the pooled analyses escape the sample size limitations of the site-specific experiments, they do not escape the limitations of covariate quality. They also raise three fundamental issues. First, in multisite clinical trials in medicine, the usual advice (Brookes et al., 2001) is not to analyze each site separately. This is for fear of capitalizing on chance and because the usual policy need is to identify interventions that are robustly effective despite site variation since it is not easy to tailor individual policies to individual sites. Such considerations lead to preferring the pooled results over the site-specific ones. Second, it is not always trivial for policymakers to learn that an experiment and observational study agree on the reliability and direction of an effect but not on its magnitude. We are not sure that policy actions should always depend on the size of an effect rather than on its direction and reliability. And third, the authors used extant data to conduct an observational study with nonequivalent groups that lacked pretest information on the same scale as the outcome. This nonequivalent comparison group design that results is of a kind that Cook and Campbell (1979) labeled as "generally (causally) uninterpretable."

If one were to design an observational study on class size from scratch, it would look quite different from Wilde and Hollister (2007). One would not take within-school contrasts seriously when the sample sizes were so low. One would seek to collect individual pretest achievement scores from kindergarten children at the beginning of the school year. One might also seek to collect multiple prior years of achievement data on the classes taught by the same teachers in the treatment and comparison groups. One would also look to select comparison classes from schools that were matched with the intervention ones on past years' performance, seeking to contrast intact treatment with intact comparison classes from maximally similar schools. And if individual student matching were necessary, one would collect voluminous data on the students, teachers, families, and neighborhoods so as to construct propensity scores that are more likely to be highly correlated with both selection and outcome, as Hong and Raudenbush (2006) did in their early childhood education study. Even if we had to accept the limitations on constructing comparison groups imposed by Wilde and Hollister, for comparison purposes it would still probably have been better to use the best matched classes that served as controls at other experimental sites instead of immediately jumping to matching at the student level. In our view, Wilde and Hollister compared an adequate pooled experiment to an inadequately designed quasi-experiment. Like Agodini and Dynarksi (2004), they have shown that propensity scores cannot compensate for bad quasi-experimental design. They have not shown that observational studies at their current best fail to reproduce the results of experiments.

CONCLUSIONS

Past within-study comparisons from job training have been widely interpreted as indicating that observational studies fail to reproduce the results of experiments. Of the 12 recent within-study comparisons reviewed here from 10 different research projects, only two dealt with job training. Yet eight of the comparisons produced observational study results that are reasonably close to those of their yoked experiment, and two obtained a close correspondence in some analyses but not others. Only two studies claimed different findings in the experiment and observational study, each involving a particularly weak observational study. Taken as a whole, then, the strong but still imperfect correspondence in causal findings reported here contradicts the monolithic pessimism emerging from past reviews of the within-study comparison literature.

The starting point for this paper was identifying three different types of nonequivalent group design that privilege the question "Under which conditions do experiments and observational studies produce reasonably similar results?" over the traditional (and more coarse-grained) question, "Do experiments and observational studies produce similar findings?" RD is one type of nonequivalent group design, and three studies showed that it produced generally the same causal estimates as experiments. This is not theoretically surprising. More surprising is that the experiments and RD studies produced essentially the same statistical significance patterns despite RD's lower efficiency. But two of the studies in question had larger sample sizes for the RD component than the experiment, and the other had over 5,000 cases in each type of design. The basic conclusion, though, is that RD estimates are valid if they result from analyses sensitive to the method's main assumptions.

We can also trust estimates from observational studies that match intact treatment and comparison groups on at least pretest measures of outcome. Explicit here is the use of a study's sampling design to minimize initial differences between the control and comparison populations. The possibility of hidden bias from unobservables still exists, of course. However, in the three instances analyzed here (Aiken et al., 1998; Bloom, Michalopoulos, & Hill, 2005; Diaz & Handa, 2006), such hidden bias either failed to operate or, less plausibly, was suppressed by selection bias of comparable magnitude operating in the opposite direction across all three examples. Diaz and Handa's Sample 2 was not completely successful in matching the intact set of households within program-eligible villages, perhaps because the village matches were not on geographical proximity as well. But although Sample 2 revealed some initial pretest differences between experimental control and comparison households on background variables, no bias was evident in the final analyses after controlling for selection. Intact group matching on pretest values seems to be a boon, probably more so when it also involves matching for geographic proximity. And it is a boon even when the match is not perfect, for it reduces the initial selection bias that subsequent individual case matching has to deal with.

Sometimes, though, researchers cannot choose intact local comparison groups and have to try to make groups equivalent, not via a study's sampling design, but via statistical procedures such as OLS, propensity scores, or instrumental variables. In within-study comparisons in job training, these methods rarely produced unbiased estimates. So why did they sometimes do so here? Two within-study comparisons (Shadish Clark, & Steiner, in press; Diaz & Handa, 2006) strongly suggest that knowing and measuring the correct selection process reduces bias. For the analysis of their Sample 1, Diaz and Handa took advantage of a fortunate situation in which the process of allocation to treatment was explicit and where the very same measures of this process were available for selecting nonexperimental comparison cases. The selection process in Shadish, Clark, and Steiner (in press) was less explicitly linked to an externally known selection process, but it seems reasonable to assume

that persons chose exposure to instruction in math over literature because they either liked math more or feared it less (see Steiner et al., under review). Measuring these motivational processes also reduced most bias. The importance of knowing the selection process is not theoretically surprising. It is, after all, the key feature of both random assignment and RD.

In most research situations the selection process is more opaque than in the two examples above, and measures of what it might be are sometimes limited in availability and quality. The results reported here indicate that statistical adjustments for selection are less useful (1) when comparison cases are selected from national data sets that differ from the intervention group in population, settings, and measurement; (2) when sample sizes in the control or comparison conditions are modest; and (3) when only demographic variables are available as covariates. One implication of these conditions is that reducing bias to a meaningful degree will be difficult when comparison groups are formed from extant survey databases. Surveys are rarely planned with a specific local selection process in mind, thereby impoverishing the quality of available covariates. A second implication is that nonexperiments will generally be less successful in reproducing experimental estimates than will prospective quasi-experiments. Quasi-experiments permit researchers to collect their own data, preferably utilizing experimental design features that were deliberately developed in the past to rule out specific alternative causal interpretations. Quasi-experimental researchers are also more likely to be able to develop and measure context-specific theories about selection processes. And they tend to value initial group matching on high correlates of the outcome, especially pretest measures of the outcome rather than proxies for it. Indeed, this review showed that use of "off-the-shelf" (mostly demographic) covariates consistently failed to reproduce the results of experiments. Yet such variables are often all that is available from survey data. They are also, alas, all that some analysts think they need.

The failure of "off-the-shelf" covariates indicts much current causal practice in the social sciences, where researchers finish up doing propensity score and OLS analyses of what are poor quasi-experimental designs and impoverished sets of covariates. Such designs and analyses do not undermine OLS or propensity scores per se—only poor practice in using these tools. Nonetheless, we are skeptical that statistical adjustments can play a useful role when population differences are large and quasi-experimental designs are weak, as happened with Agodini and Dynarski (2004) and Wilde and Hollister (2007). This means we are skeptical about much current practice in sociology, economics, and political science, where the substantive issues often preclude random assignment, RD, intact group matching, or even informed theories of selection that are specific to the causal issue and context under study. In this predicament, researchers seem to put their effort into substantive causal modeling and IV analyses of nonexperiments, including natural experiments, though these all too often inadvertently pyramid assumptions that have unclear links to the real world and to valid data about this world. As a result, too little care goes into prospective quasi-experimental design, theory-based covariate selection, and even the use of very many covariates in the shotgun fashion prevalent among users of propensity scores.

We turn now from nonexperimental to quasi-experimental practice. The fact that some kinds of quasi-experiments robustly reproduce experimental results does not imply that all or most quasi-experiments will routinely do so. The evidence we have adduced is relevant to only a narrow slice of all quasi-experimental practice. Future within-study comparison work is needed comparing experimental results to a broader range of quasi-experimental designs and practices, including sibling and cohort designs, interrupted time series with control series, and designs in which the intervention is introduced at different times into different groups. The generally positive findings reported here reflect a circumscribed set from among the larger repertoire of quasi-experimental practices.

So what might policymakers make of these findings? They do not undermine the superiority of random assignment studies where they are feasible. They are better than any alternative considered here if the only criterion for judging studies is the clarity of causal inference. But if other criteria are invoked, the situation becomes murkier. The current paper reduces the extent to which random assignment experiments are superior to certain classes of quasi-experiments, though not necessarily to all types of quasi-experiments or nonexperiments. Thus, if a feasible quasi-experiment were superior in, say, the persons, settings, or times targeted, then this might argue for conducting a quasi-experiment over an experiment, deliberately trading off a small degree of freedom from bias against some estimated improvement in generalization. For policymakers in research-sponsoring institutions that currently prefer random assignment, this is a concession that might open up the floodgates to low-quality causal research if the carefully circumscribed types of quasi-experiments investigated here were overgeneralized to include all quasi-experiments or nonexperiments. Researchers might then believe that "quasi-experiments are as good as experiments" and propose causal studies that are unnecessarily weak. But that is not what the current paper has demonstrated. Such a consequence is neither theoretically nor empirically true but could be a consequence of overgeneralizing this paper.

It is striking that past within-study comparisons in job training consistently failed to reproduce experimental results, while such comparisons had the opposite consequence in this review. Why? One possibility is that selection mechanisms are more complex in job training. Individuals go for job training for many reasons—because of court orders, time limits on welfare benefits, social worker suggestions, or advice and even hectoring from partners. They may also attend because a friend is receiving training at the same site, or the site is easily accessible, or rumors have been heard about the site's effectiveness. Individuals may also go because they are bored at home, depressed, ready to seize a last chance, or want to impress immigration officials. Individuals can attend training for many different mixes among all these single reasons. Even so, Heckman and his colleagues (Heckman & Smith, 1999; Heckman, LaLonde, & Smith, 1999; Heckman et al., 1997) have speculated that selection processes could be well measured in job training if there were information to make local case matches in the same labor market; if employment and wage data were available and measured in the same way for comparison, treatment, and control group members; and if information on dynamic prior labor market experiences were available, especially transitions from employment to unemployment over eight quarters prior to job training. Measures like these might or might not work, but they were not available or utilized in the original job-training within-study comparisons. In contrast, the selection process of Shadish, Clark, and Steiner (in press) was simpler. Individuals can prefer math or vocabulary coaching for several different reasons—out of a liking for one subject matter or avoidance of the other, because they want to better themselves, or because they are indifferent but still have to pick. But as complex as these reasons are, they pale beside the more numerous and complex selection processes in job training.

Another difference between the current studies and past ones from job training relates to the correlation of covariates with outcome. In most of the examples reviewed in this paper, the selection model was more highly correlated with outcome than is typical in job training. In education, for example, individual achievement scores are very highly correlated over most annual pretest and posttest time points. But this is not the case with annual or quarterly income data, at least not for populations undergoing job training. Although we stressed understanding the selection process here, we should not forget that bias is reduced only by those selection variables that are correlated with outcome.

A third reason for the current greater optimism may be the most crucial. It pertains to being specific about kinds of observational study. We emphasized three here, that from theory, are most likely to reproduce the results of experiments.

We suspect that few methodologically sophisticated scholars will quibble with the claim that reasonably well-executed experiments and RD studies will often produce similar causal estimates. Nor will they disagree with the claim that studies predicated on careful intact group matching on pretest measures of outcome will considerably reduce selection bias. Most will not even quibble with the notion that understanding, validating, and measuring the selection process will substantially reduce the bias associated with populations that are demonstrably nonequivalent at pretest. In contrast, the within-study comparisons from job training did not rely as much on theory of method to predict when an experiment and observational study are more likely to generate similar causal results. In the job training work, the quasi-experimental design structures were heterogeneous in form and underexplicated relative to the emphasis the researchers placed on statistical models and analytic details. It is as though the studies' main purpose was to test the adequacy of whatever nonexperimental statistical practice for selection bias adjustment seemed current in job training at the time. This is quite different from trying to test best possible quasi-experimental design and analysis practice, as we have done here.

We may never definitively understand why the types of observational study explored here often reproduced experimental results while within-study comparisons in job training did not. But the simplicity of selection models, the correlation of selection variables with outcome, and the theoretical specificity with which subtypes of observational study designs were explicated may provide some guidance. We must also not forget that the within-study comparisons in job training predated the studies reviewed here, and so the criteria for conducting a quality within-study comparison were less developed than in this review, adding even more uncertainty to the knowledge claims emanating from earlier within-study comparisons.

THOMAS D. COOK is Professor of Sociology, Northwestern University.

WILLIAM R. SHADISH is Professor of Psychology, University of California, Merced.

VIVIAN C. WONG is a Graduate Student at Northwestern University.

ACKNOWLEDGMENTS

We thank Howard Bloom, Mark Dynarski, Larry Hedges, Donald Rubin, Jeffrey Smith, and Peter Steiner for their perceptive feedback. For additional comments, we also thank attendees at a French Econometric Society conference, at the MIT Labor and Developmental Economics Workshop, at the Harvard Applied Statistics Workshop, at the Northwestern Q-Center, and at colloquia at Arizona State and Florida State universities. The work was funded by a grant from the Institute for Educational Sciences of the U.S. Department of Education.

REFERENCES

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. Review of Economics and Statistics, 86, 180–194.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. Evaluation Review, 22, 207–244.

Angrist, J., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444–472.

Black, D., Galdo, J., & Smith, J. C. (2005). Evaluating the regression discontinuity design using experimental data [electronic version]. Working paper from http://www.personal.ceu.hu/departs/personal/Gabor_Kezdi/Program-Evaluation/Black-Galdo-Smith-2005-RegressionDiscontinuity.pdf.

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), Learning more from social experiments (pp. 173–235). New York: Russell Sage Foundation.

Bloom, H. S., Michalopoulos, C., Hill, C. J., & Lei, Y. (2002). Can nonexperimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? Washington, DC: Manpower Demonstration Research Corporation.

Brookes, S. T., Whitley, E., Peters, T. J., Mulheran, P. A., Egger, M., & Davey Smith, G. (2001). Subgroup analyses in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. Health Technology Assessment, 5, 1–56.

Buddelmeyer, H., & Skoufias, E. (2003). An evaluation of the performance of regression discontinuity design on PROGRESA. Bonn, Germany: IZA.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics, and economics. Journal of Econometrics, 142, 636–654.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis for field settings. Chicago: Rand McNally.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94, 1053–1062.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator. Journal of Human Resources, 41, 319–345.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Preventive Medicine, 15, 451–474.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources, 22, 194–227.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. Annals of the American Academy, 589, 63–93.

Goldberg, L. R. (1997). Big-five factor markers derived from the IPIP item pool (short scales). International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences [online]. Available at http://ipip.ori.org/ipip/appendixa.htm#AppendixA.

Goldberger, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations. Madison, WI: Institute for Research on Poverty.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69, 201–209.

Heckman, J., Ichimura, H., Smith, J. C., & Todd, P. (1998). Characterizing selection bias. Econometrica, 66, 1017–1098.

Heckman, J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. Review of Economic Studies, 64, 605–654.

Heckman, J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. Review of Economic Studies, 65, 261–294.

Heckman, J., LaLonde, R., & Smith, J. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), Handbook of labor economics, Vol. 4 (pp. 1865–2073). Amsterdam: Elsevier Science.

Heckman, J., & Smith, J. A. (1995). Assessing the case for social experiments. Journal of Economic Perspectives, 9, 85–110.

Heckman, J., & Smith, J. A. (1999). The pre-program earnings dip and the determinants of participation in a social program: Implication for simple program evaluation strategies. Economic Journal, 109, 313–348.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X. L. Meng (Eds.), Applied Bayesian and causal inference from an incomplete data perspective (pp. 49–60). New York: Wiley.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. Journal of the American Statistical Association, 101, 901–910.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. American Economic Review, 76, 604–620.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. American Psychologist, 48, 1181–1209.

Ludwig, J. O., Duncan, G. J., & Ladd, H. F. (2003). The effects of MTO on children and parents in Baltimore. In J. Goering & J. D. Feins (Eds.), Choosing a better life? Evaluating the moving to opportunity social experiment (pp. 153–176). Washington, DC: Urban Institute Press.

Luellen, J. K. (2007). A comparison of prospensity score estimation and adjustment methods on simulated data. Memphis, TN: The University of Memphis.

McKenzie, D., Gibson, J., & Stillman, S. (2007). How important is selection? Experimental versus nonexperimental measures of income gains from migration. Washington, DC: World Bank.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? Review of Economics and Statistics, 86, 156–179.

Orr, L. L., Bloom, H. S., Bell, S. H., Doolittle, F., Lin, W., & Cave, G. (1996). Does training for the disadvantaged work? Evidence from the national JTPA study. Washington, DC: Urban Institute Press.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Services and Outcomes Research Methodology, 2, 169–188.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (in press). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. Journal of American Statistical Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Company.

Smith, J. C., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. Journal of Econometrics, 125, 305–353.

Steiner, P. M., Cook, T. D., Shadish, W.R., & Clark, M. H. (under review). The nonignorability of strongly ignorable treatment assignment: Covariates that effectively control for selection bias in observational studies.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. Journal of Educational Psychology, 51, 309–317.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. Journal of Policy Analysis and Management, 26, 455–477.

Zellner, A., & Rossi, P. (1986). Evaluating the methodology of social experiments. In A. Munnell (Ed.), Lessons from the Income Maintenance Experiments (pp. 131–157). Boston: Federal Reserve Bank of Boston.