Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

How Bias Reduction Is Affected by Covariate Choice, Unreliability, and Mode of Data Analysis: Results From Two Types of Within-Study Comparisons

Thomas D. Cook [a]; Peter M. Steiner [a]; Steffi Pohl [b]
[a] Institute for Policy Research, Northwestern University, [b] Friedrich-Schiller-Universität, Jena, Germany

Online publication date: 11 December 2009

To cite this Article Cook, Thomas D., Steiner, Peter M. and Pohl, Steffi(2009) 'How Bias Reduction Is Affected by Covariate Choice, Unreliability, and Mode of Data Analysis: Results From Two Types of Within-Study Comparisons', Multivariate Behavioral Research, 44: 6, 828 — 847
To link to this Article: DOI: 10.1080/00273170903333673
URL: http://dx.doi.org/10.1080/00273170903333673

Multivariate Behavioral Research, 44:828–847, 2009

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170903333673

Psychology Press
Taylor & Francis Group

2008 Saul B. Sells Award Address Paper

How Bias Reduction Is Affected by Covariate Choice, Unreliability, and Mode of Data Analysis: Results From Two Types of Within-Study Comparisons

Thomas D. Cook and Peter M. Steiner
Institute for Policy Research
Northwestern University

Steffi Pohl
Friedrich-Schiller-Universität, Jena, Germany

This study uses within-study comparisons to assess the relative importance of covariate choice, unreliability in the measurement of these covariates, and whether regression or various forms of propensity score analysis are used to analyze the outcome data. Two of the within-study comparisons are of the four-arm type, and many more are of the three-arm type. To examine unreliability, simulations of differences in reliability are deliberately introduced into the 2 four-arm studies. Results are similar across the samples of studies reviewed with their wide range of non-experimental designs and topic areas. Covariate choice counts most, unreliability next most, and the mode of data analysis hardly matters at all. Unreliability

Correspondence concerning this article should be addressed to Thomas D. Cook, Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208. E-mail: t-cook@northwestern.edu

> has larger effects the more important a covariate is for bias reduction, but even so the very best covariates measured with a reliability of only .60 still do better than substantively poor covariates that are measured perfectly. Why regression methods do as well as propensity score methods used in several different ways is a mystery still because, in theory, propensity scores would seem to have a distinct advantage in many practical applications, especially those where functional forms are in doubt.

Many causal studies use quasi-experimental or observational study methods whose interpretation is bedeviled by the possibility of selection bias—namely, population differences masquerading as treatment effects. It has long been known that the effectiveness of statistical controls for selection depend on two factors (Cronbach, 1982). One is the particular covariate constructs chosen to model the outcome, and this source of potential bias has been variously labeled as specification bias, omitted or hidden variable bias, or failure to meet the strong ignorability assumption (Rosenbaum, 2002; Rosenbaum & Rubin, 1983). The second source of bias stems from unreliability in the covariate constructs measured, variously called errors in covariates, pretest unreliability, and, more generally, errors in variables. Irrespective of the covariates chosen and their reliability, bias can also arise from how the outcome is modeled analytically. For example, regression-based methods have come under attack from propensity score methods (PS) whose assumptions are different and generally more flexible (e.g., Stuart & Rubin, 2007).

Using multiple data sources, this article presents analyses in order to test two things: the relative importance of each of these ways to reduce selection bias and the generality of their relative importance across a set of quite mixed substantive circumstances. These purposes require a valid causal benchmark against which to estimate the magnitude of any initial bias and to assess how much of this bias is then reduced by the covariates chosen, their unreliability, and how they are used in adjusting for selection bias. To achieve this benchmark, we rely on evidence from "within-study comparisons."

Such studies compare the effect size from a randomized experiment with the effect size from an observational study that has been statistically adjusted to try to rule out selection bias. If all the bias is eliminated, then the effect size in the experiment and adjusted quasi-experiment should be identical. There are two traditions for conducting such studies. In the first and earliest (e.g., LaLonde, 1986) there are three treatment arms: the randomly formed treatment group, the randomly formed control group, and a nonrandomly formed comparison group. Thus, the same treatment group is used with both kinds of counterfactual, either randomly formed or not. In the second tradition, respondents are randomly assigned to serve in an experiment or quasi-experiment, each of which has, say, two arms (e.g., Shadish, Clark, & Steiner, 2008). This creates four arms in total,

and each type of counterfactual is evaluated against its own treatment group that is probabilistically equivalent to the other treatment group. Our goal in this article is to use both kinds of within-study comparisons to estimate the extent to which bias reduction depends on the covariate constructs chosen to index selection, on the reliability with which these constructs are measured, and on how they are then used in the analysis of outcome.

We want to do this primarily with two sets of data that use the four-arm within-study comparison design, each from a different country (Pohl, Steiner, Eisermann, Soellner, & Cook, in press; Shadish et al., 2008). But we also rely for replication on over 30 within-study comparisons using the three-arm design. This design is inferentially weaker than the alternative because of the possibility that the two comparison groups under consideration differ in more than whether one was formed at random and the other not—for example, the two groups may also differ in when and how the measurement of covariates and even of outcomes took place. The results from these within-study comparisons are summarized in Glazerman, Levy, and Myers (2003) in the job training domain and in Cook, Shadish, and Wong (2008) for a variety of social and behavioral science domains. The three-arm designs do not deal with all three of the issues in the four-arm literature, being more concerned with covariate selection and with the comparison of regression and PS methods than with the role of unreliability in the covariates. But they nonetheless do add to the generality of the results from the two four-arm studies that we emphasize somewhat more.

Within-study comparisons depend on the validity of the causal estimate from the randomized experiment. There are some instances in the past where it is legitimate to suspect that the randomization procedure was inadequate (for examples, see Cook et al., 2008) or was at least questionable (Rubin, 2008). This entails that the value of the experimental estimate is not beyond doubt. More important perhaps is that the randomized experiment is only unbiased in expectation; in any one particular instance, the causal relationship is merely estimated because it is necessarily affected by sampling error, the more so when sample sizes are small. Clearly, the experiment requires as much critical scrutiny as the quasi-experiment to which it is yoked in a within-study comparison. Here, we guard against the uncertainty associated with the experimental benchmark not just by replication but also by conducting simulations where we control and specify the size of the effect.

At a highly abstract theoretical level, the problem of overcoming bias in quasi-experiments is trivial. As Cronbach (1982) emphasized, one only needs to know either the complete selection process into treatment or the complete model of the outcome. Economists have also shown that unbiased inference will result from identifying an instrumental variable that is only related to the outcome through the treatment (e.g., Goldberger, 1972; Heckman, 2000). In real social science practice these conditions are sometimes met. With respect to full knowledge of

selection, this is fundamental to the regression-discontinuity design (Cook, 2008; Imbens & Lemieux, 2008) but can sometimes also be found in other contexts. Thus, Diaz and Handa (2006) examined a Mexican program where the treatment was made available to poor villages on a random assignment basis and where, within these villages, allocation to treatment or control status depended on a cutoff score on a multi-item scale of material hardship. The experimental effect was then the difference between eligible villagers in the villages that had or had not been assigned to treatment. The material hardship index was also available on residents of generally more affluent comparison villages and, when it was used to adjust the nonequivalent villagers' scores and to compare them with the scores of treated villagers, the causal estimate in the quasi-experiment did not differ from the estimate in the experiment. This was because the selection process into treatment was fully known and dependent only on the material hardship score. However, truth in advertising forces us to confess that full and convincing knowledge of the selection process is very rare, though many analysts vaguely claim they have achieved it without providing much evidence to support their claim.

Even rarer perhaps is full knowledge of the outcome. But some close approximations are found. In education research on aggregates like schools, it is typical in national samples to find pretest-posttest correlations for standardized achievement measures that are in the .80–.90 range (Hedges & Hedberg, 2007). This correlation typically increases as one aggregates across prior years and corrects for unreliability, resulting in correlations in the .95 range such as those found with special education students (Cook, Wong, et al., 2008). Thus, prediction of the outcome, although not perfect, is nonetheless very close, as it will be in a number of contexts with highly aggregated scores arranged in a time-series. But this situation is hardly likely to arise with the many social and behavioral sciences mostly interested in individual level scores.

As for instrumental variables, it is clear that a few specific ones can be defined from theory. One such case is when the instrument is random assignment itself. Assuming that the treatment is at least partially implemented, random assignment is then only correlated with the outcome through the treatment. This makes it possible to address causal hypotheses of a treatment on treated form as well as an intent to treat one (Angrist, Imbens, & Rubin, 1996). An analogous case is when the cutoff score in regression-discontinuity is used as the instrument, as is implied in Hahn, Todd, and van der Klaauw (2001) so long as the assignment variable is also in the model. But in most substantive contexts, it is very difficult to convince all critics that an instrument has been discovered that only affects the outcome through the treatment.

Although theoretically impeccable circumstances like the aforementioned three can and do occur and can and will reduce all bias, they can constitute only a miniscule fraction of the contexts in which observational studies are

carried out or could be carried out. In social research practice, it is much more typical to encounter both a selection process that is not, and could not have been, fully known and an outcome that cannot be fully predicted. And determined critics have undercut most proposals about specific instrumental variables other than random assignment and regression discontinuity. So, valid theory about eliminating all selection bias provides only indirect guidance for practice. Given this mismatch between statistical theory concerning bias control and the realities of research practice, we need to know more about the empirical effectiveness of various practical strategies for bias reduction in observational studies.

One strategic concern is with the variables chosen to index the selection process into treatment, particularly the part correlated with outcome. Some covariates are bound to be superior to others because they better model the true selection process. Any covariate that is not correlated with either selection or outcome is of no value and will not reduce any selection bias; any covariate or set of covariates that indexes all of the selection process correlated with outcome is bound to be perfect, reducing all the selection bias. So there is little theoretical yield from inquiring which covariates reduce more bias.

But there is a practical payoff. In cross-sectional surveys, modal practice to index selection is probably to use some combination of demographic measures. In fields characterized by longitudinal surveys, an early wave measure of the outcome (the "pretest") is used to index all or part of the selection process. There can even be multiple such pretest waves prior to treatment, as in interrupted time-series. In various psychological studies, individual difference measures might also be used to index selection; in educational contexts where pretest measurement of the outcome might be reactive, correlates of the outcome are often measured instead and then called "proxy pretests" (Cook & Campbell, 1979). And, of course, these various sources can be combined in an attempt to make an even better model of the selection process. Because recognition is universal that covariates count, the issues for practice are (a) which covariates should one choose in any given context and (b) what kinds of covariates are generally better than others in reducing the bias in observational studies. In reanalyzing data from Shadish et al. (2008), Steiner, Cook, Shadish, and Clark (2009) have identified the specific covariates in Shadish et al. that were more or less responsible for bias reduction and have also shown—as Shadish et al. did also—that the manner of data analysis mattered little. But that was in a single short-term study in a single context and did not deal with unreliability and its effects on bias reduction.

Since Lord's (1960) seminal work and Campbell and Erlebacher's (1970) dissemination of it, recognition has been widespread that bias can result from unreliably measured covariates. That is why in structural equation modeling in Psychology, for instance, some form of multiple measurement is now the norm.

But similar practices are not as widespread in, say, Economics and Statistics where measurement error is not assigned a major role in bias reduction. Although the reasons for this disciplinary difference are not completely clear, it likely reflects the beliefs either that unreliability is merely another instance of a missing covariate (West & Thoemmes, in press) or that its effects are small given the high reliability values for many kinds of variables of special importance in these fields—for example, demographic covariates, pretest measures of nationally important outcomes in economics and finance, and the standard use of composites such as propensity scores in Applied Statistics. Regarding the latter, Rubin and colleagues (Rubin & Thomas, 1996, 2000; Stuart & Rubin, 2007) prefer selecting many covariates for potential entry into propensity scores and also using lenient probability levels for including a covariate in the final weighted score. The result is a propensity score composite made up of many variables and thus likely to be particularly reliable. But it may not be perfectly reliable, and where there are very few covariates in the composite it may hardly be more reliable at all. So imperfect reliability is a potential bias-inducing problem even with propensity scores, and no test has yet been made that propensity scores reduce bias at least in part because they are measured more reliably than other individual variables or composites. Steiner, Cook, & Shadish (2009) have such a paper, simulating different levels of unreliability using the basic data of Shadish et al. (2008). Basically, they show that unreliability detracts from the bias-reducing capacity of covariates, the more so with those covariates that have the most potential to reduce bias. But this has not been replicated, and we offer a replication here.

From a theoretical standpoint, unreliability is bound to play a role in bias reduction. If a construct that is perfectly correlated with selection is perfectly measured, no bias will result. If the same construct is measured with zero reliability, then selection will not be controlled at all because the construct has not in fact been measured and a hidden variable problem results. Of course, this holds only if selection operates on the latent rather than the observed covariates as is typical with self-selection processes, for example. Nonetheless, for social science and behavioral research practice the issue is how unreliability affects selection bias reduction within the range normally considered appropriate. This varies, but few fields are willing to tolerate measures with reliability lower than .6. We do the same here, taking constructs that vary in their ability to reduce bias and then asking how this capacity is affected as reliability decreases from 1.0 to .6 in decrements of .1.

Propensity scores were created out of a concern with the validity of regression analyses of data when the treatment and comparison groups come from manifestly different populations. Regression analyses ask how a covariance-adjusted outcome varies for persons with comparable covariate scores. To be valid, such an analysis makes strong assumptions. The most problematic are

(a) the assumption of linear functional form, (b) the need to extrapolate into the unmeasured part of the covariate where the groups are most nonequivalent, and (c) treating covariates singly and thus capitalizing on unreliability. In contrast, propensity scores methods make no assumption about the functional form of the outcome, limit the analysis to the region of overlap between the treatment and comparison groups on the propensity score (also called the region of common support), and they combine many covariates into a single and likely more reliable composite. In theory, propensity score analyses have some salient advantages over regression methods.

In practice, though, it is not clear that they routinely produce effect size estimates that are closer to those from a yoked random assignment experiment. This might be because practice in regression analysis has evolved so much by now that sensitivity to the assumptions is widespread. Or it may be that propensity score practice is often deficient in ways that deny its obvious theoretical privilege—for instance because groups are not as well balanced as they should be. To claim that propensity scores are better in theory does not necessarily entail that social and behavioral scientists routinely practice the art in superior, or even adequate, ways.

For reasons already enumerated, it is theoretically trivial to ask about the capacity of covariate choice, unreliability, and data analysis technique to affect the degree to which selection bias is controlled. But it is not trivial to ask about their relative importance in practice or to ask how the knowledge we already have helps guide practice in observational research. Practitioners look for guidance as to what kinds of covariates they should select—demographic ones, psychological ones, proxy pretests, pretests at one time, or pretests at multiple times. Which turn out to be generally better; or is some strategy of mixing to be preferred; if so, which one? Likewise, practitioners look for guidance about how much unreliability to tolerate even within the bounds between 1.0 and .6 that are traditionally considered acceptable. It should also be noted that most practitioners are better acquainted with regression methods than with propensity scores techniques and that they often have fewer cases than good propensity score analysis calls for. So even if they were to acknowledge the general superiority of propensity score methods in theory, practitioners might still look for guidance about whether these methods are absolutely needed for a given research application.

Conducting research to enlighten practice makes somewhat different sampling demands than conducting research to test theory. In particular, the domain to which we can generalize results becomes more consequential. We adopt two different strategies here, though each involves a particular form of within-study comparison. The first entails reanalysis of the two existing four-arm within-study comparisons in which university students were randomly assigned to serve in a randomized experiment with two arms or in a quasi-experiment with the

same two arms (Pohl et al., in press; Shadish et al., 2008). The intervention and comparison conditions were exactly the same in the experiment and quasi-experiment, as were the testing circumstances. These careful procedures rule out the possibility that variable causal estimates are due to population, treatment or testing procedures that differed between the experiment and adjusted quasi-experiment rather than to the choice of covariates, how well they are measured, and to whether the outcome data were analyzed with regression or PS methods.

But these two studies have limited ecological validity. They took place in a laboratory with college students and the intervention lasted less than half an hour. Fortunately, there are many more within-study comparisons of long-term interventions in "real world" contexts. But these are three-arm designs and inferentially weaker than four-arm ones because many of the nonequivalent comparison groups are constructed from archives or from geographically nonequivalent groups, thus confounding with other attributes of study design the cause of interest—whether design or statistical controls can compensate for comparison groups not formed at random. Nonetheless, we propose to use two reviews of the three-arm within-study comparison literature (Cook et al., 2008; Glazerman et al., 2003) to probe the generality of any finding we generate from our reanalysis of the two four-arm studies.

METHODS

The Four-Arm Studies

We use data from Shadish et al. (2008) and its close German replication by Pohl et al. (in press). In Shadish et al. students from introductory psychology classes in Memphis, Tennessee, were randomly assigned to be in a randomized experiment ($N = 235$) or an observational study ($N = 210$). Students in the randomized experiment were then randomly assigned to mathematics ($N = 119$) or vocabulary training ($N = 116$); those in the observational study self-selected into the training of their choice—79 students chose mathematics and 131 vocabulary (Figure 1). The short treatment consisted either of learning vocabulary or math materials. Before the random assignment to the randomized experiment or observational study, all students took the same vocabulary and math pretest and completed a detailed questionnaire. During these pretests they were treated in exactly the same way except for assignment method. After treatment, all students were tested on both mathematics and vocabulary outcomes. Hence, the math scores of those experiencing the vocabulary training served as controls for those experiencing the math training, and vice versa for the vocabulary intervention.

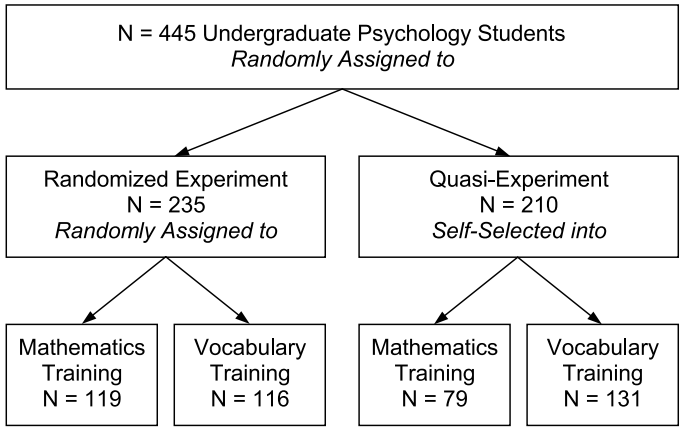


FIGURE 1 Design of the within-study comparison by Shadish et al. (2008).

Prior to treatment, covariate information was collected with a questionnaire of 156 items covering 23 constructs of five substantive domains: (a) demographics including student's age, sex, ethnicity, marital status, and credit hours completed at the university; (b) proxy-pretest measures on vocabulary and math; (c) prior academic achievement scores in university and high school; (d) topic preference, that is, the motivation to learn more about math or vocabulary and a math-anxiety scale; and (e) psychological predisposition measured by the big five variables from personality psychology and the Beck depression scale. In addition we reanalyze the Pohl et al. (in press) data because no reanalyses have yet been conducted with this data set.

Pohl et al. (in press) replicated the within-study comparison of Shadish et al. (2008) in Berlin, Germany, with 202 students of Psychology and Education. Except for the translation into the German language, they used the same Math training, proxy-pretests, and posttest. But instead of the vocabulary training, they administered an English training and measured corresponding pretests and posttests in English. Of the 202 students, 99 were randomly assigned to the randomized experiment and 103 to the observational study (Figure 2). Those in the experiment were then randomly assigned into the Math ($N = 55$ students) and English training ($N = 44$). Of the students in the observational study 55 chose Math and 48 English. Because the study was introduced as an evaluation of training for improving the Math and English skills needed for studying Psychology and Education, a different selection mechanism than in Shadish et al. took place. Students who liked English but showed low pretest values mainly chose the English training. In Shadish et al., students weak in Math avoided the Math training by choosing the vocabulary condition. Prior to

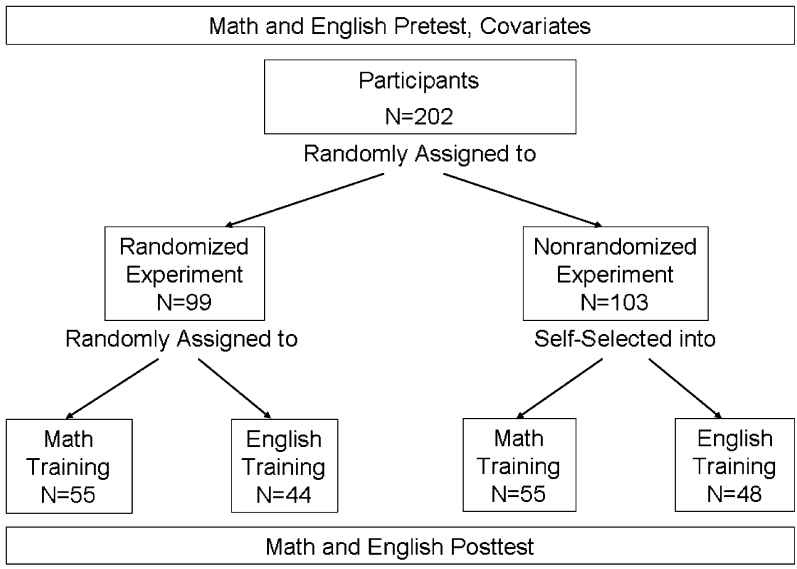


FIGURE 2 Design of the within-study comparison by Pohl et al. (in press).

treatment, Pohl et al. measured 25 constructs from the same five domains as in Shadish et al. Moreover, almost all constructs closely resembled Shadish et al.'s constructs.

Both within-study comparisons applied the same methods in adjusting for selection bias (Lunceford & Davidian, 2004; Morgan & Winship, 2007; Rubin, 2006; Schafer & Kang, 2008): (a) PS stratification with five equal-size strata on the PS; (b) PS-ANCOVA with the logit of the PS included as linear, quadratic, and cubic term; and (c) simple analysis of variance (ANCOVA) with all covariates included as main effects. Shadish et al. (2008) also reported results on PS weighting. They applied PS methods in a "doubly robust" way, that is, PS adjustments are combined with an additional covariance adjustment in the outcome model (Cochran & Rubin, 1973; Robins & Rotnitzky, 1995). Although combining the two adjustments protects against the misspecification either of the PS model or the outcome model, it does not guarantee that more bias is removed than by a PS adjustment alone if both models are misspecified (Kang & Schafer, 2007).

Using the different analytic methods and the five construct domains described earlier (demographics, proxy-pretests, prior academic achievement, topic preference, and psychological predisposition), we investigated the potential of a domain to reduce bias both singly and in various combinations with other domains. In assessing the effect of measurement error on bias reduction for the Pohl et al.

(in press) data we followed Steiner, Cook, and Shadish (under review), who used the Shadish et al. (2008) data for simulating the effect of measurement errors in covariates. They based their simulations on the estimated outcome models assuming that all observed covariates were measured with perfect reliability ($\rho_X = 1$) and that the inclusion of all constructs completely removes all the selection bias. They then investigated the influence of measurement error by systematically decreasing the reliability of each covariate from 1.0 to .9, .8, .7, .6—except for demographics because they are typically measured with high reliability.

The Three-Arm Reviews

Glazerman et al. (2003) meta-analyzed 12 within-study comparisons from the job training evaluation literature, mostly to ask if the experiments and adjusted quasi-experiments resulted in comparable effect sizes. Their main conclusion was that they did not when examined from study to study but did when the average effect size in all the experiments was compared with the average effect size of all the adjusted nonexperiments. But they also examined which kinds of covariates and which kinds of data analysis did better in general, and that is of special interest to us here.

Cook et al. (2008) examined 12 comparisons of experiments and adjusted quasi-experiments and found comparable effects if the quasi-experiment was a regression-discontinuity study; or one with careful selection of a comparison group from a very focal, local population; or if the selection process was almost completely known. They too compared how well the PS and regression methods corresponded in their results, but did so impressionistically and used a vote count method rather than formal meta-analysis.

RESULTS

Reanalyses of Shadish et al. (2008) Data

Figure 3 presents the remaining bias for vocabulary using different sets of constructs. When all the constructs are used, no bias remains and the methods of analysis did not differ, exactly as Shadish et al. (2008) found. However, the construct domains differ in their ability to rule out bias. The topic preference and proxy pretest measures do as well as all the constructs together, and the academic history and psychology variables do hardly better than chance. There is no evidence of a systematic difference due to analytic method either within the three PS scores methods or between the PS and regression (ANCOVA) methods. Figure 4 presents the comparable math results, and we see the same thing. All

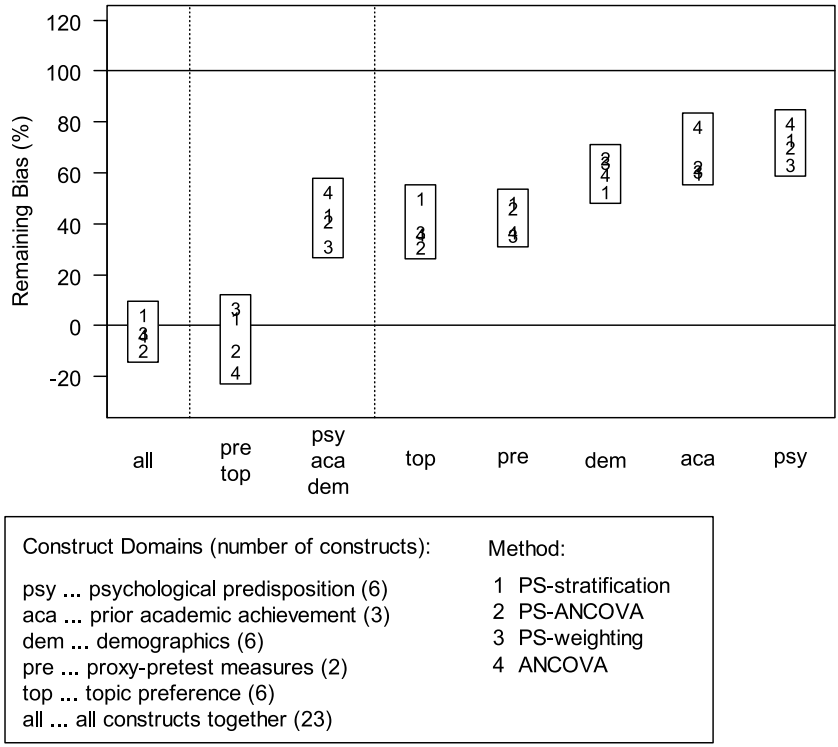


FIGURE 3 Remaining bias (in percentage) in the vocabulary treatment effect by construct domains for the Shadish et al. (2008) data.

the constructs taken together are effective; some construct domains, particularly topic preference, are much better than others; and the type of analysis makes no difference.

Next we turn to the data with the different degrees of measurement error added. Figures 5 and 6 show the bias this causes for the vocabulary treatment first and math next. The results are similar. The biasing effects of measurement error are greater with the constructs that are most effective at reducing bias. Constructs that are not effective are not affected by the bias due to unreliable measurement. The effects of bias from unreliability are not so strong, though, so that the total bias reduction is greater for the poorly measured good constructs than for the well-measured poor constructs. It seems, then, that the choice of covariates counts more within the limits of unreliability manipulated here, the normal range in the behavioral sciences. However, the reliable measurement of constructs is more important than the choice of a specific adjustment method because no

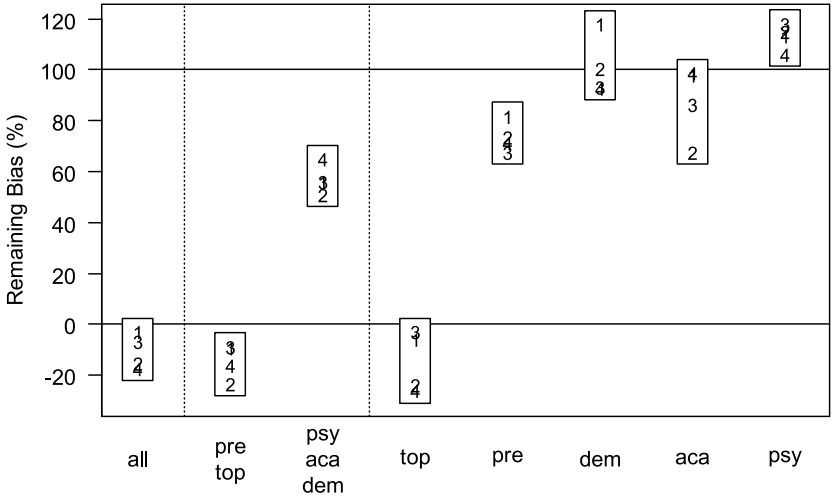


FIGURE 4 Remaining bias (in percentage) in the Math treatment effect by construct domains for the Shadish et al. (2008) data. For construct domains, see Figure 3.

analytic method performed uniformly well or poorly whereas measurement error definitely results in more remaining bias.

Reanalyses of Pohl et al. (in press)

In Pohl et al. (in press), there was no initial bias associated with self-selection into math, and so there is no point to seeing how well covariates reduced bias when none was observed. However, it is worth noting that none of the construct domains induced bias where there was none to start with.

As for English, the results are in Figure 7. Using all the constructs reduced all of the bias, and the mode of analysis did not matter. These two findings replicate the results of Shadish et al. (2008). Also, the specific constructs chosen made a large difference to the amount of bias reduction achieved, as in Steiner, Cook, Shadish, et al. (2009). Combining the pretest and topic preference reduced all bias whereas the academic, demographic, and psychological measures achieved very little bias reduction by themselves or when combined.

Adding the different degrees of unreliability in the simulation again had the same effect. As Figure 8 shows, the better a construct domain's potential for bias reduction the more it was negatively affected by measurement error. But again, the worst measured good constructs did better than the best measured poor constructs. Covariate selection counted most, and as in the American study, the mode of analysis made no systematic difference.

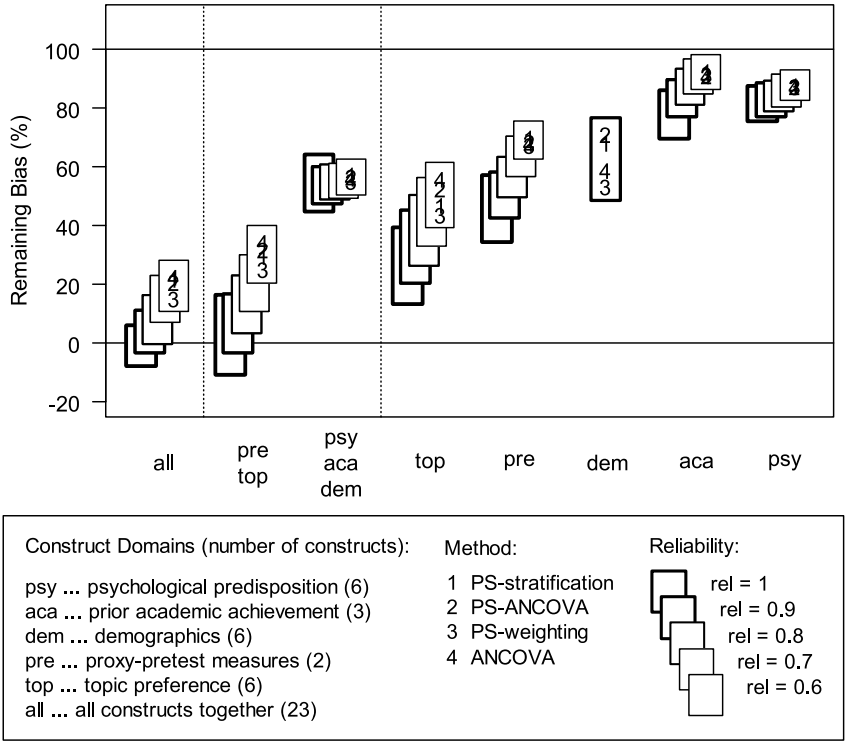


FIGURE 5 Remaining bias (in percentage) in the vocabulary treatment effect by construct domains and reliability for the Shadish et al. (2008) data.

The Generality of These Results in Three-Arm Studies

Both Glazerman et al. (2003) and Cook et al. (2008) reported no discernible differences in effect sizes between the regression or propensity score analyses. Glazerman et al. state that "bias reduction associated with the most common methods—regression, propensity score matching, or other forms of matching—did not differ substantially" (p. 86). So the clear theoretical advantage of propensity scores turns out to be of little practical advantage across the two dozen tests conducted to date.

The totality of these analytic differences has not yet been subjected to thorough statistical meta-analysis to examine whether there is small advantage for one method over the relevant studies and to probe whether one method is superior to the other across the conditions under which one would expect PS or regression to have a comparative advantage. It is clear that propensity scores have some significant theoretical advantages, especially concerning balancing baseline

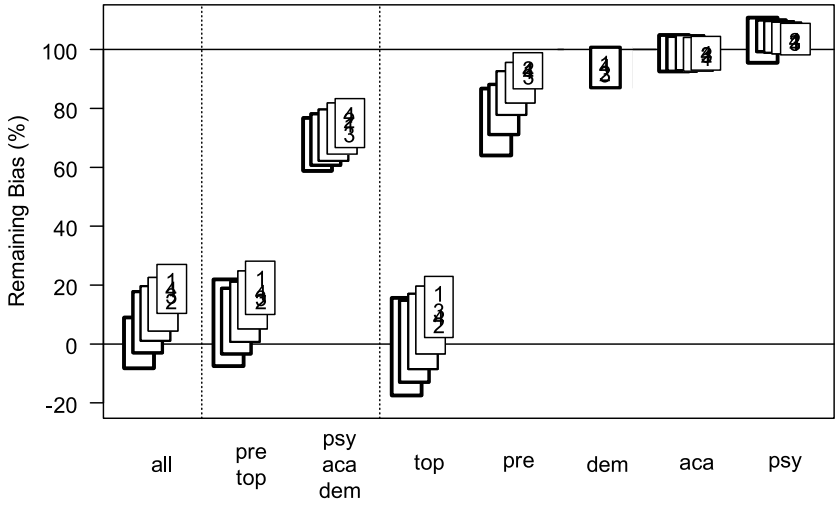
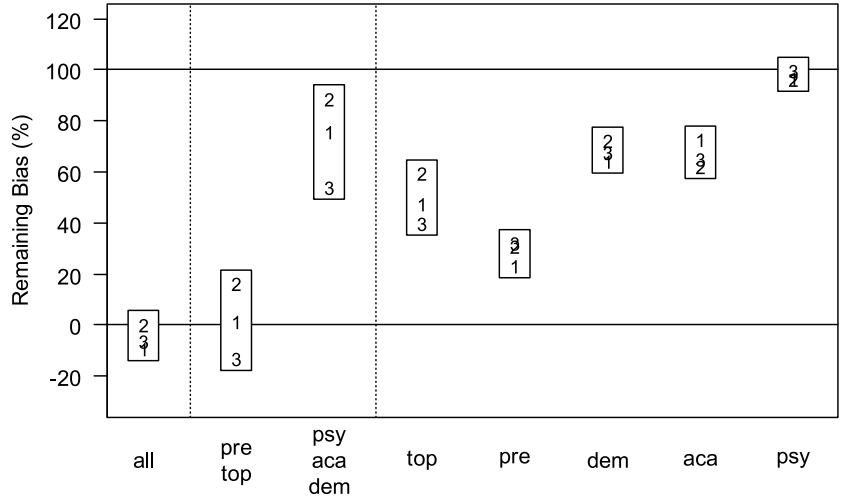


FIGURE 6 Remaining bias (in percentage) in the Math treatment effect by construct domains and reliability for the Shadish et al. (2008) data. For construct domains, see Figure 5.

group differences without consideration of the outcome variable and creating a nonparametric estimation of the treatment effect. But even so, any differences by analysis are likely to be small on average and limited in their conditions of application by the assumptions built into the theories of both regression and PS methods. Although these two dominant modes of contemporary data analysis differ in theory, it is not clear that these differences have major consequences for bias reduction in scientific practice—at least not in the practice of the sample of within-study comparisons examined here.

The same is not true of the effects of covariate selection. In theory, covariate selection should matter considerably, depending on how well a covariate is correlated with both selection and outcome, a fact now demonstrated in two four-arm within-study comparisons in two countries and in the reviews of many three-arm studies. The bias reductions achieved were greatest of all when pretest measures of the outcome were available and when the nonequivalent comparison cases were local rather than distant. Correspondence was also greater when the selection process was better known and identical measures of it existed for the treatment and the nonequivalent comparison group (Diaz & Handa, 2006) and also when the treatment and comparison groups were deliberately selected from populations known to be minimally different on pretest means. Design tradition and empirical results indicate that some covariates are better than others, with the consistent worst being when just standard demographic data were used to index the selection process.



Construct Domains (number of constructs):

psy ... psychological predisposition (2)
aca ... prior academic achievement (8)
dem ... demographics (8)
pre ... proxy-pretest measures (2)
top ... topic preference (5)
all ... all constructs together (25)

Method:

1 PS-stratification
2 PS-ANCOVA
3 ANCOVA

FIGURE 7 Remaining bias (in percentage) in the English treatment effect by construct domains for the Pohl et al. (in press) data.

The usual advice in covariate selection is to collect pretest data on many variables. That is why Shadish et al. (2008) had 23 constructs based on 156 questionnaire items and Hong and Raudenbush (2006) had 206. But we demonstrated here that a small subset of constructs (pretest and motivational measures) would have sufficed in Shadish et al. and Pohl et al. (in press), *had they been known in advance.* The practical problem is to know them in advance, and so there is no substitute for being the anthropologist of the situation and knowing as much as possible about how units get into treatments and which constructs predict the outcome. Sometimes it will be reasonably transparent but at other times quite opaque. In any event, the covariate selection clearly counts, and that is evident in both Glazerman et al. (2003) and Cook et al. (2008) as well as in the two four-arm studies analyzed here.

Both of the four-arm studies also showed that the effects of unreliability on bias reduction were greater with effective than ineffective constructs. But the

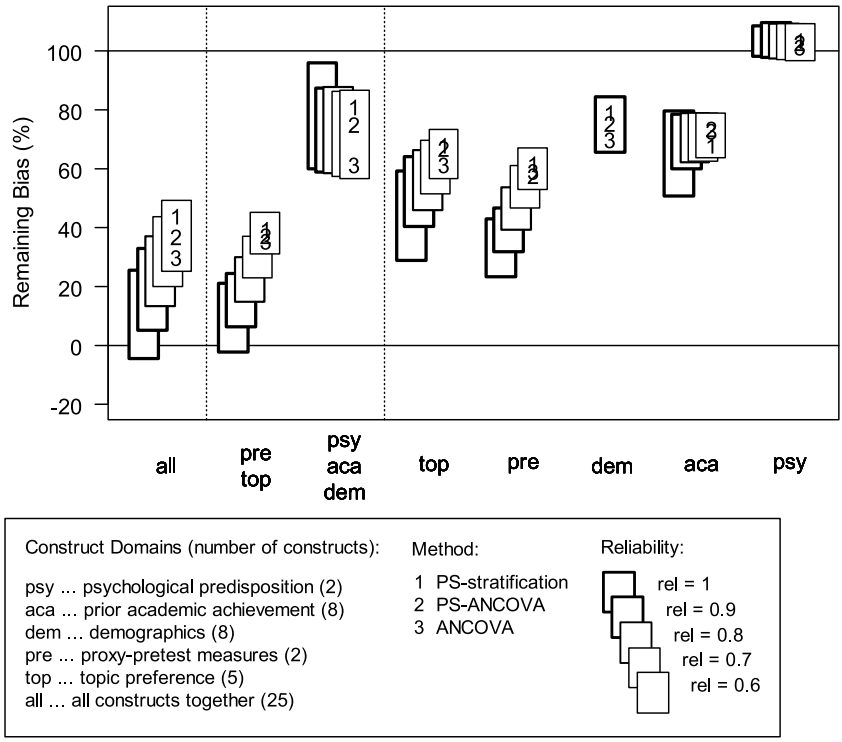


FIGURE 8 Remaining bias (in percentage) in the English treatment effect by construct domains and reliability for the Pohl et al. (in press) data.

role of unreliability was never investigated in any of the other three-arm studies published to date. So we have not been able to replicate that the best constructs were more affected by poor measurement or that unreliability counted less for bias reduction than covariate choice. But these are two clear findings in the technically superior four-arm within-study design literature.

DISCUSSION

One question we asked concerned the relative importance of covariate selection, unreliability of measurement, and mode of data analysis. The evidence currently available is overwhelming that, in within-study comparisons as they are now conducted and the data from them analyzed, covariate selection is most important, unreliability next, and mode of data analysis least. Even reducing unreliability to an average of .60 fails to make the best covariates function as

poorly as the worse covariates did when they were measured perfectly. So the art of observational study design and practice is to discover the best covariates and to measure them really well. So long as there is some sensitivity to major assumptions, how the data are analyzed seems to matter very little. This finding is corroborated by two meta-analyses of observational studies in epidemiology that also concluded that PS and regression methods do not differ in practice (Shah, Laupacis, Hux, & Austin, 2005; Stürmer et al., 2006). However, propensity score methods are theoretically superior for many of the most obvious conditions of application.

This article sought to embed its findings within a multiple replication perspective, albeit one limited to within-study comparisons because of their ability to provide a true causal baseline in expectation even if not in each individual experiment. We were successful in this for comparisons of covariate choice versus mode of data analysis, but less so for comparisons involving measurement error since neither Glazerman et al. (2003) nor Cook et al. (2008) dealt with such error. The full data indicate quite securely that covariate choice matters considerably and that choice of data analysis technique does not, at least not within the quite heterogeneous range of applications studied to date. Data from just the two four-arm studies suggests, somewhat less securely, that measurement error counts even with propensity scores but that its role is quite modest when compared to that of covariate choice.

ACKNOWLEDGMENTS

Thomas D. Cook and Peter M. Steiner were supported in part by Grant R305U070003 from the Institute for Educational Sciences, U.S. Department of Education. Peter M. Steiner was also supported by grants from the W. T. Grant Foundation and Spencer Foundation. Thanks to David Kenny for a critical and helpful reading.

REFERENCES

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.

Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), *The disadvantaged child: Vol. 3. Compensatory education: A national debate* (pp. 185–210). New York: Brunner/Mazel.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, 35,* 417–466.

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 issues for field settings*. Chicago: Rand-McNally.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27,* 724–750.

Cook, T. D., Wong, V. C., Taylor, J., Gandhi, A., Kendziora, K., Choi, K., et al. (2008). *Impacts of school improvement status on students with disabilities: Technical work group materials.* Washington, DC: American Institutes for Research.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

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

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

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica, 40,* 979–1001.

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. J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics, 115,* 45–97.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlations for planning group-randomized experiments in education. *Educational Evaluation and Policy Analysis, 29,* 60–87.

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.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice *Journal of Econometrics, 142,* 615–635.

Kang, J., & Shafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science, 26,* 523–539.

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

Lord, F. M. (1960). Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55,* 307–331.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via propensity score in estimation of causal treatment effects: A comparative study. *Statistical Medicine, 23,* 2937–2960.

Morgan, S. L., & Winship C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge, UK: Cambridge University Press.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (in press). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis.*

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90,* 122–129.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge, UK: Cambridge University Press.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2,* 808–840.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Schafer, J. L., & Kang, J. (2008). Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103,* 1334–1344.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of Clinical Epidemiology, 58,* 550–559.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2009). *On the importance of reliable covariate measurement in selection bias adjustments using propensity scores.* Manuscript submitted for publication.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2009). *The importance of covariate selection in controlling for selection bias in observational studies.* Manuscript submitted for publication.

Stuart E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.), *Best practices in quantitative methods* (chap. 11, pp. 155–176). Thousand Oaks, CA: Sage.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology, 59,* 437–447.

West, S. G., & Thoemmes, F. (in press). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*.