Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Bias Reduction in Quasi-Experiments With Little Selection Theory but Many Covariates

Peter M. Steiner, Thomas D. Cook, Wei Li & M. H. Clark

To cite this article: Peter M. Steiner, Thomas D. Cook, Wei Li & M. H. Clark (2015) Bias Reduction in Quasi-Experiments With Little Selection Theory but Many Covariates, Journal of Research on Educational Effectiveness, 8:4, 552-576, DOI: 10.1080/19345747.2014.978058

To link to this article: https://doi.org/10.1080/19345747.2014.978058

Accepted author version posted online: 09 Dec 2014.
Published online: 02 Jul 2015.

Journal of Research on Educational Effectiveness, 8: 552–576, 2015

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2014.978058

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

Bias Reduction in Quasi-Experiments With Little Selection Theory but Many Covariates

Peter M. Steiner
University of Wisconsin–Madison, Madison, Wisconsin, USA

Thomas D. Cook and Wei Li
Northwestern University, Evanston, Illinois, USA

M. H. Clark
University of Central Florida, Orlando, Florida, USA

Abstract: In observational studies, selection bias will be completely removed only if the selection mechanism is ignorable, namely, all confounders of treatment selection and potential outcomes are reliably measured. Ideally, well-grounded substantive theories about the selection process and outcome-generating model are used to generate the sample of covariates. However, covariate selection is more heuristic in actual practice. Using two empirical data sets in a simulation study, we investigate four research questions about bias reduction when the selection mechanism is not known but many covariates are measured: (1) How important is the conceptual heterogeneity of the covariate domains in the data set? (2) How important is the number of covariates assessing each domain? (3) What are the *joint* effects of this conceptual heterogeneity and of the number of covariates per domain? (4) What happens to bias reduction when the set of covariates is deliberately impoverished by removing the covariates most responsible for selection bias, thus ensuring a slightly smaller but still heterogeneous set of covariates? The results indicate: (1) increasingly more bias is reduced as the number of covariate domains and the number of covariates per domain increase, though the rate of bias reduction is diminishing in each case; (2) sampling covariates from multiple heterogeneous covariate domains is more important than choosing many measures from fewer domains; (3) the most heterogeneous set of covariate domains removes almost all of the selection bias when at least five covariates are assessed in each domain; and (4) omitting the most crucial covariates generally replicates the pattern of results due to the number of domains and the number of covariates per domain, but the amount of bias reduction is less than when all variables are included and will surely not satisfy all consumers of causal research.

Keywords: Observational study, causal inference, propensity score, covariate selection

Address correspondence to Peter M. Steiner, University of Wisconsin–Madison, Department of Educational Psychology, 1025 W. Johnson Street, Madison, WI 53706, USA. E-mail: psteiner@wisc.edu

INTRODUCTION

Many philosophers of science contend that causation is irreducible and so not capable of a single definition (Collingwood, 1940). However, most fields of study have their own preferred theory or theories of causation. Currently dominant in statistics and much of the social sciences is Rubin's Causal Model (Holland, 1986; Rubin, 1974, 1979), abbreviated here to RCM. It asserts that the effect of a cause can be theoretically described as the difference between the same subject's potential treatment and potential control outcomes, that is, the outcomes one would observe after the subject simultaneously selected into both the treatment or control conditions. Because this cannot happen, RCM abjures individual causal estimates in favor of average causal effects computed across samples of subjects exposed to, say, a treatment and a control condition. But for this average causal effect to be valid requires that the process of selection into the treatment or control group is fully known or at least ignorable (Rosenbaum & Rubin, 1983).

One way of ensuring such ignorability is to assign treatments at random, a procedure whose value applied statisticians recognized long before RCM. But random assignment is not universally applicable, for example, for ethical or organizational reasons. Moreover, it sometimes breaks down in specific applications, for example, when attrition differs across the treatment and control groups. Nonetheless, it can serve as the blueprint for designing observational studies and explicitly plays this role in RCM's theory on propensity score methods for causal inference.

Seen more broadly, random assignment is but one instance of a widely recognized general principle for warranting unbiased causal inference. This is when the process of selection into treatment is completely known, validly measured, and used to adjust for a potential nonequivalence between the treatment and control groups (Cook, 2008; Goldberger, 1972; Shadish, Cook, & Campbell, 2002). Another instance of this same principle is the regression discontinuity design (RDD), in which treatment assignment depends only on whether a study unit scores at or above a known value on a continuous assignment variable (Cook, 2008; Thistlewaite & Campbell, 1960). The same principle also applies in some very rare policy applications where the treatment assignment mechanism is fully deterministic (e.g., Diaz and Handa, 2006). But this last situation is so rare that it hardly applies to practical research action.

Outside these contexts, full knowledge of the selection process is not possible, forcing researchers to rely on weak or incomplete theories about the selection and outcome model (i.e., the data-generating process) or to equate groups on whatever covariates have been observed. But without knowing the data-generating process it is impossible to know whether bias remains due to unobserved characteristics that are related to the selection process and study outcome (Cochran & Rubin, 1973; Rosenbaum, 2002; Rubin, 2006).

Nonetheless, clues to the selection process are sometimes available from systematic past research on the selection process in question, from direct observation of it, from interviews with experts, from armchair theorizing, and even from "common sense." Such procedures improve the knowledge about the data-generating process and should guide the choice of covariates in hopes of increasing bias reduction. Graphical models that visualize the data-generating process according to a researcher's knowledge and belief can then be helpful for selecting the confounding covariates or realizing that measures of some confounders are lacking or fallible (Pearl, 2009). At this last point, researchers should probably end the analysis. But often they do not, so strong is the temptation to proceed with whatever other pretreatment covariates are available, however poor or unclear their theoretical relation to selection might be. Even worse, little or no thought might have been

given to plausible selection processes. Instead, reliance is placed on the available covariates in the belief that they are "rich" enough to establish an ignorable selection mechanism when used to analyze the study outcome.

This article explicates one understanding of a "rich" set of covariates. It characterizes "richness" in terms of: (1) how much conceptual heterogeneity there is across the covariate domains observed; (2) how much variation there is in the number of covariates observed within each domain; and (3) how the heterogeneity of domains and the reliability of domain assessments combine to affect bias reduction in observational studies. We understand covariates as variables that can be used to control for selection, whether as single items or as multi-item scale totals that are typically more reliable and thus more likely to reduce bias (Steiner, Cook, & Shadish, 2011). We understand domains to be more general than covariates. They are the labeled categories into which individual covariates can be fit because they have face validity within that category. For instance, covariates about age and gender fit better into the domain of "demography" than "academic achievement," while covariates assessing reading or math are more validly assigned to "academic achievement." Although covariate domains are not necessarily independent, the presumption is that the conceptual heterogeneity is greater the more domains a set of covariates represents. Within a domain, the covariates are typically correlated with each other. So a domain is more reliably assessed as new measures are added that contribute unique variation relevant to the domain's category label. We understand "rich" covariate data in terms of either the heterogeneity of domains, the number of items per domain and, most plausibly, as the combination of both domain heterogeneity and domain reliability.

This conceptualization of "richness" requires access to data sets with many covariate domains as one factor and many items within each domain as the second factor. The research we present uses two such data sets, each estimating learning impacts in education. The data used in Shadish, Clark, and Steiner (2008, hereafter SCS) assess the effects of undergraduate training in math or vocabulary, while the Early Childhood Longitudinal Study (hereafter ECLS-K) assesses the effects of being retained in kindergarten (Hong & Raudenbush, 2006). Each data set has at least 150 covariates measured prior to the intervention. We partition them into different numbers of domains (up to 12), with up to seven covariates in each. This permits a replicated simulation study that systematically constructs propensity scores (PS) from different numbers of covariate domains and covariates per domain in order to estimate how well they reduce the selection bias in an observational study.

The two data sets we use are different from each other in some ways that serve to probe the robustness of any findings. With the SCS data set, the covariate measures were explicitly selected from theory and experience in the belief that they were likely to remove the bias due to self-selection into math or vocabulary training. In contrast, the ECLS-K data are from a national longitudinal survey and were presumably chosen by a committee that gave no thought to testing a hypothesis as specific as how grade retention affects achievement scores. Both data sets have some covariates representing traditional single-item constructs such as race, age, and gender, and both also have some multi-item scale totals such as math achievement. But while the SCS data set allows us to break out the individual items from such totals in order to vary the number of covariates per domain, only scale totals are available to us from ECLS-K. So when we vary the number of covariates per domain in ECLS-K these covariates are more often scale totals than in SCS. As such, they will usually be more reliable. And finally, the SCS design involved randomly assigning students to serve in a randomized experiment or a quasi-experiment. In the former, they were then randomly assigned to treatments (either a math or vocabulary training), while in the latter they self-selected themselves into treatments. This design creates a clear experimental

benchmark against which the extent of bias reduction can be assessed. But the costs of this are a brief intervention lasting less than 15 minutes and laboratory-like control over all the testing circumstances. In contrast, ECLS-K has no causal benchmark from an experiment. Instead, the most plausible benchmark is quasi-experimental and comes from the use of an unusually large set of 208 well-measured covariates over two time periods prior to intervention in a context where the selection process is plausibly yet imperfectly known. However, the advantage of ECLS-K is that the intervention lasts a year and takes place in circumstances such as those to which we want to generalize. So any replicated findings about bias reduction that are achieved across such disparate research settings should increase confidence about the robustness of the role of domain heterogeneity and domain reliability in reducing selection bias.

The two data sets have been used before to identify the covariates most responsible for bias reduction. Steiner, Cook, Shadish, and Clark (2010) analyzed the SCS data set to identify which individual covariates were responsible for removing the demonstrated selection bias. They discovered that two covariates removed almost all the selection bias from the vocabulary outcome: the single-item preference for mathematics over literature and the 36-item vocabulary proxy-pretest total. For the mathematics outcome, the critical covariates removing almost all the bias were two items from a single domain: liking for mathematics and preferring mathematics over literature. We designate these as the "critical covariates" for each outcome and retain them when analyzing all covariates. However, we deliberately omit them in other tests comprised of covariates that are all individually less effective. The issue is whether their cumulative association with the best approximation to the true selection process is nonetheless high enough that they remove all or most of the selection bias in PS analyses. In other words, do the heterogeneity of domains and the reliability of domains still significantly reduce bias when the very best covariates are deliberately omitted?

The same PS analyses with and without the "critical covariates" is also possible for ECLS-K. Hallberg (2013) identified two sets of time-varying covariates that remove nearly as much selection bias as all the 208 covariates combined—the two pretest waves of student math and reading assessed by standardized tests, and teacher judgments of student performance over the same two waves. Either combination was as effective as when all 208 covariates had a chance to form the PS, perhaps because the propensity to be retained a grade is largely determined by academic performance or teacher judgments of such performance. So we replicate PS analyses both with and without the critical covariates of math and reading pretests and teacher judgments.

Omitting the critical covariates creates not just the usual under-explicated selection theory but a theory that is manifestly incomplete. This then permits testing whether the remaining and still very heterogeneous set of covariates can compensate for the missing critical covariates by capturing their association with the omitted covariates. If the heterogeneity and reliability of construct domains help with poorly explicated selection theory, how well do they help in the even more potentially debilitating circumstance of an explicitly compromised selection theory?

In summary, this study explicates a "rich" set of covariates to be those that are both heterogeneous in domain coverage and reliable when measured. It then tests how variation in the number of domains and of covariates per domain affects the amount of selection bias remaining after PS adjustment in two quite disparate quasi-experiments. More specifically, we seek: (a) to describe how selection bias is related to the heterogeneity of covariate domains and the reliability of domain assessments when each is examined individually and in combination; (b) to identify the particular combination of heterogeneity and reliability that reduces most selection bias; (c) to estimate the relative importance of heterogeneity and

reliability in reducing bias; (d) to solve the practical trade-off that arises when researchers are faced with the dilemma of using their limited resources either to sample more domains or to sample more covariates within a domain; and (e) to assess what happens to bias reduction when the critical covariates are deliberately omitted but many covariates still remain even though they are individually fallible.

METHODS

Data Sources

The Shadish et al. (2008) study contrasts a random assignment study and an observational study with self-selection into treatment. Figure 1 shows the design. Overall, 445 volunteer undergraduate students from introductory psychology classes at a large mid-south public university participated. All took an extensive pretest measurement battery that covered demographic and educational characteristics, vocabulary and arithmetic aptitude tests, vocabulary and math preference tests, and personality tests. Archival data were also collected on past academic performance in high school and college. After the pretest, students were randomly assigned into a randomized experiment or an observational study. The latter required them to choose the treatment to which they would be assigned. The 235 students assigned to the randomized experiment were then further randomly assigned, this time to receive either mathematics training ($N = 119$) or vocabulary training ($N = 116$). The 210 students in the observational arm of the study were asked to choose their preferred training: 79 chose mathematics and 131 vocabulary. Students in the random assignment or self-selection group then attended the same training sessions and were assessed at posttest in the same way. The posttest consisted of 30 vocabulary items and 20 mathematics items, half of which were directly linked to the content of the instruction and the others were

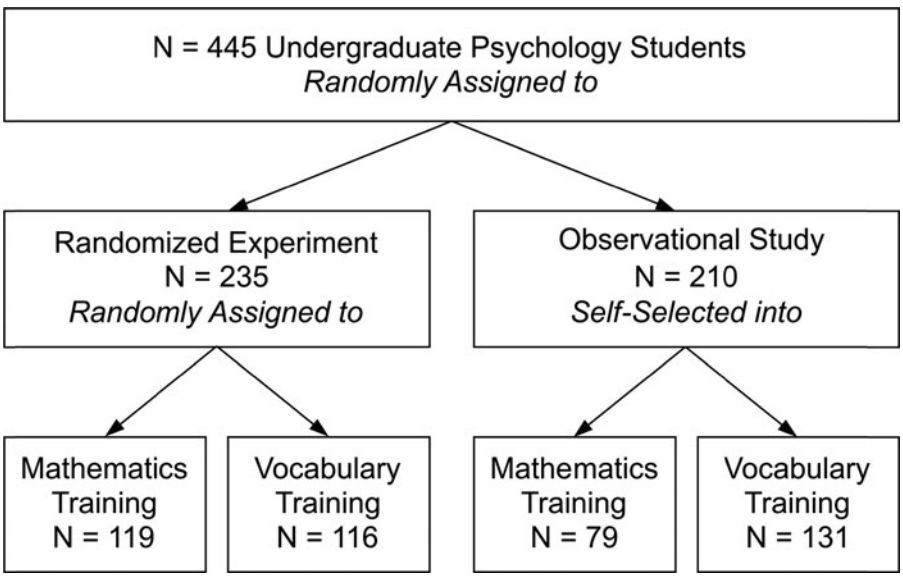

Figure 1. Within-study comparison of Shadish, Clark, & Steiner (2008).

generalizations therefrom. Thus, the posttest tapped into highly specific domain knowledge while the pretest assessed general aptitude for the domain (Clark, 2000). This entails that the pretest is a proxy measure of the outcome, not a true measure of it.

This Shadish et al. (2008) design ensures that (a) assessments during the pretest period can function as covariates because they were collected before assigning treatments; (b) the treatment and testing procedures are identical for the randomized and nonrandomized groups in both mathematics and vocabulary training, thus ruling out inadvertent differences between the two design types in treatment content, implementation, or outcome assessment methods; and (c) each treatment group can be used as the control group for the other, that is, the vocabulary group serves as the control group for estimating math training effects while the mathematics group serves as the control group for estimating vocabulary effects under the assumption that brief training in the one topic does not affect ratings of the other.

To remove selection bias, we classified the 157 questionnaire items of the SCS data into twelve heterogeneous covariate domains:

1. Demographics (five constructs, each with a single covariate item): age, sex, race (White, African American, and other), marital status, and number of college credit hours.
2. Vocabulary proxy pretest (one construct based on 36 individual items; Educational Testing Service, 1962).
3. Math proxy pretest (one construct based on 15 covariate items; Educational Testing Service, 1993).
4. Prior academic achievement (three covariates, each based on multiple items but available only as a collapsed total): high school GPA, current college GPA, and the ACT college admission scores.
5. Mathematics Anxiety Rating Scale (one construct based on 25 covariate items; Faust, Ashcraft, & Fleck, 1996).
6. Topic preference (five constructs but a total of 10 covariate items) assessing: the number of prior mathematics courses taken, liking literature, liking mathematics, preference for mathematics over literature, and whether the major field of study is math intensive or not.
7. Short Beck Depression Inventory (one construct based on 13 covariate items; Beck & Beck, 1972)
8. Extroversion (one construct based on 10 covariate items; Goldberg, 1992)
9. Emotional stability (one construct based on 10 covariate items; Goldberg, 1992)
10. Agreeability (one construct based on 10 covariate items; Goldberg, 1992)
11. Openness to experience (one construct based on 10 covariate items; Goldberg, 1992)
12. Conscientiousness (one construct based on 10 covariate items; Goldberg, 1992)

As in SCS and Steiner et al. (2010), we use some single items as constructs within a domain (e.g., respondent sex within demographics, or the preference for math over literature in topic preference). But in contrast to the earlier studies that used test totals for most constructs, we randomly sample single items from the corresponding multi-item constructs (e.g., each item of the Short Beck Depression Inventory represents a covariate in these analyses). Such items are usually less reliable than the totals from which they are sampled. As noted earlier, Steiner et al. (2010) identified two covariates for math and two others for vocabulary that reduced all selection bias. These we treat as the critical covariates and include them in some analyses but omit them from others.

Overall, fewer than 1% of all covariates were missing. We imputed missing covariate values via the Markov Chain Monte Carlo method as implemented in the multiple imputa-

tion procedure MI in SAS 9.2 (Li, 1988). Given the rather small number of missing values and our purely methodological interest in the covariates' importance in removing selection bias we used a single rather than multiple imputation procedure.

The second data set we use is the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K). This involves a national sample of the 1998–1999 cohort of kindergarten students who were followed through their eighth grade. Various indicators of a child's early learning experiences were collected twice a year—including whether a child was retained after the first kindergarten year instead of being promoted (Hong & Raudenbush, 2006). We examine the effect of such retention on subsequent academic performance.

For the covariates, we use 208 variables, some collected at a single time, but others over the fall and spring of the kindergarten year prior to the retention decision. Following Hallberg (2013), we classified these 208 covariates into 10 covariate domains:

1. Mathematics and reading pretest performance (12 multi-item covariates consisting of two waves each of math and reading achievement scores plus two waves each of four teacher proxy assessments of student performance).
2. Child cognitive skills (6 covariates)
3. Child demographics (10 covariates)
4. Child social skills (22 covariates)
5. Classroom demographic composition (10 covariates)
6. Classroom learning environment (28 covariates)
7. School demographic composition (6 covariates)
8. School structures and supports (72 covariates)
9. Teacher demographics (18 covariates)
10. Home environment (18 covariates)

Within these construct domains, some covariates are single-item but most are multi-item. In the latter case, the information available to us precludes access to the item-level data. For example, pretest math and reading are only available as multi-item totals. However, each is assessed on two occasions prior to retention as well as once after it, and each is assessed using both objective test scores and teacher assessments of performance. And as noted earlier, Hallberg (2013) found that two pretest waves of either the teacher reports or objective assessments of performance reduced as much bias as when all 208 covariates had a chance to form the PS. So we replicate the PS analyses both with and without these critical covariates.

Although about 88% of the 208 variables in ECLS-K had at least one missing value, only 6% of the data was missing in total. In imputing the missing values we followed Hallberg (2013) and used her singly imputed data set (Markov Chain Monte Carlo imputation).

Treatment Effects and Estimators

To investigate the importance of a large set of covariates with coverage of heterogeneous covariate domains, we systematically sample covariate domains and subsets of covariates from within domains and estimate the bias remaining after the PS-adjustment. For both data sets, we aim at estimating the average treatment effect for the overall study population (ATE).

Using the Rubin Causal Model and its potential outcomes notation (Holland, 1986; Rubin, 1974), we can define the average treatment effect (τ) as the expected difference between potential treatment and control outcomes,

$$\tau = E\left[Y(1) - Y(0)\right], \tag{1}$$

where $Y(0)$ is a subject's potential control outcome, which we would observe if a subject selects into the control condition, and $Y(1)$ is a subject's potential treatment outcome, which we would observe in case the subject selects into the treatment condition. In general, $Y(Z)$ indicates the potential outcome under the treatment condition Z, where $Z = 0$ for the control condition and $Z = 1$ for the treatment condition. Note that Equation (1) cannot directly be used to derive an estimator for the average treatment effect because we never observe both potential outcomes simultaneously. Depending on the treatment condition Z, we either observe the potential control outcome or the potential treatment outcome: $Y = ZY(1) + (1 - Z)Y(0)$. Using the observed outcomes, we can estimate the prima facie effect, that is, the unadjusted difference between the treated subjects' average outcome and the control subjects' average outcome, $\hat{\tau}_{PF} = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$. But due to differential selection into treatment and control conditions, the prima facie effect is in general a biased estimator of the average treatment effect.

However, under certain conditions we can obtain unbiased estimates of the average treatment effect as defined in Equation (1). One such condition is random assignment of subjects to treatment conditions. Because randomization establishes independence between the pair of potential outcomes ($Y(0)$, $Y(1)$) and treatment assignment Z, the prima facie effect, $\hat{\tau}_{PF}$, is an unbiased estimator of the average treatment effect. For observational data, where subjects systematically select or get assigned into treatment conditions, we can estimate an unbiased treatment effect only if the strong ignorability assumption is met (Rosenbaum & Rubin, 1983). Treatment selection is said to be strongly ignorable if the potential outcomes ($Y(0)$, $Y(1)$) and treatment selection Z are independent, given a set of observed covariates $\mathbf{X}$, that is,

$$(Y(0),\ Y(1)) \perp Z|\mathbf{X} \text{ with } 0 < \mathrm{P}(Z = 1|\mathbf{X}) < 1.$$

Assuming strong ignorability, Rosenbaum & Rubin (1983) then proved that the average difference in conditional expectations, $\mathrm{E}_X(\mathrm{E}(Y(1) \mid Z = 1, \mathbf{X})) - \mathrm{E}_X(\mathrm{E}(Y(0) \mid Z = 0, \mathbf{X}))$, is equivalent to the average treatment effect τ as defined in Equation (1). Rosenbaum and Rubin (1983) also showed that if the strong ignorability is met, we can replace the set of covariates with the corresponding PS, $e(\mathbf{X}) = \mathrm{P}(Z = 1 \mid \mathbf{X})$, and still achieve the same conditional independence between potential outcomes and treatment selection: $(Y(0), Y(1)) \perp Z \mid e(\mathbf{X})$.

The strong ignorability assumption is defined in terms of the true PS, which is unobserved in practice. Thus, we need to estimate the PS from observed covariates $\mathbf{X}$ using logistic regression, for instance. Whether the estimated PS succeeds in balancing pretreatment group differences on the observed covariates can then be probed with balancing tests such as the standardized mean differences in covariates, including the logit of the PS, and corresponding t-tests (Steiner & Cook, 2013). Given that the estimated PS succeeds in balancing the treatment and control group's distribution of observed covariates, we can use the PS for estimating a PS-adjusted average treatment effect. Most techniques suggested for estimating PS-adjusted effects belong to one of four main PS methods: PS matching, PS stratification, inverse-propensity weighting, and PS regression adjustments (for an overview, see Imbens, 2004; Schafer & Kang, 2008; Steiner & Cook, 2013).

We will estimate the PS using logistic regression and will determine the treatment effect via PS regression estimation with the logit of the PS included as a quadratic polynomial, that is,

$$Y_i = \beta_0 + \tau_{PS} Z_i + \beta_1 \hat{l}_i + \beta_2 \hat{l}_i^2 + \varepsilon_i, \tag{2}$$

where $\hat{l}_i = \log(\frac{\hat{e}_i(\mathbf{X})}{1-\hat{e}_i(\mathbf{X})})$ is the logit of the estimated PS, $\hat{e}_i(\mathbf{X})$, the βs represent the regression coefficients of the corresponding linear and quadratic term of the PS-logit, and τ_{PS} is the PS-adjusted treatment effect. The error term ε_i is assumed to be independent and identically distributed. We used the PS-logit instead of the PS in the regression model since it is usually more linearly related to the outcome of interest (Rubin, 2001). For both of our data sets, the inclusion of a second order polynomial of the PS-logit is based on exploratory analyses using Akaike's Information Criterion (AIC).

The PS regression estimation turned out to be more robust in our simulations than PS matching, stratification, or inverse-propensity weighting. Particularly for small or medium-sized samples such as the SCS data, PS regression is less sensitive to a lack of overlap within the PS logit distribution (as opposed to at the tails of the distribution). This is because the parametric functional form allows PS regression to inter- and extrapolate to regions of non-overlap. Nonetheless, the results from using PS regression estimation were similar to those from PS stratification and inverse-propensity weighting, although the latter techniques had greater variance of the simulated treatment effects. Such similarity of causal estimates is not surprising because simulation studies and within-study comparisons have regularly shown that the choice of data-analytic method is of little importance in removing selection bias when compared to the nature of the covariates available and how reliably they are measured (Cook & Steiner, 2010; Steiner et al., 2010). Nonetheless, for each subset of covariates sampled, we checked the overlap between the treatment and control group on the PS logit and deleted non-overlapping cases. To maintain comparability with the benchmark estimates, we also deleted cases with correspondingly extreme PSs (predicted from the estimated PS model) in the randomized experiment (SCS data) or the benchmark analysis with all 208 covariates (ECLS-K). This procedure leads to slight variations in the underlying target population across iterations of our simulation, but it ensures that the same reference population is used for the benchmark estimates (even if the treatment effects vary across the truncated target populations, the comparability of results is maintained because we present the results in terms of relative instead of absolute bias).

Design of Simulation Study

To investigate the effect of an increasing number of covariate domains and of covariates per domain we simulated variation in the number of domains and of covariates per domain. Let M be the total number of covariate domains in the data set and N_{j} the number of covariates in the jth domain, $j = 1, \ldots, M$. Then, our goal is to determine the extent of bias reduction as the number of domains (m) and the number of covariates (n) within domains increases. For our two data sets, it would be possible, at least theoretically, to evaluate the extent of bias reduction for all possible combinations of domains and covariates within domains. However, the large number of available covariates, particularly for the ECLS-K data set, results in too many combinations to be evaluated within a reasonable time. Thus, we resorted to a sampling strategy that investigates all possible combinations of domains but randomly samples covariates within each single combination of domains.

More formally, in choosing m covariate domains from all M available domains we get $H = C(M, m)$ possible combinations of domains, where $C(M, m) = \frac{M!}{m!(M-m)!}$ is the

binomial coefficient. For each selected domain j, we then have $C\left(N_j, n\right)$ possible ways to select n covariates from the N_j covariates forming domain j. Thus, for m selected domains and n selected covariates within each domain, the total number of possible combinations is $Q\left(m, n\right) = \sum_{h=1}^{H} \prod_{j \in G_h} C\left(N_j, n\right)$, where G_h is the index set of selected covariate domains for a specific combination of domains $h \in \{1, \ldots, H\}$. Assuming that the number of observed covariates is the same for all domains, $N_j = N$, the number of possible combinations (Q) grows faster than $O(N^M)$. Consequently, even for small numbers of selected domains, m, and selected covariates within domains, n, it is nearly impossible to evaluate the bias reduced by each covariate combination within reasonable computational time.

Instead, we assess the average bias reduction achieved by increasing the number of covariates within domains and do so by randomly drawing sets of n covariates from the N_j covariates of each sampled domain. With respect to the covariate domains, we explore all possible $H = C(M, m)$ combinations, that is, no random sampling is involved at the domain level. Thus, the sampling procedure for a given number m of domains and a given number n of covariates per domain represents a stratified sampling strategy where the strata are formed by the domains within a specific domain combination G_h:

i. Given a selected number of domains m from the M available domains, generate all $H = C(M, m)$ combinations of domains and enumerate them from $h = 1, \ldots, H$. Then, for a selected combination h, let G_h denote the set of the m domain indices $j \in \{1, \ldots, M\}$. For example, with $M = 12$ and $m = 3$ for the SCS data set we obtain $H = C(12, 3) = 220$ combinations of domains: $G_1 = \{1, 2, 3\}$, $G_2 = \{1, 2, 4\}$, $\ldots$, $G_{220} = \{10,11, 12\}$.
ii. For each domain index set G_h, independently and randomly select n covariates without replacement from each domain $j \in G_h$.
iii. Repeat the random sampling of covariates in Step ii for each combination G_h r times, resulting in $r \times H$ sampled data sets of m domains and n covariates per domain.

Steps (i) to (iii) are independently repeated for each (m, n)-pair of m domains and n covariates per domain. For the SCS data set, we systematically selected $m = 1, \ldots, 12$ domains of the $M = 12$ covariate domains, and sampled up to 7 covariates per domain, $n = 1, 3, 5, 7$. For the ECLS-K data set we selected $m = 1,\ldots,10$ domains of the $M = 10$ covariate domains, and sampled up to 9 covariates per domain, $n = 1, 3, 5, 7, 9$. We were able to sample 9 instead of 7 covariates since each covariate domain of the ECLS-K data set contains more covariates than the 12 domains of the SCS data.[1] For the SCS data set we replicated the sampling for a given (m, n)-pair $r = 300$ times; for the ECLS-K data set we used $r = 100$ replications due to the larger number of observations that require much more computational time in estimating the selection model and treatment effect.

For each sampled data set, we estimated the PS-logit from the PS model, which we selected using stepwise forward regression (with AIC as inclusion criterion, i.e., a covariate was included if AIC decreased). Although this is not a commonly suggested covariate selection strategy for model selection in practice (one better uses balance criteria for

[1]If a domain contained less than n covariates (not all domains of both data sets actually consist of at least 7 or 9 covariates), all covariates were sampled. Thus, sampling 7 or 9 covariates does not necessarily mean that 9 covariates were actually available for each covariate domain. In fact, the real average numbers of covariates drawn from each domain are $n^* = 1, 3, 4.6, 5.8, 7$ at the designated levels of $n = 1, 3, 5, 7, 9$ for the SCS data, and $n^* = 1, 3, 5, 6.9, 8.4$, for $n = 1, 3, \ldots, 9$ for the ECLS-K data, indicating that the deviations from the designated levels are minor.

specifying the PS model), it was at least a feasible strategy for our simulation. Alternatively, we could have included all the sampled covariates in our PS model, but for large samples of covariates this would have resulted in a severe overfitting and a lack of overlap between treatment and control subjects. We computed the PS-adjusted treatment effect, $\hat{\tau}_{PS}$, via PS regression as described in Equation (2). For assessing a sampled covariate set's success in removing selection bias, we compare the PS-adjusted treatment effect, $\hat{\tau}_{PS}$, to the treatment effect obtained from the randomized experiment, $\hat{\tau}_{RE}$. We use the percent of remaining bias after PS-adjustment as a standardized measure:

$$\hat{b} = \frac{\hat{\tau}_{PS} - \hat{\tau}_{RE}}{\hat{\tau}_{PF} - \hat{\tau}_{RE}} \times 100.$$

The percentage of remaining bias is calculated as the ratio between the remaining bias in the adjusted treatment effect, $\hat{\tau}_{PS} - \hat{\tau}_{RE}$, and the initial bias of the prima facie effect, $\hat{\tau}_{PF} - \hat{\tau}_{RE}$. The prima facie effect, $\hat{\tau}_{PF}$, is the unadjusted difference between the treatment and control group's average outcome. Since we have a benchmark effect estimate from a randomized experiment only for the SCS data, this definition of remaining bias does not directly apply to the ECLS-K data set for which we do not have a corresponding randomized experiment. Instead we use the estimated treatment effect when all available 208 covariates are used for estimating the PS (this does not mean that all 208 covariates actually entered the PS model, but in selecting the PS model, balance was optimized for all 208 available covariates). Thus, remaining bias refers to the situation where all 208 covariates were available for estimating the unknown PS. Given the large set of covariates, including two waves of pretest measures of the outcomes and proximal pretest measures (i.e., teacher assessments of students), we presume that the benchmark estimate is close to what we would have obtained from a randomized experiment. However, even if this presumption is not true, the remaining bias $\hat{b}$ still allows us to assess the relative importance of the number of domains and the number of measurements within domains, given that the benchmark estimate is the least biased estimate we can obtain. It is also important to note that the benchmark estimates, which we use for assessing the initial and remaining bias, do not represent the true causal effects because they are estimates on their own and thus are subject to sampling error. Although the standard errors of both the benchmark estimate and the PS-adjusted treatment effect for the SCS data are rather large in comparison to the initial bias, including all covariates removes in both outcomes almost 100% bias so that sampling error is not an issue here.[2] Since the corresponding standard errors for the ECLS-K data are rather small, sampling error is not of major importance.[3] Moreover, the sampling error does not compromise the comparison of results across selected numbers of domains and covariates because we did not sample subjects in our simulations (only domains and covariates were sampled).

[2]For the SCS data, the initial bias in the vocabulary outcome amounts to .82 points (.24 *SD*), the standard error of the benchmark estimate from the randomized experiment is .39 points and .49 points for the PS-adjusted treatment effect when all covariates are considered for inclusion in the PS model. The math outcome has an initial bias of .95 points (.30 *SD*) and the corresponding standard error for the benchmark estimate and the PS-adjusted estimate are .36 and .39 points, respectively.

[3]For the ECLS-K reading outcome, the initial bias with respect to the model where all covariates were considered amounts to –10.6 points (–.79 *SD*), the standard error of the benchmark estimate is 1.09 points. For the math outcome, the initial bias is –6.6 points (–.74 *SD*), the standard error of the benchmark estimate is .85.

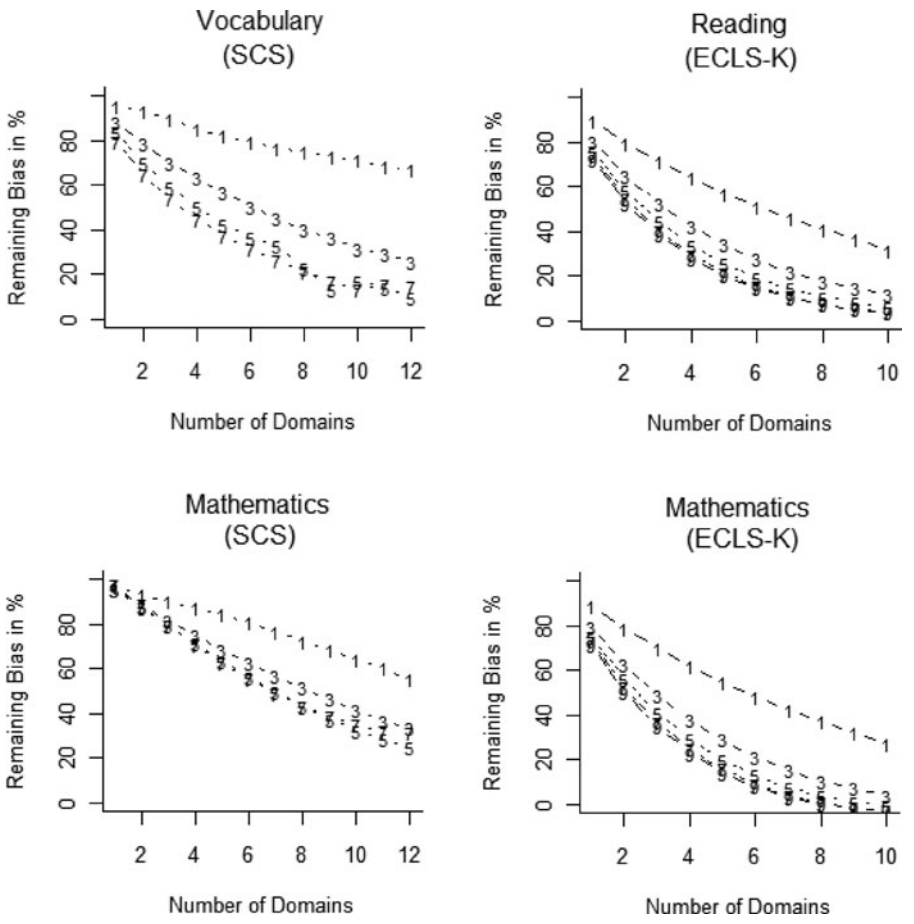

Figure 2. Average remaining bias in percent from all covariate samples (including samples with critical covariates) of the Shadish et al. and ECLS-K data. Line numbers 1, 3, 5, and 7 represent the number of covariates sampled from each domain.

Finally, the average remaining bias $\bar{b}_{m,n}$ for a specific (m, n)-pair of domains and covariates is obtained by averaging the remaining biases across the corresponding $r \times H$ sampled data sets:

$$\bar{b}_{m,n} = \frac{1}{rH} \sum_{h=1}^{H} \sum_{i=1}^{r} \hat{b}_i.$$

An alternative estimator of the average remaining bias would take the stratified nature of the sample into account, that is, by weighting stratum-specific estimates of the remaining bias by the number of covariates (Lohr, 1999). Because each covariate domain contains covariates that are highly correlated with each other and are not exhaustive with respect to the universe of all possible covariates, there is no clear rationale for putting more weight on domains represented by a larger number of observed covariates. Assuming a hypothetical superpopulation of covariates, we may argue that the number of covariates that could have

been measured for each covariate domain approaches infinity. In any case, the results we present below differ only slightly from the weighted averages of remaining bias.

Because there is variation in the bias remaining across replications, we also report the 10% and 90% quantiles (i.e., the lower and upper deciles) of the distribution of remaining bias. The deciles describe the spread of remaining bias we face when covariates are blindly sampled, as in our simulation. Sometimes, we might be lucky and have access to a set of covariates including those that are rather effective in removing selection bias, but sometimes the data set will have poorer covariates capable of removing only a small fraction of the total selection bias, however large the number of covariates. Indeed, it is even possible for the bias to increase if a specific adverse set of covariates is observed that we briefly outline in the discussion section.

We ran two sets of simulations. One included all the covariates in the covariate sampling pool. The other excluded from the pool the previously discussed critical covariates that remove a large part of the overall selection bias on their own, that is, all the covariate items that belong to the covariates identified by Steiner et al. (2010) for the SCS data set and Hallberg (2013) for the ECLS-K data set. The obviously worse scenario without critical covariates helps to investigate the extent to which increasing the number of covariate domains and of covariates within each domain protects against selection bias in this particularly adverse covariate-selection context.

RESULTS

Bias Reduction When All Covariates Are Sampled, Including the Critical Ones

The relevant simulation results are summarized in the plots of Figure 2. They display the average percent of remaining bias $\bar{b}_{m,n}$ (on the ordinate) as a function of the number of domains (on the abscissa) and number of sampled covariates per domain (as separate lines). Not surprisingly, the plots indicate that bias reduction increases with the number of domains and of covariates per domain.

Holding constant the number of domains, the average amount of bias removed by an additional covariate diminishes as the number of covariates per domain increases. While increasing the number of covariates from one to three has a strong effect on bias reduction, increasing it from five to seven or from seven to nine results in an almost negligible average reduction of the selection bias (i.e., in Figure 2, for the SCS data, the lines with five and seven covariates are rather similar; for the ECLS-K data, the lines with seven and nine covariates do not even reliably differ).

With regard to the relative importance of domains and covariates per domain, increasing the number of domains is more important. This is best seen from the first five columns in Tables 1 through 4. In Table 3, for the SCS vocabulary outcome, 45% of the bias remains on average when sampling seven domains and utilizing three within-domain covariates (21 covariates in total), while 54% bias is left when sampling three domains and seven covariates (also a total of 21 covariates). The effect is even stronger for the ECLS-K reading outcomes shown in Table 1, where nine domains with three covariates removes 86% of the average bias while three domains with nine covariates removes 61%. So holding the total number of covariates constant, covariate sets with more construct domains do better than covariate sets with fewer domains but more covariates per domain.

Table 1. Average remaining bias in percent for the reading outcome of the ECLS-K data

Number of Domains (m)	All Samples					Samples Without Critical Covariates				
	Number of Covariates Per Domain (n)					Number of Covariates Per Domain (n)				
	1	3	5	7	9	1	3	5	7	9
1	89 [63, 102]	80 [20, 102]	76 [19, 102]	74 [18, 102]	72 [18, 102]	95 [88, 102]	88 [59, 102]	84 [49, 102]	82 [20, 102]	80 [20, 102]
2	80 [28, 102]	65 [11, 101]	58 [4, 100]	55 [0, 99]	53 [0, 98]	90 [66, 102]	79 [26, 101]	72 [21, 100]	68 [21, 100]	66 [21, 99]
3	72 [26, 100]	52 [9, 95]	44 [2, 91]	40 [0, 87]	39 [0, 85]	86 [63, 101]	70 [23, 98]	61 [21, 95]	57 [20, 93]	54 [20, 91]
4	64 [24, 98]	42 [7, 88]	34 [1, 82]	30 [0, 76]	28 [0, 71]	82 [57, 100]	62 [23, 92]	53 [20, 86]	48 [20, 82]	45 [19, 79]
5	57 [22, 95]	34 [6, 81]	26 [1, 72]	21 [0, 66]	20 [0, 61]	78 [49, 98]	56 [23, 86]	45 [20, 79]	40 [19, 74]	38 [19, 68]
6	51 [21, 91]	27 [4, 68]	19 [1, 57]	15 [0, 51]	15 [0, 47]	75 [29, 96]	50 [22, 80]	39 [20, 73]	34 [19, 66]	32 [18, 61]
7	46 [19, 86]	22 [4, 54]	14 [1, 36]	11 [0, 22]	10 [0, 21]	72 [28, 94]	45 [22, 74]	33 [19, 63]	28 [18, 57]	27 [18, 50]
8	41 [18, 74]	18 [3, 41]	11 [1, 22]	7 [0, 20]	7 [0, 19]	67 [27, 92]	41 [22, 64]	28 [19, 49]	23 [18, 43]	22 [17, 41]
9	37 [16, 65]	14 [3, 24]	8 [1, 18]	5 [1, 15]	5 [1, 12]	66 [28, 92]	37 [21, 54]	24 [19, 40]	19 [17, 21]	19 [17, 20]
10	31 [13, 57]	12 [2, 19]	6 [2, 12]	4 [0, 9]	3 [1, 8]					

Note. The 10% and 90% percentiles of the remaining bias are given in the square brackets. For the samples without critical covariates the maximum number of domains is nine since the 10th domain represents the critical pretest domain (from which no covariates were sampled).

Table 2. Average remaining bias in percent for the mathematics outcome of the ECLS-K data

	All Samples					Samples Without Critical Covariates				
	Number of Covariates Per Domain (n)					Number of Covariates Per Domain (n)				
Number of Domains (m)	1	3	5	7	9	1	3	5	7	9
1	89 [55, 102]	79 [17, 102]	75 [16, 102]	73 [15, 102]	71 [15, 102]	95 [88, 102]	88 [52, 102]	84 [45, 102]	82 [17, 102]	80 [17, 102]
2	79 [25, 102]	63 [4, 101]	56 [−2, 99]	52 [−4, 99]	50 [−4, 98]	90 [58, 102]	77 [27, 101]	70 [17, 100]	66 [16, 99]	64 [16, 98]
3	70 [23, 101]	49 [2, 96]	41 [−3, 92]	37 [−5, 88]	35 [−5, 86]	85 [54, 102]	68 [19, 98]	59 [16, 95]	54 [15, 92]	51 [15, 91]
4	62 [20, 99]	38 [0, 89]	29 [−3, 82]	25 [−5, 77]	23 [−5, 69]	80 [50, 101]	59 [18, 92]	49 [15, 87]	44 [14, 83]	41 [13, 80]
5	55 [18, 96]	29 [0, 81]	20 [−3, 71]	16 [−5, 63]	14 [−6, 58]	76 [41, 99]	52 [17, 87]	41 [14, 80]	35 [13, 74]	33 [12, 65]
6	48 [16, 91]	21 [−1, 66]	13 [−4, 52]	9 [−5, 46]	8 [−6, 40]	72 [30, 97]	45 [17, 81]	33 [13, 72]	28 [12, 64]	26 [11, 58]
7	42 [14, 86]	15 [−1, 44]	8 [−4, 25]	4 [−5, 16]	3 [−6, 15]	69 [29, 94]	39 [16, 73]	27 [13, 60]	22 [11, 52]	20 [10, 45]
8	37 [12, 73]	11 [−2, 33]	4 [−4, 16]	1 [−5, 13]	0 [−6, 13]	64 [28, 92]	34 [16, 61]	21 [12, 41]	17 [11, 35]	16 [10, 33]
9	32 [12, 62]	7 [−2, 18]	1 [−4, 10]	−2 [−5, 4]	−2 [−5, 3]	64 [28, 94]	30 [16, 45]	17 [11, 31]	13 [10, 15]	12 [10, 14]
10	27 [10, 48]	4 [−2, 10]	−1 [−4, 3]	−3 [−5, −1]	−3 [−5, −1]					

Note. The 10% and 90% percentiles of the remaining bias are given in the square brackets. For the samples without critical covariates the maximum number of domains is nine since the 10th domain represents the critical pretest domain (from which no covariates were sampled).

Table 3. Average remaining bias in percent for the vocabulary outcome of the SCS data

	All Samples				Samples Without Critical Covariates			
	Number of Covariates Per Domain (n)				Number of Covariates Per Domain (n)			
Number of Domains (m)	1	3	5	7	1	3	5	7
1	96 [82, 109]	88 [64, 113]	84 [60, 113]	80 [49, 111]	97 [85, 110]	91 [70, 114]	87 [65, 115]	84 [60, 114]
2	93 [70, 112]	79 [46, 110]	70 [36, 105]	65 [30, 102]	95 [76, 113]	84 [56, 113]	77 [49, 110]	73 [41, 108]
3	90 [62, 113]	70 [34, 106]	59 [22, 96]	54 [15, 92]	94 [70, 115]	77 [48, 110]	69 [38, 102]	64 [31, 99]
4	85 [52, 113]	63 [24, 100]	50 [10, 88]	45 [3, 85]	91 [63, 115]	72 [41, 105]	63 [30, 96]	58 [23, 94]
5	82 [44, 113]	57 [16, 94]	42 [2, 82]	37 [−6, 79]	89 [60, 115]	67 [36, 100]	57 [24, 91]	53 [17, 90]
6	79 [39, 112]	51 [9, 89]	36 [−5, 77]	31 [−13, 75]	87 [57, 115]	62 [31, 95]	52 [19, 87]	50 [13, 87]
7	77 [36, 112]	45 [3, 84]	33 [−12, 73]	26 [−20, 71]	85 [54, 116]	57 [27, 90]	54 [15, 83]	47 [10, 84]
8	75 [33, 111]	40 [−2, 80]	23 [−17, 69]	21 [−26, 67]	84 [51, 115]	54 [24, 86]	45 [11, 80]	44 [7, 81]
9	73 [31, 110]	36 [−6, 75]	13 [−22, 66]	17 [−31, 65]	82 [49, 114]	52 [22, 82]	41 [8, 76]	43 [6, 78]
10	71 [29, 109]	32 [−9, 71]	17 [−25, 62]	13 [−34, 61]	81 [46, 114]	48 [18, 77]	38 [6, 71]	40 [5, 74]
11	68 [26, 106]	29 [−13, 69]	14 [−32, 59]	15 [−36, 59]	78 [42, 112]	45 [16, 75]	36 [4, 67]	34 [−4, 71]
12	67 [22, 104]	26 [−18, 67]	9 [−25, 54]	15 [−37, 62]				

Note. The 10% and 90% percentiles of the remaining bias are given in the square brackets. For the samples without critical covariates the maximum number of domains is 11 since the 12th domain represents the critical vocabulary pretest domain (from which no covariates were sampled).

Table 4. Average remaining bias in percent for the mathematics outcome of the SCS data

	All Samples				Samples Without Critical Covariates			
	Number of Covariates Per Domain (n)				Number of Covariates Per Domain (n)			
Number of Domains (m)	1	3	5	7	1	3	5	7
1	97 [85, 115]	96 [75, 120]	96 [69, 122]	97 [71, 129]	100 [85, 115]	103 [85, 121]	106 [86, 123]	109 [88, 130]
2	93 [73, 115]	89 [20, 122]	88 [1, 128]	89 [−7, 134]	98 [80, 115]	101 [79, 123]	106 [83, 130]	110 [84, 136]
3	90 [52, 116]	82 [9, 122]	80 [−6, 129]	80 [−14, 135]	97 [77, 116]	99 [72, 124]	105 [78, 133]	109 [79, 140]
4	87 [33, 116]	75 [5, 121]	71 [−8, 128]	71 [−15, 134]	96 [75, 117]	95 [65, 124]	102 [72, 133]	107 [75, 141]
5	84 [26, 116]	69 [3, 118]	64 [−9, 124]	63 [−15, 129]	95 [73, 118]	92 [58, 123]	98 [66, 132]	104 [70, 140]
6	81 [23, 115]	63 [2, 114]	56 [−10, 118]	56 [−15, 123]	93 [69, 117]	88 [52, 120]	94 [60, 129]	100 [66, 137]
7	77 [20, 113]	57 [1, 109]	50 [−10, 110]	49 [−14, 116]	90 [63, 116]	83 [46, 117]	89 [54, 125]	97 [61, 134]
8	73 [19, 111]	52 [0, 103]	43 [−10, 102]	43 [−14, 108]	86 [56, 113]	78 [41, 114]	84 [48, 120]	93 [57, 130]
9	68 [17, 107]	46 [−1, 96]	38 [−11, 92]	39 [−13, 98]	82 [51, 110]	73 [36, 109]	79 [42, 114]	89 [53, 125]
10	64 [15, 104]	42 [−2, 88]	33 [−12, 80]	35 [−13, 86]	78 [47, 107]	67 [31, 103]	73 [37, 107]	85 [49, 119]
11	60 [12, 99]	37 [−2, 77]	28 [−14, 71]	32 [−12, 74]	75 [44, 102]	60 [26, 96]	65 [28, 97]	79 [40, 113]
12	55 [5, 95]	33 [−2, 66]	25 [−20, 66]	32 [−11, 69]	70 [43, 99]	51 [20, 78]	49 [17, 88]	57 [43, 76]

Note. The 10% and 90% percentiles of the remaining bias are given in the square brackets.

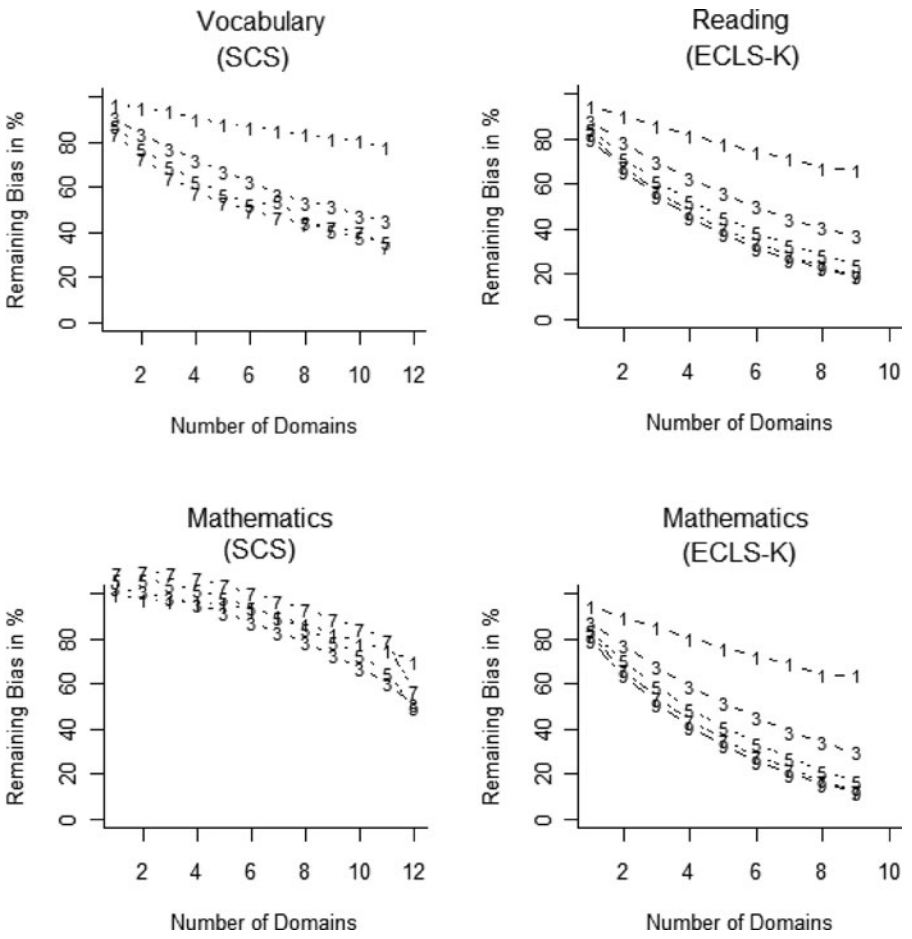

Figure 3. Average remaining bias in percent from covariate samples not including any critical covariates. Line numbers 1, 3, 5, 7, and 9 represent the number of covariates sampled from each domain.

Bias Reduction When Critical Covariates Are Unobserved

Figure 3 and the second block of columns in Tables 1 to 4 show the average remaining bias after the critical covariates have been deliberately removed.[4] But otherwise the same domains are analyzed. As the plots indicate, for three of the four analyses, the same basic pattern of results is obtained as above, the SCS math outcome being the exception. That is, (a) the number of covariate domains and the number of covariates per domain both make a difference, (b) bias reduction diminishes on average as the number of domains and covariates within domains increases, and (c) very little or no bias is reduced with one covariate per domain.

However, the overall level of bias reduction is less than when all the covariates could potentially contribute to the PS. With all domains and the maximum number of covariates

[4]With the exemption of math outcome of the SCS data, removing the critical covariates reduces the number of domains by one because some of the covariates actually form an entire domain.

per domain, but without the critical covariates, the remaining bias is 34% on average for vocabulary and 57% for math with the SCS data set, and it is 19% for reading and 12% for math in ECLS-K. These levels of correspondence are strikingly lower for the SCS data whose benchmark comes from an experiment than for the ECLS-K data whose benchmark does not. Within the limits of our simulation study, it is not easy to contend that a heterogeneous set of quite well-measured domains will result in a tolerable level of remaining bias when the crucial covariates are not in the item sample. Bias will be reduced, as here, but certainly for SCS it was not to a tolerable level.

Effectiveness of Bias Reduction at the 10th and 90th Percentiles

The results above only hold when averaged across all the simulated random draws of domains and of covariates of a given size (m and n, respectively). But the domains and covariates sampled for each replication differ considerably and, thus, might vary in how well they control for bias. The numbers in square brackets in Tables 1 through 4 present the bias of the 10th and 90th percentiles of the remaining bias and provide an estimate of the variability of the remaining bias. Percentages over 100% indicate that bias actually increased compared to the unadjusted prima facie effect, and negative percentages indicate an over-adjustment.

The results show consistently large variability between the 10th and 90th percentiles, indicating that bias reduction depends heavily on the particular set of domains of a given size and on the particular set of within-domain covariates of a given size that happens to get selected. Because domains and covariates within domains differ in their ability to reduce bias, the success in bias reduction attributed to sampling as many heterogeneous domains as possible cannot be dissociated from which specific domains and covariates are selected. In actual research practice, it is rare to have no clues to the actual selection process, even if full knowledge is denied us. Those clues need to be used to guide the choice of domains and covariates instead of relying, as here, on a blind choice process.

Another striking result from the 10th and 90th percentile statistics is that the variability of remaining bias is lower for the ECLS-K data than for the SCS data when the number of domains and covariates within domains is large. Consider the case with eight domains and seven covariates per domain: for the reading outcome of the ECLS-K data set, the remaining bias ranges from 0% to 20% and for math from –5% to 13%. The corresponding values for the SCS data are –26% to 67% and –14% to 108%. Although the percentiles for the ECLS-K data set might be acceptable to many persons, the percentiles of the SCS data could never be.

DISCUSSION

Summary of the Results

First, holding constant the number of covariates within a domain, the results from both data sources show that the amount of bias reduction increased with the number of construct domains. However, the increase was not linear because the influence of adding domains diminished as the number of domains increased. These diminishing returns presumably reflect the fact that less bias remains to be reduced as the number of domains increases.

Second, holding the number of domains constant, both data sets showed that the amount of bias reduction increased as the number of covariates in a domain increased. This relationship also involved diminishing returns.

Third, with both data sets most bias reduction was observed with the maximal number of domains and at least five covariates per domain. When all the covariates were included in the sampling pool, the bias remaining was trivial for both ECLS-K outcomes (3% with nine covariates per domain) and perhaps also for the SCS vocabulary outcome (15% with seven), although as much as 32% of the initial bias remained for the math outcome.

Fourth, the number of covariate domains was more important than the number of covariates per domain—a replication of the earlier finding that the nature of the covariates is more important for bias reduction than is the reliability with which they are measured as long as reliability is above the often-accepted .60 level (Steiner et al, 2011). For both data sets, these findings were confirmed by a formal analysis of variance (not reported here), where the variance of the remaining bias was modeled as a function of the number of domains and the number of covariates within domains (including quadratic and interaction terms).

Omitting the critical covariates that remove almost all of the bias on their own hardly affected the functional form of bias reduction compared to when all covariates were in the analysis: all effects were positive and characterized by diminishing returns. However, the total bias reduction was noticeably lower when the effective covariates were excluded. Some will see the bias reduction achieved as better than nothing, while others will see it as a harbinger of dangerously misleading causal conclusions.

Finally, we sought to replicate data patterns across two quite different data sets. Although the overall data trends are quite similar, there are some striking differences in the ability to reduce bias. In general, the ECLS-K results were superior in the amount of bias reduced on average, and its bounds were tighter between the 10% and 90% percentiles for the amount of bias remaining. The best scenario for maximizing bias removal in ECLS-K was with 12 domains, seven covariates per domain, and no excluded critical covariates. Then, only 3% of the bias remained for both math and vocabulary. In the closest corresponding SCS scenario, only 15% of the bias remained for the vocabulary outcome in SCS, although more than 30% did for math. This last seems inconsistent with previous reports that almost 100% of the selection bias for math was removed when all covariates were used (Shadish et al., 2008; Steiner et al., 2010). One likely reason for the difference touches on the level of covariate measurement. In our analyses of the SCS data set, a within-domain covariate might represent a single item from within, say, the original 15-item math proxy pretest. But in the earlier studies, only test totals were used as covariates and these have considerably higher reliability than a single item. Another reason is that sampling seven covariates from all 12 domains still resulted in sampling only 84 of all the 156 covariates available in SCS. This means that the present study investigates the *average* bias reduction due to *randomly sampled subsets* of covariates, whereas earlier studies investigated the bias reduction due to all the available covariates. In contrast, the analyses of ECLS-K used more of the available covariates because so many of them were scale totals rather than single items from within multi-item constructs.

All these results are in line with the findings of Shadish et al. (2008), Steiner et al. (2011), and Steiner et al. (2010), who demonstrated that the two most important factors for removing (almost) all the selection bias are (a) the choice of an appropriate set of covariates and (b) their reliable measurement. Our results highlight that failing to measure the most crucial covariates or covariate domains for removing selection bias may result in a remaining bias of nonnegligible size. In particular, selecting only a few domains reduces the chance to cover the most crucial domains for bias reductions. We also showed

that if covariate domains are unreliably measured (as for the SCS data where we used single items of multi-item constructs) bias reduction is attenuated. Selecting only a few single items from an important covariate domain might severely compromise a domain's reliability.

Implications of the Results

This study sought to understand the consequences of increasing the heterogeneity and reliability of the construct domains used to control for selection bias. It did so under two conditions. The first is when no explicit theories of selection were used for choosing study covariates. Instead, chance was used to select covariates from an otherwise large pool that, in one case, was designed to contain good covariates to assess the study selection process, while in the other case the item pool came from a longitudinal survey that was not designed to measure the selection process into being retained a grade. The second condition was when the pool of covariates was deliberately compromised by omitting the most important covariates for measuring selection into math or vocabulary training or into grade retention. We have demonstrated that the number, heterogeneity, and reliability of covariates clearly matter.

But even so, our blind covariate choice models led to considerable variability in the amount of bias reduction achieved for a fixed size of domains and covariates. Since different sets of covariates of the same size differ considerably in their ability to reduce bias, no guarantee of unbiased causal inference can be offered when using blind sampling. The best advice is to get the number of domains as high as possible and the number of covariates within domains to be at least five. There should be as little curtailment of the heterogeneity of the covariate sampling pool as possible.

In real research applications, explicit attention to covariate choice is needed at the design stage. Fortunately, it will be very rare not to have some knowledge of selection. In the two cases just analyzed, we presume that almost anyone would have hypothesized that kindergarten retention decisions are made on the basis of student academic performance and teacher judgments of such performance. These constructs would have been part of anyone's desired selection model. Also, consensus is presumably widespread that self-selection into vocabulary and mathematics is a function of prior knowledge of each, of preference for one over the other, and, in particular, of dislike for math. However imperfect, such clues to selection should be utilized so as to increase the odds of selecting the better covariates into the pool for analysis. However, there is still a role for the blind collection of many other heterogeneous and modestly well-measured covariates. It will help to protect against some of the unobserved selection elements not identified in the pre-intervention analysis or not capable of quality measurement. It is a modest insurance policy against hidden bias and not a substitute for the overt strategy of selection specification(s) guiding the choice of covariates. Even hundreds of covariates cannot guarantee that the bias in the treatment effect can be reduced to a negligibly small size if they do not tap into the domains crucial for bias reduction. Theory and direct investigations of the selection process are required to inform us about whether the available covariates might succeed in removing most of the bias. Instead of blindly assuming strong ignorability, applied researchers need to defend this assumption on theoretical grounds, otherwise they should abstain from a causal interpretation of the estimated effects.

Few past applications have deliberately tailored covariate selection to explicit and critically appraised analyses of possible selection processes supplemented by other covariates

from multiple domains that might also be related to selection and outcome. In job training, it has been rare for within-study comparisons to reduce all selection bias (Glazerman, Levy, & Myers, 2003), but there have not been measures of the motivation to seek job training, of the informal pressures that family and friends put on individuals to attend such training, of the formal pressures exercised by social workers and the courts, or of the changes in local availability of training opportunities. The impression is that covariates were selected more for their links to the outcome than to selection—for example, measures such as past employment and wage history—and more for their availability in data archives than for their theoretical fit to likely true selection processes. For instance, Peikes, Moreno and Orzol (2008) also found that they could not recreate an experimental estimate, but they also frankly noted that their list of covariates failed to include plausible motivational covariates and included only structural attributes that, they contended, analysts might be inclined to use as comprehensive selection controls. Yet a crucial class of theoretically identified covariates was clearly missing from the attempt to control for selection (see Shadish, Steiner, & Cook, 2012). Finally, in a voting context, Arceneaux, Gerber, and Green (2010) failed to recreate the experimental estimate obtained in a study of the effects of a persuasive message to turn out to vote. Here selection depends on the propensity to be persuaded to vote, but the covariates used assess the propensity to vote. They are not the same, and each would require a different set of covariates. The authors again argued that the covariates they used were those that analysts might use to try to control for selection; but they are not those that careful analysis suggests should be used.

In our simulation, bias sometimes increased even after covariates were applied to reduce it. Some of the estimates of more than 100% bias remaining might be due to sampling error in both the benchmark estimate and the randomly drawn data. But even so, there are more than we would expect by chance and they are systematic in distribution. For both ECLS-K outcomes and for vocabulary in the SCS data set, the fewer the domains and covariates within them, the more frequently bias increased. However, with more than two domains, additional bias hardly occurs. The exception is for math in SCS where additional bias is more prevalent with fewer domains and covariates but does not disappear until the number of domains exceeds nine and the number of covariates per domain exceeds three. It is discomforting that bias can increase because of attempts to reduce it, even if this is less likely with data sets containing numerous and heterogeneous covariates.

There are theories that explain the possibility of bias-inducing and bias-amplifying covariates (Greenland, 2003; Pearl, 2009, 2010; Wooldridge, 2005). Instrumental and near instrumental variables that are strongly related to selection but that are unrelated, or only weakly related, to the outcome cannot induce bias on their own, but they can amplify any bias remaining after a regression or PS adjustment. In contrast, collider variables that may be unrelated to selection and the outcome may be related to unobserved variables influencing treatment selection or outcome, and then these collider variables can induce bias on their own (M-bias; Greenland, 2003). To completely identify bias-amplifying (near) instrumental variables and bias-inducing colliders, a rather strong substantive theory is needed about the data-generating selection and outcome models. Lacking such strong theories, the practical question is: how harmful can such covariates actually be? The analyses here suggest that, although (near) instruments and colliders might be harmful, an increase in bias (with respect to the initial bias) seems less likely as the size of the covariate set increases. Even so, this does not mean that no bias-amplifying or bias-inducing effects are present; they might just be weak because pure instruments and colliders are presumably rare in practice and because including other interrelated covariates in the PS model may block or weaken bias induction and amplification.

A PS analysis based on all observed pretreatment covariates often remains the best a researcher can do when strong theories about the selection and outcome model are missing. However, PS analyses should always be supplemented by sensitivity analyses that assess the effect of violating the ignorability assumption in specific applications (Rosenbaum, 2002). These sensitivity tests have their own assumptions, of course, so that, although they are useful, they are still not definitive. Analysts of observational data have to live with the reality that the process of selection into treatment is almost never fully known. However, design elements such as nonequivalent outcomes, multiple comparison groups, or unaffected cases of the treatment group can be used to probe potential violations of the strong ignorability assumption (Rosenbaum, 1984; Shadish et al., 2002; Steiner, 2012).

Limitations of the Results

The generalizability of results is limited because of the use of only two data sets. More would be desirable. Particularly desirable would be data sets where the selection process into treatment is more opaque than it is for kindergarten retention and for self-selection into math versus vocabulary training. We tried to simulate a more opaque setting by omitting the identified crucial covariates, but this is not the same as having data from real applications where the true selection processes are less accessible. Another limitation comes from the different covariate sampling designs in the two studies. Optimal conditions would be if each replicate had a data structure with many single items per construct, many constructs per domain, and many domains. Instead, the SCS data set had many items for most constructs, many domains, but few constructs per domain, while ECLS-K had no single items for most constructs, but many domains and many constructs per domain. Future research may benefit from a better-planned covariate structure than the opportunistic ones of which we took advantage. A final limitation worth mentioning is the absence of an experimental benchmark for the ECLS-K data. Instead, one has to accept the argument that having multiple pretest measures of student achievement and of teacher assessments of such achievement is sufficient for removing (almost) all the selection bias together with all the other covariates. An experimental benchmark would definitely be more convincing, although it is clearly difficult to randomly assign to kindergarten retention.

ACKNOWLEDGMENTS

This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R305D100033 (Peter M. Steiner and Thomas D. Cook) and R305D120005 (Peter M. Steiner). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

Arceneaux, K., Gerber, A. S., & Green, D. P. (2010). A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociological Methods & Research*, *39*(2), 256–282.

Beck, A. T., & Beck, R. W. (1972). Screening depressed patients in family practice: A rapid technic. *Postgraduate Medicine*, *51*, 81–85.

Clark, M. H. (2000). *A laboratory experiment comparing assignment methods using propensity scores* (Unpublished master's thesis). University of Memphis, Memphis, TN.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, A*, *35*, 417–446.

Collingwood, R. (1940). *An essay on metaphysics*. Oxford, UK: Clarendon Press.

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*(2), 636–654.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of the pretest as a covariate, unreliable measurement and mode of data analysis. *Psychological Methods*, *15*(1), 56–68.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator. *The Journal of Human Resources*, *XLI*(2), 319–345.

Educational Testing Service. (1962). *Vocabulary Test II (V-2). Kit of factor referenced cognitive tests*. Princeton, NJ: Author.

Educational Testing Service. (1993). *Arithmetic Aptitude Test (RG-1). Kit of factor referenced cognitive tests*. Princeton, NJ: Author.

Faust, M. W., Ashcraft, M. H., & Fleck, D. E. (1996). Mathematics anxiety effects in simple and complex addition. *Mathematical Cognition*, *2*, 25–62.

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

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26–42.

Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper #123). Madison, WI: Institute for Research on Poverty, University of Wisconsin.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding versus collider-stratification bias. *Epidemiology*, *14*, 300–306.

Hallberg, K. (2013). *Identifying conditions that support causal inference in observational studies in education: Empirical evidence from within study comparisons* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (#3563726), http://gradworks.umi.com/35/63/3563726.html

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–970.

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. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29.

Li, K. H. (1988). Imputation using Markov Chains. *Journal of Statistical Computation and Simulation*, *30*, 57–79.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (UAI, 2010), 425–432. Retrieved from http://event.cwi.nl/uai2010/papers/UAI2010_0120.pdf

Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity score matching: A note of caution for evaluators of social programs. *The American Statistician*, *62*, 222–231.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, *79*, 41–48.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag.

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. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

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.

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

Schafer, J. L., & Kang, J. (2008). Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 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*, 1334–1343.

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

Shadish, W. R., Steiner, P. M., & Cook, T. D. (2012). A case study about why it can be difficult to test whether propensity score analysis works in field experiments. *Journal of Methods and Measurement in the Social Sciences*, *3*(2), 1–12.

Steiner, P. M. (2012). Using design elements for increasing the severity of causal mediation tests. Commentary on Hong, G., & Nomi, T. (2012). Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness*, *5*, 296–298.

Steiner, P. M., & Cook, D. L. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 237–259). New York, NY: Oxford University Press.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213–236.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250–267.

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.

Wooldridge, J. M. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory*, *21*, 1026–1028.