Published in final edited form as:

Prev Sci. 2018 April ; 19(3): 274–283. doi:10.1007/s11121-016-0732-6.

Pretest Measures of the Study Outcome and the Elimination of Selection Bias: Evidence from Three Within Study Comparisons

Kelly Hallberg¹, Thomas D. Cook^{2,3}, Peter M. Steiner⁴, M. H. Clark⁵

¹University of Chicago, Chicago, IL, USA

²Mathematica Policy Research, Chicago, IL, USA

³Northwestern University, Evanston, IL, USA

⁴University of Wisconsin, Madison, WI, USA

⁵University of Central Florida, Orlando, FL, USA

Abstract

This paper examines how pretest measures of a study outcome reduce selection bias in observational studies in education. The theoretical rationale for privileging pretests in bias control is that they are often highly correlated with the outcome, and in many contexts, they are also highly correlated with the selection process. To examine the pretest's role in bias reduction, we use the data from two within study comparisons and an especially strong quasi-experiment, each with an educational intervention that seeks to improve achievement. In each study, the pretest measures are consistently highly correlated with post-intervention measures of themselves, but the studies vary the correlation between the pretest and the process of selection into treatment. Across the three datasets with two outcomes each, there are three cases where this correlation is low and three where it is high. A single wave of pretest always reduces bias across the six instances examined, and it eliminates bias in three of them. Adding a second pretest wave eliminates bias in two more instances. However, the pattern of bias elimination does not follow the predicted pattern—that more bias reduction ensues as a function of how highly the pretest is correlated with selection. The findings show that bias is more complexly related to the pretest's correlation with selection than we hypothesized, and we seek to explain why.

Keywords

Within-study comparison; Propensity score matching; Randomized experiment; Causal inference

Kelly Hallberg hallberk@uchicago.edu.

Compliance with Ethical Standards

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required. This article does not contain any studies with animals performed by any of the authors.

Informed Consent This study only included de-identified, secondary data analysis. For this type of study, formal consent is not required.

Conflict of Interest The authors declare that they have no conflict of interest.

Introduction

Estimating causal effects is crucial in all the sciences seeking to learn how to prevent social problems. Sometimes, the causal research necessary to achieve this learning requires quasi-experiments rather than randomized control trials (RCTs). Analysts must then deal with the selection bias that arises because the quasi-experimental treatment and comparison groups come from different populations and so confound population differences with treatment effects—often called selection bias. In theory, the extent to which selection bias can be reduced depends on how well the data meet the strong ignorability assumption (Rosenbaum and Rubin 1983). This condition is met when the covariates used in the study impact analysis are perfectly correlated with all those parts of the selection process that are related to the study outcome. The practical problem, of course, is to know when these conditions are met with the available covariates. This paper explores how a pre-intervention measure of the study outcome or pretest influences bias.

The case for privileging the pretest in bias control is clear. First, it is usually a better predictor of the outcome than any other single type of pre-intervention measure. In education, for example, measures of academic performance tend to be highly correlated over time. Second, pretest performance levels are often (but not always) related to the reasons why school administrators assign individuals or schools to treatment, and also why individuals self-select into treatment. For instance, students are retained a grade because of poor earlier school performance, and students who want more instruction in a particular topic may do so because they already do well on that topic. It is no surprise, therefore, that pretests are highly recommended for the design of quasi-experimental studies to eliminate selection bias (e.g., Campbell and Stanley 1963).

Empirical studies also find that pretests can play a large role in accounting for selection and reducing bias (Glazerman, Levy, and Myers 2003; Bloom et al. 2005; Smith and Todd 2005), sometimes even reducing all of it (Bifulco 2012; St. Clair et al. 2014). However, it is also clear that they do not routinely eliminate all of the bias (Wong et al. 2016). The need, therefore, is to understand the conditions under which pretests reduce more or less of the bias. The studies we report here vary the strength of the relationship between the pretest and selection in order to test the hypothesis that the size of this relationship affects the amount of bias reduction achieved. We use a design replication study for this purpose, also known as a within study comparison (WSC) (LaLonde 1986; Cook et al. 2008). WSCs estimate the extent to which a specific quasi-experiment reproduces the causal estimates of an RCT when both designs have the same treatment group. The aim is to assess how similar are the posttest means of the non-equivalent comparison group and the experimental control group after steps have been taken in the quasi-experiment to reduce selection bias. When the adjusted posttest means are identical, the RCT and quasi-experimental results are interchangeable. However, when the final adjusted means differ, bias has not been eliminated, even though it might have been reduced from what it was prior to any adjustment efforts. This study varies the correlation between pretest and selection in order to examine how it affects final bias. We particularly want to determine whether the difference between RCT and quasi-experimental final estimates is so close to zero that it is plausible to assume minimal or zero

bias since sampling error in each design precludes obtaining estimates that are identical even when causal parameters are identical.

The paper proceeds as follows. We first lay out theoretical arguments for and against privileging pretests as a way to reduce selection bias in quasi-experiments. We then present our methodological approach, the hypotheses, the datasets we use, and the findings. We conclude by discussing the implications of our main finding that the pretest's effects on bias are not perfectly predicted by the pretest's correlation with selection.

Conceptual Framework

The Campbell tradition of causal inference has traditionally privileged observational study designs that include a pretest measure of the study outcome (Campbell 1957; Campbell and Stanley 1963). Shadish et al. (2002) even assert that “no single variable will usually do as well as the pretest for these [bias reduction] purposes” (p. 136). This assertion follows because no other single variable is likely to be as highly correlated with the outcome and also because selection processes are often the result of pretest differences or correlates thereof. Campbell was deeply skeptical about the validity of matching approaches without a pretest measure and also about those using an unreliably measured pretest that cannot function as intended (Campbell and Boruch 1975; Campbell and Erlebacher 1970). He also advocated pretest assessment at more than one pre-intervention time point, where possible in order to (1) compensate for some of the unreliability in a single pretest measure (Campbell and Boruch 1975; Steiner et al. 2011), (2) assess whether group performance is unusual immediately prior to the intervention and so increases the odds of regression to the mean (e.g., Ashenfelter 1978), and (3) to help identify and control for group differences in maturation or growth rates (Campbell and Boruch 1975).

Other scholars dispute a general bias-reduction role for pretests. Cronbach (1982, p. 188) argues that “a pretest twinned to the posttest does not necessarily give the best initial information.” Instead, he claims that pretreatment covariates should be chosen to measure both readiness for the treatment and the probability that an individual or institution will improve on the outcomes measured. He invokes a hypothetical study comparing two approaches to teaching physics to students with little prior exposure to the field, arguing that it makes less sense to administer a pretest measure of physics knowledge and more sense to measure student characteristics that are likely to predict success in learning physics. Rubin and his colleagues also do not grant special status to the pretest (e.g., Rubin and Thomas 1996). Instead, they emphasize selection bias control through a large array of covariates that are correlated with the selection process and measured prior to treatment. The pretest might be among them if it enters into the estimated propensity score, and it might be weighted more heavily therein than any other single variable. But it need not. The key is meeting the crucial strong ignorability assumption, and how it is met is secondary.

Proponents of “collider variables” take the critique of pretests a step further, implying that conditioning on them will sometimes increase bias rather than reduce it or leave it unaffected (Pearl 2009; Elwert and Winship 2014). Collider variables are not correlated with either the selection process or the outcome, but are correlated with other variables that are so

correlated. This sets up an endogenous selection process that opens up a “back door” and increases bias. Wooldridge (2009) showed that conditioning on an instrumental variable can also increase bias. So bias inflation is possible because of a pretest as well as the bias reduction on which we focus. Some light can be shed on bias increases. If the final quasi-experimental and RCT estimates hardly differ, this suggests that bias-inflation forces did not operate or did so very weakly. If only bias-inflating forces operated, then the final quasi-experimental estimate should be even more distant from the RCT effect than the raw posttest difference between the treatment and non-equivalent comparison groups—sometimes called the naïve bias estimate. Otherwise, it is very difficult to use data to differentiate the roles of bias reduction and increase. An incidental purpose of the present study is to examine whether use of a pretest leads to causal estimates that are either not different from the RCT value, suggesting that bias did not increase, or that are even more distant from it than the naïve estimate, suggesting that only bias-inflating forces operated.

The study’s main purpose, though, is to examine the causal contingency prediction that the pretest reduces bias by more when it is highly correlated with the selection process into treatment as opposed to when the correlation is smaller or zero. Statistical theory is clear that the elimination of bias depends on the extent to which a covariate captures all of the selection process that is correlated with the study outcome (Rosenbaum and Rubin 1983). A high correlation between the pretest and selection indicates that the selection process can be understood as a product of the pretest so that adjusting for group pretest differences will account for more of the initial bias. On the other hand, a low or null correlation between the pretest and selection process indicates that the pretest does not account for selection in the context studied, that other forces are responsible for whatever bias is initially observed. Adjusting for such a pretest should hardly affect bias reduction. We use three educational datasets to test this contingency hypothesis. One contrast between high and low correlations of the pretest with selection is achieved across studies, while the other is achieved within the same dataset and so holds constant irrelevant features that might differ between studies and affect the level of bias attained.

Methodological Approach and Data Sources

Each of the three datasets we present has a valid and credible estimate of the causal effect of the program or policy under study. Two come from an RCT, the usual and non-controversial benchmark in within study comparisons. The third comes from a particularly strong quasi-experiment, the validity of whose causal estimate rests on repeated observations of the selection process into treatment, reliable measurement of the elements from those observations, plus 140 other covariates designed to account for selection factors. This is obviously a weaker warrant than random assignment. Notwithstanding, each dataset has pretest and posttest achievement measures that are similarly and highly correlated. However, the pretest’s relationship with the selection process varies considerably across the datasets. In one application, the grade retention study, the pretest plays a substantial role in selection. In the second, the Indiana study, schools adopt a benchmark assessment system and do so without apparent regard for prior performance. The third study, the Memphis study, provides an even stronger test since two interventions are evaluated and the pretest plays a stronger role in selection with one of them than the other.

Dataset 1: Indiana Benchmark Assessment Study

The Study

The first dataset we examine is a cluster RCT (Konstantopoulos et al. 2013) that was designed to study how Indiana's benchmark assessment system affected student achievement in mathematics and English Language Arts (ELA). The treatment involved teachers receiving regular feedback about student performance that was disaggregated in a variety of ways in order to inform the teachers about the performance on specific learning tasks of the whole class and of individual students within it. The expectation was that teachers would use this feedback to improve their teaching. The study outcome was performance on the annual Indiana state test, for which data were available at both the school and student levels. In the 2009–2010 school year, 56 K-8 schools volunteered to implement the system. Of these, 34 were randomly assigned to the state's benchmark assessment system while 22 served as controls. Here, we analyze only 5th grade data because no state data exist for grades K through 2, most Indiana primary schools have a K through 5 grade structure and so 6th through 8th grade data are sparse, and grades 3 and 4 could not provide the pretest assessments required for this study's purpose. We use both student and school level pre-intervention data, the latter from the past 5th grade cohorts.

The non-experimental comparison group was constructed from all 1007 schools in the state that served 5th graders. Of these, 326 were excluded since they were already implementing something close to the state's benchmark assessment system. As a result, the pool for selecting no-treatment comparison schools consisted of the 681 schools serving 5th grade students that did not implement a benchmark assessment system during the 2009–2010 school year.

We matched treatment and comparison schools on the basis of observable pre-treatment school characteristics. These were annual school math and reading scores over five pre-intervention years and school-level demographic information (including the portion of student by race/ethnicity, free and reduced price lunch status, special education status, and English Language Learner status). In addition, the Indiana Department of Education (DOE) and the Common Core of Data (CCD) also provided multiple years of prior data on school size, school structure, attendance rates, levels and growth rates for achievement on state tests, average teacher and administrator salaries, and whether the school is a Title I, charter, or magnet school.

Pretest Correlations with Selection and Outcome

Schools had to apply to the DOE if they wanted to implement the new program in the 2009–2010 school year. Of those applying, most were invited to participate in the RCT and schools that finished up in the control group were promised access to the program next year. The process of selection into the study reflects primarily the principal's interest in implementing the new benchmark assessment system. A sample of study school principals was asked why they had decided to apply, and their responses suggest a wide variety of reasons. Some saw the intervention as an opportunity to take advantage of free resources; others cited a pre-existing interest in data-driven decision-making; and yet others mentioned

knowing other schools that had implemented the program in an earlier pilot test. Past school performance was never explicitly mentioned as a reason for volunteering to implement the new assessment system.

Correlations between pretest measures and selection were trivial. The correlation between selecting into the study and school-level average ELA performance in the spring before the study was -0.041 , and the analogous correlation for mathematics was -0.012 . Neither was reliably different from zero. So just before the intervention, schools that volunteered to implement the program were not reliably different in achievement from the rest of the state. Selection bias is still theoretically possible from other sources that are not correlated with pretest achievement but that are correlated with the true but unknown selection process. However, it is not easy to imagine forces that are related to posttest but not to pretest achievement, given the usually high correlation between them—the spring to spring correlations in math and ELA achievement were consistently close to 0.80 at the school level.

The Causal Benchmark

The causal benchmark comes from the 5th grade students in the RCT. We examined balance on all pretreatment covariates and discovered that treatment and control groups were not significantly different from one another at the 0.05 level on any of the 27 school-level pretreatment covariates measured at the spring pretest time point closest to the intervention. However, these analyses had modest power. The pretest math difference approached statistical significance ($p = 0.10$) and four of the 27 balance tests showed non-reliable differences larger than .25 standard deviation units. So we examined the pretest differences over five pre-intervention years, and these showed consistent differences favoring the treatment schools over the control ones. Such consistency in the direction and temporal pattern of pre-intervention treatment/control differences points to the distinct possibility of an imbalanced RCT. To address this imbalance, our RCT outcome model included the four school-level covariates with a standardized mean difference greater than 0.25 standard deviations at the immediate pretest time—current best practice in RCT analysis according to the What Works Clearinghouse (2011).¹

Dataset 2: Kindergarten Retention

The Study

The second dataset involves the study of the effects of kindergarten retention (Hong and Raudenbush 2005, 2006). Data were drawn from the Early Childhood Longitudinal Study—Kindergarten Class of 1998–1999 (ECLS-K). The study follows a nationally representative sample of kindergarten students through eighth grade. From fall and spring of the kindergarten year prior to grade retention, we used covariate measures of children's cognitive, social, emotional and physical development, as well as measures of their home environment, home educational activities, school structures and supports, classroom learning

¹To ensure the congruence of effect estimates across models, only school-level covariates were included in the models presented here. However, inclusions of pretreatment student level covariates in the outcome models did not substantively change the analysis results.

environment, and teacher qualifications. The outcomes were reading and math assessed in the spring of the year after kindergarten but also available as pretests from the Spring and Fall of the pre-retention year. We limited the analysis to schools where more than one student was retained, thus 1080 schools with 10,726 students.

Correlations with Selection and Outcome

Pretest achievement and selection into grade retention are likely to be highly related since it is common practice to retain children in kindergarten “to remedy inadequate academic progress” (Jackson 1975, p. 614). Alexander et al. (2003) further show that “the risk exposure of such children is not to grade retention only. It is a ‘high risk’ profile generally – for academic setbacks in the near-term, for a lifetime of struggle over the longer term” (p. 68). Such comments suggest the high likelihood of negative bias due to retained students performing less well academically than their non-retained counterparts.

The reading pretest measured most proximally to the introduction of treatment is correlated -0.185 with selection into being retained; the corresponding correlation of the math pretest is -0.179 . Both are statistically significant at the 0.05 level. But as point-biserial correlations, their upper bound depends on the extremity of the split on the dichotomous variable—in this case, 447 retained students versus 9995 non-retained ones. Correcting for this extreme split, the upper bound of the correlation between selection and the pretest is -0.381 for math and -0.366 for reading (Demirtas and Hedeker 2011). The test-retest correlations between pretest and outcome were 0.744 and 0.756 in reading and mathematics, respectively. Given the pretest measures are highly related to both retention in kindergarten (selection) and the study outcome, we would expect the pretest to remove much of the initial bias, though how much is not clear.

The Causal Benchmark

The Hong and Raudenbush study (2005; 2006) did not have an RCT benchmark. Instead, we use the causal estimate achieved when all 144 of the available covariates were used to construct propensity scores. Among these are the two achievement pretests for math and reading in fall and spring, as well as proxy pretest assessments at the same time from teachers who judged the kindergarten reading and math performance of each child. The crucial assumption is that using all 144 observed covariates results in less selection bias than when any subset of them is used alone. Since some crucial covariates might nonetheless still be missing, and since some of those observed might inadvertently function as “colliders” or falsely controlled instrumental variables (Pearl 2009; Wooldridge 2009), these 144 covariates do not meet the strong ignorability assumption in the same way random assignment does. Instead, one has to make the case that the causes of kindergarten retention are well understood, that prior performance and teacher judgments of this performance are major causes of retention, that repeated reliable measures of this past performance and of teacher judgments of it are among the 144 covariates, and that the 140 other measured covariates cover many heterogeneous domains that might account for being retained for reasons other than past performance and teacher judgments of this performance. It is on the basis of these assumptions that we use as benchmark estimates of cause the results of a

propensity score analysis with 144 covariates, contrasting this with the results when pretest measures alone are used.

To estimate the propensity to be retained, all 144 pre-retention covariates were used in a logistic regression. Because characteristics of both students and their schools might influence the retention decision, measures at both levels were included in the model for each student. The final estimated propensity score only included those covariates that were correlated with selection and maximized balance. To select the final propensity score, we began by including only those covariates correlated with retention at a p value less than 0.2 in a forward stepwise regression function. For this first estimate of the propensity score, we then checked balance on all 144 covariates using standardized mean differences and variance ratios to test whether pretreatment group differences remained on the observed covariates (Rubin 2001; Shadish et al. 2008). If balance was not then achieved on all 144 covariates, we improved the model by including additional covariates and higher order terms (interaction and quadratic effects) until satisfactory balance was obtained. In the final model, which included 100 covariates, the standardized mean difference between the retained and non-retained groups was less than 0.1 across all 144 covariates.

An examination of the overlap between the treatment and comparison groups on the logit of the propensity score revealed some non-overlap, mostly due to some students scoring so high they never had any realistic chance of being retained and who were omitted—284 cases out of 9995. Of the 471 retained students, 24 scored so low that they too fell outside the area of common support and were excluded. The resulting analytic dataset includes 10,442 students, of whom 447 were retained. This sample was used in all the analyses described below.

We estimated our benchmark treatment effects for the math and vocabulary outcomes with a weighted two-level hierarchical linear model. The weights used to estimate the hierarchical linear model controlled for the differential distribution of treatment and control cases across propensity score strata (Hong and Raudenbush 2005). This weighting scheme reflects the stratification approach for estimating the average treatment effect for retained students. Promoted students who were overrepresented in a particular stratum relative to the retained students were down-weighted, while promoted students who were underrepresented were weighted upwards.

Dataset 3: Vocabulary and Mathematics Training

The Study

The ECLS-K and Indiana data sets provide a between-study contrast of pretest measures that vary in their correlation with selection. But they also differ in other factors that might correlate with the capacity of pretests to reduce bias—e.g., the amount of bias originally obtained or, to a lesser degree, in the pretest correlation with outcome. So the third data set was chosen to provide a WSC of the effects of differences in the pretest correlation with selection. As such, it provides the conceptually strongest test of the role of the pretest.

The data come from Shadish, Clark, and Steiner (2008). In this study, conducted in Memphis, individual students were first randomly assigned to serve in an RCT or a quasi-experiment and then they were assigned to treatment, either randomly in the RCT or by self-selection in the quasi-experiment. Treatment was exposure to a mathematics or vocabulary training session. A cross-over design was used such that the vocabulary scores of those exposed to math serve as the control observations for the vocabulary intervention, while the math scores of those exposed to vocabulary serve as controls for the math intervention.

The available covariates for adjustment were 156, representing 23 constructs in five domains entitled demographics, proxy pretests, prior academic achievement, topic preference/motivation, and psychological predisposition. The study lasted 50 min in total. Because of the short pretest-posttest period and the highly specific curricula, no true pretest assessments could be collected for fear of posttest responses being affected by memory of pretest responses. The proxy pretests assessed were general measures of math or vocabulary whereas the posttest outcomes dealt with knowledge of logarithmic functions and strategies for learning vocabulary, the former relevant to the math intervention and the latter to the vocabulary one.

Correlations of Pretests with Selection and Outcome

Participants in the quasi-experiment were asked why they selected either the mathematics or vocabulary training. Forty-two percent of vocabulary and 47% of math participants reported selecting their treatment for self-improvement; 17% of vocabulary and 11% of mathematics participants reported selecting their training because they were good at the subject; 18% of vocabulary and 30% of mathematics participants said they selected their training because they liked the subject; and 21% of vocabulary and 8% of mathematics participants selected their training in order to avoid the other subject.

The key issue is how such perceptions of the selection process are correlated with performance on the proxy pretest measures. The vocabulary pretest measure was positively and significantly correlated with selection into vocabulary training (bi-serial correlation = 0.169), while the pretest measure of mathematics was not reliably correlated with selection into math training (-0.09). The correlations with outcome were 0.446 for the math measures and 0.468 for the vocabulary measures, thus not as high as for the true achievement pretests in the other WSC studies but still quite high and essentially equal across the two outcomes. This differential pattern of within-study correlation between pretest and selection complements the between-study contrast that the kindergarten retention and Indiana datasets together provide.

The Causal Benchmark

The effect of the intervention was estimated at the student level, and 445 students were assigned to treatment. Despite the RCT being subject to some sampling error, balance in the RCT was good. So the RCT outcome analysis included all of the available pre-treatment covariates in a backward stepwise regression.

Analytic Approach

For each dataset, a propensity score approach was implemented to identify a matched comparison group. In the Indiana dataset, each treatment school was matched with replacement to the comparison school with the closest propensity score.² In the remaining two datasets, a propensity score stratification approach was employed to maximize balance across all available pre-treatment covariates.

Findings

Table 1 presents the results for all three datasets. Initial bias reflects the observed and unobserved pre-intervention differences between treatment and comparison schools. Final bias—the purpose of this paper—is calculated as the standardized difference at posttest between the RCT benchmark whichever pretest-adjusted quasi-experimental estimate is compared to it.

The first row gives the benchmark results by study and outcome. For the two studies with an RCT benchmark, all four treatment estimates differ from zero, indicating that each treatment had a replicated effect and that the task of the quasi-experiment is to see how well it replicates the same positive effect size. For the non-RCT benchmark with a well-known and well-measured selection process plus 140 other covariates, the benchmark effect is also statistically significant and indicates that retaining students a year lowered subsequent performance.

The next row indicates the naïve bias, the posttest difference between the treatment means and the non-equivalent comparison means absent any controls. It also reports how far this difference is from the benchmark treatment effect, the next column labelled as “bias”. All datasets show some naïve estimate bias, though its magnitude differs by study. It is largest (almost .80 standard deviation units) for the grade retention study. This is because the raw achievement difference between retained and promoted students is obviously replete with selection bias due to promoted students scoring systematically higher than retained ones. The naïve bias is smallest for the Indiana study where principals volunteered to be in the study. The estimate shows that the treatment and comparison schools had similar posttest means. Since the RCT showed reliably higher treatment than control means, this indicates that the naïve comparison would have failed to detect a true treatment effect while the RCT would have detected it. The final bias is of about .011 standard deviations for ELA and of .013 for math.

The next row of Table 1 indicates the posttest mean difference in the quasi-experiment after controlling for the pretest, and also the final bias remained after the quasi-experimental and RCT estimates were differenced. In the Indiana study, all the bias is removed, with the difference between the RCT and the pretest-adjusted quasi-experiment being under .10 SDs for each outcome. The same is true for the vocabulary intervention in the Memphis study. But it is not true for either outcome in the retention study or for the math outcome in the

²The estimates presented here rely on 1:1 nearest neighbor matching, but are robust to alternative specifications.

Memphis study. In all six cases, a single pretest reduces some bias but does not consistently remove all of it. Some treatment effect coefficients differ by more than .20 SDs between the RCT and the pretest-adjusted quasi-experiment.

However, the retention study includes two pre-intervention time waves for each outcome. The next row shows the results of controlling for each pretest time point separately. The bias then shrinks below .10 for each outcome, indicating total bias removal. But that bias removal is a function of two waves of pretest data rather than the one.

The Memphis study only one wave of pre-intervention data, and the (proxy) pretest failed to reduce all the math bias. Something other than math achievement presumably accounted for the bias initially obtained with the self-selection into math exposure. However, a total of 156 covariates are available for use in the outcome analyses, and the next row shows that using all of them in a propensity score analysis eliminates all the math bias in the quasi-experiment.

Discussion

Across six tests based on three independent datasets and two dependent achievement outcomes, bias was completely removed in all instances. That is good news for those wanting to know if all the bias in quasi-experiments can be reduced, at least in the academic achievement context. However, this conclusion depends on two major assumptions. The first concerns the criterion that the RCT and final quasi-experimental estimates should differ by less than .10 SDs to be considered functionally equivalent, given sampling error in each design. In fact, the final bias results were consistently smaller than this and a lower criterion could have been invoked. In the Indiana study with one pretest wave, the final difference in estimates were .03 for ELA and .02 for math; in the retention study with two pretest waves the final bias estimates were .06 and .05 respectively; and in the Memphis study with all 156 covariates the differences between the RCT and quasi-experimental estimates were .00 and .03. Readers can judge whether these bias effects are small enough to treat the RCT and quasi-experimental estimates as functionally similar; we believe they meet that challenge. The second assumption is that the benchmarks were adequate. In the Memphis study there is no question of bias in the RCT. In the small sample Indiana study, there is evidence of a small selection bias over time for which we attempted to control the final RCT results, but to which we return later. The biggest problem is with the retention study, for there is no benchmark RCT and reliance is placed instead on the argument that (1) the selection into grade retention is independently known to be largely determined by pretest achievement and teacher ratings, each of which was measured twice in the dataset we analyzed; and (2) the other 140 heterogeneous covariates significantly increase the odds of accounting for all the otherwise hidden biases that the achievement tests and teaching ratings do not capture. Fortunately, reliance on this argument is not absolute, for the retention study is only one instance of a high correlation between the pretest and selection process. The vocabulary manipulation in the Memphis study provides a replication of this correlation and an adequate RCT benchmark.

The main purpose of this paper is not to identify the different covariate sets that eliminate bias from one within study comparison to another. It is about the role the pretest plays in bias reduction. We discovered that, when the pretest was used alone, it reduced some of the original bias in all six tests. However, its relationships to bias *elimination* were more complex. A single pretest removed all the bias in three cases; a second wave did so in two other cases; and in the final case, much bias still remained even after pretest adjustment. Viewed unconditionally, a single pretest consistently reduced bias and mostly but not always eliminated it, though two waves of pretest data were sometimes necessary for the elimination.

However, this study postulates a conditional relationship between bias removal and how highly the pretest is correlated with selection into treatment. The higher this correlation is, the more the bias removal should be. Unfortunately, the pattern of results does not fit this prediction. In the three planned cases where the correlation was high, only once did a single pretest totally remove all the bias—the vocabulary manipulation in Memphis—while in the grade retention example two pretest waves were required for both reading and math. The puzzle here is not the failure to reduce all the bias; it is that two waves of pretest data were sometimes needed instead of one. Where the planned correlation of the pretest with selection was low or essentially zero, a single pretest should not remove much bias since it is not responsible for the observed selection process. This is exactly what happened in Memphis with the math intervention. However, if bias removal is observed with no correlation between the pretest and selection, then factors other than the pretest must be responsible for it. The pretest and selection were not related in Indiana, yet the pretest still removed all of the bias in each outcome. The puzzle is why the pretest controlled for a selection process of which it is not a part. Even if the pretest were correlated with some hidden variable responsible for selection, should this shared source not lead to a correlation between pretest and selection?

So why did it take two pretest measurement waves for all the bias to be eliminated in the retention study? Noteworthy in that study was the large size of the initial bias. Students who are to be retained a year are quite different in achievement levels to students who are not retained, even after our modest trimming of under 5% of all promoted students who scored too high to be matched and the same proportion of retained students who scored too low to be matched. Indeed, the initial population difference was between .70 and .80 SDs across the two study outcomes. Our speculation is that bias elimination is more difficult, the larger is the selection bias to be adjusted. One pretest wave brought considerable bias reduction, from the .80/.70 range to the .30/.20 range, but it took two waves to bring it down to the .05/.06 range. This might be due to the size of the initial selection bias or to the higher reliability that two measures offer over one. More likely is a third possibility. Inspection of the two pre-intervention time trends shows that the promoted students outperform their retained counterparts, not just in mean performance but also in temporal slope. This creates a pattern of time-varying inequality growth—selection-maturation—that is common with academic achievement scores examined over a broad population of students and is sometimes characterized as a “fan spread pattern” (Campbell and Stanley 1963). That is, the means and variances increase over time within groups and also between them. Controlling for two

separate pretest waves in the analysis increases the chances of picking up the selection-maturation part of the total selection process.

The bigger puzzle is why the single wave of pretest reading and math scores in Indiana was not correlated with selection but nonetheless eliminated all the selection bias. The results for the math intervention in Memphis that also had a low correlation between the pretest and selection are as expected—little bias reduction from using the math pretest. An earlier study (Steiner et al. 2010) has shown that students self-selected into math instruction for reasons having less to do with cognitive factors like those the pretest picks up and more to do with motivation because they liked math, had less fear of it, and preferred it over language arts. Any analysis using just the pretest would not capture these motivational reasons for treatment exposure since they are not highly correlated with the cognitive factors the pretest assesses. But these motivational factors were included among the 156 covariates in our study, and all the bias was eliminated when they were used in the resulting propensity score that deliberately capitalizes on the factors most correlated with selection. In Steiner et al. (2010), fewer than 156 covariates were needed since the motivational factors alone sufficed to eliminate all bias. It was they that met the demands of the strong ignorability assumption, which could be inferred from comparing the RCT and quasi-experimental results. Alas, this cannot be one in the usual stand-alone quasi-experiment where it is opaque how well the strong ignorability assumption has been met.

The surprise is that the Indiana results do not follow the Memphis math pattern. There was selection bias in the study, but only a modest amount of it—between .011 and .013 SDs by outcome. Since the pretest's relationship to selection was essentially nil for each outcome, bias removal should not have been observed. Yet the final bias shrank from .011 to .03 for ELA and from .013 to .02 for math. Was this unexpected bias removal due to some unobserved, hidden source that was correlated with the pretest but that “somehow” did not increase the correlation between the pretest and selection? Was there a pattern of pretest correlation with selection that would have been of one sign without the hidden variable but that became zero because the hidden variable had an equal correlation with the pretest, but of the opposite sign? A convoluted explanation like this is possible; but its plausibility remains hard to fathom. One version of it stems from the earlier description of pre-intervention time trends in the Indiana RCT. It revealed a more stable pattern of possibly imbalanced time trends favoring the treatment group than one pretest wave could capture in the modestly powered RCT with only 56 schools. The treatment group in that RCT also served in the quasi-experiment, and the possibility exists that the true score pretest (assessed better over five pre-intervention time points than one) was more involved in selection than the obtained pretest observed at only one time immediately prior to treatment. Were the pretest and comparison schools closer to each other at the last pretest time point than at earlier points, raising the specter of regression bias in the treatment group that the single pretest measure failed to account for? Did time-varying uncertainty in the treatment group inflate the estimate of treatment group performance in the RCT? Would a different random assignment procedure or are larger sample of schools have resulted in more of a pretest correlation with selection than we observed with the single pretest observation? We cannot be certain, but it is clear that the Indiana results vitiate our initial hypothesis.

This paper demonstrates that the pretest may be special for the frequency with which it reduces selection bias. But it is not special because it routinely or simply eliminates such bias. In that sense, the Campbell tradition's advocacy of pretest measures of the study outcome is too simple. Traditions emphasizing the use of multivariate covariate data are more relevant, as Cronbach (1982) pointed out they would be and as Rubin's work on covariate choice reinforces. The best way to select covariates is to use field observation, key informants, literature reviews and shared common sense to develop multiple theories of the selection process and then to collect reliable measures of all the concepts in all the theories. This is far from current practice in quasi-experimental research, and we acknowledge the practical limitations to making it happen on a regular basis. But still, it is the best theory-linked practice for covariate choice. The next best practice is to collect as conceptually heterogeneous a set of covariates as possible, and the pretest should probably always be among them because it is often part of the selection process and is especially highly correlated with the study outcome. But as a stand-alone covariate choice, it is not dependable. As we have seen here, pretests assessed at more than one time are also important. This increases reliability and, as we saw, may also account for some time-varying selection processes. Moving beyond two or three prior pretest waves to a comparative interrupted time-series design is an even better strategy of covariate choice, for then pre-intervention time trends can be estimated and compared across the treatment and comparison groups. To date, within study comparisons of RCT versus comparative interrupted time series results have consistently shown strong correspondence in the educational context (St. Clair et al. 2014).

It is theoretically possible for pretests (and other covariates) to increase bias rather than reduce it. It is not easy to test when forces inducing bias-inflation occur, since the two forces might countervail with the relative power of each being unclear. However, if bias-inducing forces were strong, we should expect to find few cases where the RCT and quasi-experimental estimates coincide since bias inflation would prevent bias reduction that finished up close to zero. But in all the analyses presented, all the estimates of final bias varied between .00 and .06 SDs. These low values suggest, but do not definitively prove, that any bias-inflating forces operating in these three datasets were quite weak relative to the bias-reduction the covariates achieved. The theoretical possibility of bias increases is real, but we do not yet know the frequency of such bias, the strength of it, or the conditions under which it is most likely to occur in the social sciences as opposed to in the engineering contexts in which collider variables originated.

The lessons of the analyses presented above are that (1) a pretest can help in bias reduction; (2) two waves of a pretest are better than one, especially where selection-maturation or modest reliability are expected; (3) a large and heterogeneous set of covariates is likely to reduce more bias more consistently than a single pretest wave; (4) within such a set, the pretest will often play a major role and so its omission requires justification; (5) proxy pretests can substitute for "true" pretests; and (6) while adding pretests to the analysis may sometimes increase bias, but this did not occur in any systematic way in the datasets examined where the final bias estimates were systematically close to zero. These conclusions apply only to educational interventions designed to raise academic achievement, and it is unclear how broadly they apply to other fields. The bottom line is that while pretests

are very useful in bias reduction, they cannot be depended on to eliminate selection bias in observational studies. If bias elimination is to come under better researcher control, then pretests need to be complemented by other covariates measured prior to the intervention that are explicitly derived from thoughtful observation and analysis of what the true selection bias might be, from otherwise unusually rich sets of covariates, and preferably from combining these two strategies of covariate choice.

Acknowledgments

Funding This work was supported by the National Science Foundation Grant DRL-1228866.

References

- Alexander KL, Entwisle RD, & Dauber SL (2003). *On the success of failure: A reassessment of the effects of retention in the primary school grades* Cambridge: New York.
- Ashenfelter O (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 67, 47–57.
- Bifulco R (2012). Can nonrandomized estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, 31, 729–751.
- Bloom H, Michalopoulos C, & Hill C (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In Bloom H (Ed.), *Learning more from social experiments* New York: Russell Sage.
- Campbell DT (1957). Factors relevant to the validity of experiments in social setting. *Psychological Bulletin*, 54, 297–312. [PubMed: 13465924]
- Campbell DT, & Boruch RF (1975). Making the case for randomized assignment to treatments by considering the alternatives. In Bennett CA & Lumsdaine AA (Eds.), *Evaluation and experiments: Some critical issues in assessing social programs* New York: Academic.
- Campbell DT, & Erlebacher AE (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In Hellmuth J (Ed.), *The disadvantaged child: Vol. 3, Compensatory education: A national debate* New York: Brunner/Mazel.
- Campbell DT, & Stanley J (1963). *Experimental and quasi-experimental designs for research* Boston, MA: Houghton Mifflin Company.
- Cook TD, Shadish WJ, & Wong VC (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management*, 27, 724–750.
- Cronbach L (1982). *Designing evaluations of educational and social programs* San Francisco, CA: Jossey-Bass Publishers.
- Demirtas H, & Hedeker D (2011). A practical way for computing approximate upper and lower correlation bounds. *The American Statistician*, 65, 2.
- Elwert F & Winship C (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *The Annual Review of Sociology*
- Glazerman S, Levy D, & Myers D (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy*, 589, 63–91.
- Hong G, & Raudenbush SW (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Education Evaluation and Policy Analysis*, 27, 205–224.
- Hong G, & Raudenbush SW (2006). Evaluation kindergarten retention: A case study for causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101, 901–910.
- Jackson GB (1975). The research evidence on the effects of grade retention. *Review of Educational Research*, 45, 613–635.
- Konstantopoulos S, Miller S, & van der Ploeg A (2013). The impact of Indiana's interim assessments on mathematics and reading. *Educational Evaluation and Policy Analysis*, 35, 481–499.

- LaLonde R (1986). Evaluating the econometric evaluations of training programs with experimental data. *Annual Economic Review*, 76, 604–20.
- Pearl J (2009). The structural theory of causation. In McKay Illari P, Russo F, & Williamson J (Eds.), *Causality in the sciences* (pp. 1–30). Oxford: Clarendon.
- Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rubin DB (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2, 169–188.
- Rubin DB, & Thomas N (1996). Characterizing the effect of using linear propensity score methods with normal distributions. *Biometrika*, 79, 797–809.
- Shadish WR, Clark MH, & Steiner PM (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 WR, Cook T, & Campbell D (2002). *Experimental and quasi-experimental designs for generalized causal inference* Boston, MA: Houghton Mifflin Company.
- Smith J, & Todd P (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 305–353.
- Clair T St., Cook TD, & Hallberg K (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*
- Steiner PM, Cook TD, Shadish WR, & Clark MH (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250–67. [PubMed: 20822251]
- Steiner PM, Cook TD, & Shadish WR (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36, 213.
- Wong V, Valentine JC & Miller-Bains K (2016). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness*
- Wooldridge JM (2009). Should instrumental variables be used as matching variables? Working paper

Table 1

Experimental and quasi-experimental effect estimates by study and outcome

Indiana				Kindergarten Retention						Shadish et al.					
ELA				Reading			Math			Vocabulary			Math		
TE	Bias	TE	Bias	TE	Bias	TE	TE	Bias	TE	TE	Bias	TE	TE	Bias	Bias
Causal benchmark	6.04(2.97)	-	12.03(5.43)	-	-9.20(1.09)	-	-5.29(0.85)	-	8.18(0.39)	-	4.06(0.36)	-	4.06(0.36)	-	-
Naïve effect	0.14(3.38)	-0.11	3.13(4.97)	-0.13	-19.80(0.38)	-0.79	-11.86(0.17)	-0.74	9.00(0.51)	0.24	5.01(0.55)	0.30	5.01(0.55)	0.30	0.30
One pretest wave	7.80(3.52)	0.03	13.59(5.23)	0.02	-12.70(1.10)	-0.26	-7.21(0.77)	-0.22	8.44(0.42)	0.08	4.73(0.48)	0.21	4.73(0.48)	0.21	0.21
Two pretest waves	-	-	-	-	-10.07	-0.06	-5.76	-0.05	-	-	-	-	-	-	-
All covariates	-	-	-	-	(1.12)-	-	(0.74)-	-	8.19(0.49)	0.00	3.98(0.39)	-0.03	3.98(0.39)	-0.03	-0.03

Treatment effects (TE) for each analytic approach are presented in units of the outcome used in each study (ISTEP+ for Indiana, ECLS-K math and reading scores for kindergarten retention, and study-created measures for the Shadish et al. study). Bias is the difference between each of the observational study estimates and causal benchmark. Bias is presented in standard deviation units to allow for comparison across the data sets