EdPolicyWorks

Working Paper:

Empirical Performance of Covariates in Education Observational Studies

Vivian C. Wong[1], Jeffrey Valentine[2], & Kate Miller-Bains[1]

This paper summarizes results from 12 empirical evaluations of observational methods in education contexts. We look at the performance of three common covariate-types in education observational studies where the outcome is a standardized reading or math test. The covariate-types are: pretest measures on the outcome, local geographic matching, and rich covariate sets with a strong theory of treatment selection. Overall, the review demonstrates that although the pretest often reduces bias in observational studies, it does not always eliminate it. Its performance depends on the pretest's correlation with treatment selection and the outcome, and whether pre-intervention trends are present. We also find that although local comparisons are prioritized for matching, its performance depends on whether comparable no-treatment cases are available. Otherwise, local comparisons may produce badly biased results. In cases where researchers have a strong theory of selection and rich covariate sets, observational methods perform well, but additional replication studies are needed. Finally, observational methods that rely on demographic covariates without a theory of selection rarely produce unbiased treatment effects. The paper concludes by offering education researchers empirically-based guidance on covariate selection in observational studies.

[1]University of Virginia
[2]University of Louisville

Updated April 2016

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

Acknowledgements: The work reported herein was supported in part by the U.S. Department of Education[1]s Institute of Education Sciences (contract EDIES12C0084). However, the views expressed are the authors and do not necessarily represent the positions or policies of the Institute of Education Sciences or the U.S. Department of Education.

EMPIRICAL PERFORMANCE OF COVARIATES IN EDUCATION OBSERVATIONAL STUDIES

Vivian C. Wong, Jeffrey Valentine, & Kate Miller-Bains

Introduction

Despite recent emphasis on the use of randomized control trials (RCTs) for evaluating education interventions, in most areas of education research, observational methods remain the dominant approach for assessing program effects. Among studies reviewed by the What Works Clearinghouse, only 30% of these studies were RCTs. The most common method for assessing program impacts was observational studies, followed by regression-discontinuity and single case-study designs (Institute of Education Sciences, 2015).[1] And, in subject areas that the Obama Administration has prioritized for improvement in teacher and student proficiency, there is even greater reliance on non-experimental methods for identifying what works. In a review of all federal STEM initiatives that included summative evaluations, the National Science and Technology Council observed that only 8% used random assignment for assessing program impacts, while 24% used a non-equivalent comparison group design with matched comparison groups, 15% used a non-equivalent comparison group design without matching, and 53% used single group pre-post designs (National Science and Technology Council, 2011). Given the widespread use of observational methods for assessing program impacts, it is critical that education researchers understand contexts and conditions under which these approaches produce trustworthy results, and as importantly, when they fail.

Much of the observational literature has been devoted to the development and improvement of statistical theory for estimating unbiased treatment effects (Rubin, 1974; Rosenbaum & Rubin, 1983, 1985). In the study of methods, statistical theory is critical for identifying conditions (i.e. assumptions) that are needed for an approach to estimate causal effects, but these conditions often refer to relationships and quantities that are unobserved by the researcher. Simulations provide insights about the performance of a method when specific conditions (e.g. the strong ignorability assumption) are violated, but they rarely capture all of the complexities that are likely to arise in real world evaluation settings. Moreover, simulations often fail to address the many practical questions that researchers face in field settings. How should comparison pools be selected for drawing propensity score matches? What covariates are needed for covariate adjustment or propensity score

[1] What Works Clearinghouse does not review single group pre-post design studies.

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

matching? Does geographical proximity matter in forming better comparisons in education settings? These are questions about the empirical performance of non-experimental methods in field settings.

Over the last three decades, the within-study comparison (WSC) design has emerged as a method for empirically evaluating the performance of non-experimental methods. In a traditional WSC design, treatment effects from an RCT are compared to those produced by a non-experimental approach that shares the same target population and intervention. The non-experiment may be a quasi-experimental approach, such as a regression-discontinuity (RD) design study; it may be an observational study, where comparison units are matched to treatment cases using different statistical procedures or covariates. The goals of the WSC are to determine whether the non-experiment replicates results from a high-quality RCT in field settings.

Recent WSCs have examined contexts and conditions under which observational methods produce valid results in field settings. In particular, these studies have looked at the performance of covariates for addressing selection bias in education observational studies. They have looked at how well: 1) *pretest measures* perform when the outcome is students' reading or math achievement scores (Tables 2 and 3); 2) matching units within the *same geographic area* reduces bias (Table 4); and 3) observational methods perform *when there is a rich covariate set* and a theory of selection (Table 5). Relatedly, WSC studies have looked at when the covariate set is poor and there is no theory of selection, such as when only generic demographic variables are available (Table 6). Combined, these WSC studies outline three common strategies that education researchers employ for choosing covariates in observational studies.

As standalone enterprises, results from WSC studies may have little to say about the general performance of covariates in observational studies. But results from *multiple* empirical evaluation studies provide insights as to how well these methods perform more generally for outcomes and settings of particular interest. There have been four reviews of WSC results. Three of these reviews have focused on the performance of methods in the context of the job training interventions (Bloom, Michalopoulos, & Hill, 2005; Glazerman, Levy, & Myers, 2005; Smith & Todd, 2005). They asked: "Do non-experimental methods succeed in replicating experimental benchmark results?" Glazerman et al. (2005) conducted a quantitative meta-analysis of 12 WSCs in job training; Bloom et al. (2005) provided a qualitative review of results; and Smith and Todd's (2005) qualitative review highlighted methodological weaknesses in the early WSC evaluations of methods. Overall, the reviews concluded that although non-experimental approaches sometimes replicated experimental

benchmark results, they often produced effects that were "dramatically different from the experimental benchmark" (Glazerman et al., p. 86).

Cook, Shadish, and Wong (2008) provided an updated qualitative synthesis of WSC results not included in the Glazerman et al. (2005) meta-analysis, but spanned the fields of international development, education, job training, and health. Their goal was to scan the WSC literature to develop hypotheses about conditions under which non-experimental approaches appear to replicate benchmark results. Their review generated three hypotheses about when non-experimental methods appeared to perform well in field settings: when individuals were assigned into treatment using an assignment variable and cutoff, as in the case of an RD design; when intact groups were matched on covariates that are highly correlated with the outcome and are geographically local; and when the researchers were able to model units' treatment selection.

Since 2008, researchers have designed WSCs to test hypotheses proposed by Cook et al. (2008). More empirical validation studies of the RD design emerged (Berk, Barnes, Ahlman, & Kurtz, 2010; Shadish, Galindo, Steiner, Wong, & Cook, 2011; Wing & Cook, 2013), as well as studies that examined intact group matching (Diaz & Handa, 2006; Michalopoulos, Bloom, & Hill, 2004), and matching or covariate adjustment based on the pretest (Cook & Steiner, 2010).

This paper extends the prior WSC review literature in three important ways. First, for substantive researchers in education, this is the first paper to provide empirically-based guidance on the performance of three common covariate sets (pretests and proxy pretests, local geography, and rich covariates plausibly related to treatment selection) used in observational studies with reading and math achievement outcomes. As more WSCs become available, quantitative syntheses are also needed, but this review is qualitative to provide readers with more nuanced understandings of contexts and conditions under which these covariates appear to "work" in field settings. Second, unlike prior qualitative reviews of the WSC literature, the paper summarizes results as standardized bias estimates. This allows readers to compare and interpret results across multiple studies using the same measure of correspondence, as well as to ascertain the magnitude of non-experimental bias that remains, even after matching on or adjusting for covariate types. Moreover, these bias estimates may provide helpful benchmarks for researchers assessing the robustness of their own observational treatment effects, especially in cases where an RCT is not available. Third, as new statistical methods emerge, WSCs provide a valuable tool for assessing whether stringent assumptions required for these methods are met in field settings. This paper demonstrates how ongoing reviews of WSC

results are central for examining hypotheses about non-experimental performance in field settings, as well as highlights areas where further empirical evidence on observational methods is needed.

The remainder of the paper is organized as follows. We begin by discussing common statistical methods for addressing bias in education observational studies, and the role that covariates play in implementing these approaches. We then consider results from empirical evaluations of the pretest, local matching, and rich covariate sets for forming observational comparison groups in education settings. We conclude by offering readers with practical guidance for covariate selection in the design of education observational studies.

Excising Control in Observational Studies

In observational studies, the researcher begins with a comparison group that is not equivalent to the treatment group because individuals have self-selected – or were selected by a third party – to receive the intervention. The challenge is overcoming selection bias that occurs when treatment is not randomly assigned to units. For observational methods to produce valid results, two stringent conditions must be met. First, all covariates related to treatment assignment *and* units' potential outcomes must be identified and addressed by the researcher. For example, in an evaluation of an afterschool tutoring program for improving students' math skills, the researcher must consider all covariates related to students' participation in the program and their math scores. Some of these covariates may be observed and measured by the researcher (e.g. prior math experiences, demographic characteristics), but other factors are unobserved by the researcher, including students' latent math ability, personality factors, and parental attitudes towards academics. An observational study must account for all observed and unobserved factors related to treatment assignment and the outcome. The second requirement is that the probability of treatment selection must be between zero and one for each unit. This means that there are no units for which treatment assignment into the tutoring program is impossible, and there are no units for which participating in the control condition is impossible. If both of these conditions are met, then treatment assignment is independent of units' potential outcomes given observed covariates, and treatment assignment is said to be "strongly ignorable" (Rosenbaum & Rubin, 1983). In practice, the researcher never knows the full data generating process that links treatment assignment to the outcome. Moreover, there are no direct empirical tests to help the researcher ascertain whether these assumptions are met.

The observational study literature suggests methods for addressing selection on both observed and unobserved covariates. Matching methods and covariate adjustment yield valid estimates of causal effects when confounders related to both treatment selection and units' potential

outcomes are known and reliably measured by the researcher. In covariate adjustment procedures such as regression, treatment effects are estimated by fitting linear regression models that include all causally relevant background characteristics and indicators for treatment. However, these approaches can be badly biased when the model lacks key confounders, when the model is mis-specified, and when prediction occurs outside the range of values actually observed in the covariates (Rubin, 1997; Shafer & Kang, 2008). An alternative observational approach is to match treatment and comparison units with similar values on pretreatment covariates to improve balance and overlap, avoiding issues with extrapolation in regression. One challenge with matching, however, is that as the number of matching variables increases, so does the dimensionality of matches, making it difficult to find suitable matches for each treated unit. Propensity score matching (PSM) addresses the curse of dimensionality problem. In PSM, the researcher creates a single index of the treatment probability for each unit using a logit or probit of all relevant characteristics related to treatment assignment and the outcome. Treatment effects may be estimated in a number of ways, including matching treatment and comparison units with the "nearest" propensity score and taking the mean difference in outcomes between the matched units; weighting units by the inverse of the estimated propensity scores; stratifying treatment and comparison members into the five subclasses, calculating mean differences within strata, and weighting strata appropriately for an overall treatment effect; or using the propensity score as a covariate in the outcome model (see Stuart (2010) for review of PSM approaches). PSM's advantage over regression is that it encourages researchers to check for covariate overlap and balance to avoid problems with extrapolation. Still, for valid treatment effects to be produced, both regression and PSM rely on the crucial assumption that all covariates related to treatment selection and the outcome are observed and correctly modeled.

The observational literature also suggests methods for addressing selection on *unobserved covariates*, including difference-in-differences (DID) approaches when pretest measures of the outcome are available. Here, the researcher compares pre-post changes in the outcome for the treatment group with pre-post changes in the outcome for the comparison group. One benefit of DID is that time-invariant covariates are "differenced out" of the model, reducing the threat of selection from unobserved confounders that remain constant within units across time. Moreover, the inclusion of a comparison group can address programmatic, policy, and/or compositional changes that may occur at the same time as the intervention, possibly confounding results. Thus, the ideal comparison group is subject to the same confounding factors as the treatment group, but not to the treatment itself. Treatment effects are estimated by differencing pre-post changes in the

treatment and comparison groups (see Angrist and Pishke (2009) for discussion of DID approaches).

The identifying assumption of DID with a single pretest time point is that the trend in the outcome is the same for both the treatment and comparison groups. For example, if treatment units experienced a large drop in test scores the year before the intervention was introduced (perhaps encouraging schools to sign up for treatment), a DID approach that compares differences in pre-post scores for the treatment and comparison group is likely to overestimate effects. If two or more observations are available prior to intervention, one can assess whether both groups exhibit the same pre-intervention time points. When multiple pretests are available, then a comparative interrupted time series (CITS) may be used to estimate treatment effects. CITS help rule out selection due to unobserved time invariant covariates within units, as well as allow for adjustment of pre-intervention group-specific trends over time (see Shadish, Cook, & Campbell (2002) for discussion for ITS and CITS methods).

The observational literature also recommends comparing outcomes for units within the same geographic area (Heckman, Ichimura, Smith, & Todd, 1998; Smith & Todd, 2005). The idea here is to eliminate variation by matching students or units within the same school, school district, and metropolitan area. The early empirical literature on non-experimental methods suggested promise for this approach. For example, in WSCs that used experimental data from the job training literature, Heckman et al. (1998) found that observational approaches produced less biased results when units were matched within the same labor market. The result was replicated by Smith and Todd (2005) and by Bloom et al. (2005) in their WSCs using job-training datasets. The intuition for matching units within the same geographic area is similar to the use of pretest data in DID approaches described above. However, here, unobserved confounders that are shared across units within the same geographic area are "differenced" out of the model.

Finally, researchers may combine covariate adjustment and/or matching with DID or CITS approaches to address multiple threats to validity. For example, researchers may adopt a CITS approach in which treatment and comparison members are matched based on their pre-intervention trends on the pretest measure, or on other observable characteristics. The goal here is to reduce spurious changes in treatment and comparison units over time by making groups as similar as possible on observable characteristics at baseline.

Empirical Evaluations of Covariate Performance in Observational Studies

The credibility of observational results often rely on the researcher's own assessment of how well their covariates succeed in addressing selection bias. Above, we outline statistical methods for which the pretest, geographic locale, and rich covariate sets are central for estimating observational effects. But how well do these covariates perform in field settings, and under what contexts and conditions do each of these methods perform better? How many pretest measures should be used? Should the pretest be used as covariates in regression or matching models, or as repeated measures in DID or CITS models? How geographically local is "local matching" in education settings – within schools, school districts, metropolitan areas or the state? Does the inclusion of rich covariates representing multiple domains succeed in addressing selection bias? And, do demographic covariates commonly found in administrative datasets produce unbiased effects?

Using WSC designs that compare observational results to a causal benchmark estimate on the same target population, analysts have examined these questions and others about the performance of covariate-types in educational field settings. However, for the WSC design to produce valid estimates of non-experimental bias, several requirements must be met.[2] First, the WSC design must have variation in treatment assignment between the benchmark design and the observational method, where for example, units in the benchmark condition are randomly assigned into treatment conditions and units in the observational condition self-select into treatment. Second, units in the benchmark and observational conditions should not react to their own assignment status in the WSC design, nor to the assignment status of others. A violation of this assumption occurs if there are randomization effects in the benchmark arm of the WSC, where units react to being randomly assigned into treatment conditions. Third, the benchmark design must provide a credible causal estimate for evaluating the observational method. The WSC benchmark design does not require an RCT – it may be a well-implemented regression-discontinuity design – but the approach should provide a valid causal estimate of the same quantity for which the researcher is comparing in the observational study. Fourth, there should be no differences between observational and benchmark results, except for the difference in treatment assignment between the two WSC conditions. These requirements imply that WSC researchers should attend to all possible differences that may confound results, ensuring that outcomes in both study conditions are measured at similar times on the same scales, and that the same causal quantity should be compared.

[2] Cook, Shadish, and Wong (2008) introduce criteria for a causally interpretable WSC design. Wong and Steiner (under review) formalize the design and identify assumptions for three types of WSC design.

Methods

WSC Studies

For this review, we examined all published and unpublished studies that met our WSC criteria, and compared at least one observational estimate with one benchmark estimate to assess bias. The "benchmark" could be from an experiment, or from another approach that the researchers believe to have sufficient validity to qualify for benchmark status. Given that WSCs represent heterogeneous backgrounds, selection mechanisms, outcomes, and populations, we limit our sample to include only studies that are informative for understanding covariate performance in education observational studies with math and/or reading achievement outcomes. To this end, we include the following studies in our review: WSCs that include education interventions that take place in preschool to post-secondary settings; have math and/or reading achievement scores as outcomes; and meet requirements for a WSC evaluation outlined above. All eligible studies are summarized in Table 1.

Search Strategy

Our search strategy used databases such as Ebsco, PsycINFO, Sociological Abstracts, Science Direct, ERIC, and Web of Science: Cited Reference. Search terms included "within-study comparison*" and "design replication*". We also reviewed references from existing WSCs to identify studies we might have missed in the online search. Moreover, we talked to researchers interested in WSCs from various disciplines, asking them to review our list of identified studies and suggesting others that we may not have identified.

Bias Metric

To compare non-experimental bias across WSC studies, we created a measure that is based on a common metric, which we define as the estimated standardized difference in the benchmark and observational results, such that: $\hat{B} = \frac{\hat{\tau}_{ne} - \hat{\tau}_{re}}{\hat{s}}$, where $\hat{\tau}_{ne}$ and $\hat{\tau}_{re}$ are the non-experimental and experimental treatment effect estimates, and $\hat{s}$ is the standard deviation of the outcome estimated from the experimental control group. We adopt the effect size difference as our bias metric because it provides readers with a clear measure of the magnitude of bias that remains for each WSC result. When sufficient information is available, we provide standard errors of the estimated bias metric. Finally, we may refer to additional measures of correspondence in the narrative text, such as percent bias remaining in the observational study, as additional descriptive information.

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Empirical Performance of Covariate-types in Education Observational Studies

Summary of Studies

Our search yielded 12 WSC studies that examined covariate performance in education observational studies. Except for one study that used an RD as the benchmark (Somers, Zhu, Jacob, & Bloom, 2013), all others used an RCT or a cluster RCT as the benchmark for evaluating the observational study. Study settings took place in preschools, as well as in elementary and middle, and post-secondary schools. For WSCs that took place in elementary and middle schools, the outcomes were standardized math and/or reading scores. In post-secondary settings, the outcomes included the Test of Standard Written English, and measures of writing, vocabulary and math. Because of space constraints, we address two WSCs only briefly because these studies have already been summarized in prior WSC reviews (Aiken, West, Schwalm, Carroll, & Hsiung 1998), or because the data have been examined by other studies that we discuss (Hallberg, Cook, & Steiner, under review). For completeness, we include all quantitative results in the tables.

Performance of Pretest

The earliest WSCs examined the performance of pretest measures in the context of job training interventions (LaLonde, 1986), where the outcomes of interest included earnings and employment status. Results from these empirical tests were mixed. Although pretest measures appeared to perform better than regression or matching methods alone, it did not consistently yield unbiased results. However, earnings and employment status tend to be unstable over time (Morgan, Dickinson, Dickinson, Benus, & Duncan 1974; Duncan, Coe, Corcoran, Hill, Hoffman, & Morgan, 1984), and susceptible to validity threats such as history, maturation, and regression to the mean (Ashenfelter, 1978). More recent reanalyses (Dehejia & Wahba, 1999, 2002; Smith & Todd, 2005; Diamond & Sekhon, 2013) of the LaLonde data found that results are sensitive to sample choice, and that observational methods perform better when multiple pretest time points are available and when sufficient covariate balance is achieved.

But what of achievement outcomes that are common in education evaluations? These outcomes are designed to be psychometrically reliable and valid, and have high year-to-year correlations. Aiken, West, Schwalm, Carroll, and Hsiung (1998) conducted an early WSC with a sample of undergraduate students at a large American university. For an observational study with a single pretest score, Aiken et al. (1998) estimated non-experimental bias of -0.10 and -0.02 sds for two writing outcomes. They concluded that the method succeeded in replicating benchmark results. In this section, we review the empirical performance of pretest measures from more recent WSCs in

education settings. Here, pretest measures may be used as covariates in PSM or regression analyses; or, it may be used to address unobserved selection in CITS models. We begin by examining the performance of a single pretest covariate for addressing bias. We then consider the performance of multiple pretest measures, particularly cases in which pre-intervention trends may be accounted for in the model.

Single Pretest Measures. Fortson, Verbitsky-Savitz, Kopa, and Gleason (2012; 2015) used lottery data for admission into charter schools (treatment N = 635: control N =304). The experimental benchmark (Gleason, Clark, Tuttle, & Dwoyer, 2010) consisted of a lottery in which applicants were randomly chosen for admission to a charter school, while the observational comparison consisted of students who were not in the lottery but enrolled in the same feeder schools as charter school applicants (N=20,407). To be eligible for inclusion in the WSC sample, students in both conditions had to be enrolled in the feeder schools and have at least one year of pre-intervention data available. The outcomes were achievement scores in reading and math, and treatment effects were estimated using regression with one pretest score included as a control or matching variable. The authors assessed bias by comparing RCT estimates from the lottery study with those obtained using the observational approach. Across multiple comparison samples, they found that the effect size differences between observational and RCT results ranged from 0.15 to 0.16 sds for the reading outcome, and 0.09 and 0.10 sds for the math outcome (compared to .54 sds for both outcomes in the unadjusted comparison). Still, 17% of the bias remained[3] for math and 28% for reading, even after controlling for the pretest. So, Fortson et al. (2012) concluded that although the pretest was critical for reducing bias, it was not sufficient for eliminating all of the bias in the observational study.

St. Clair, Cook, and Hallberg (2013) utilized data from a cluster RCT (Konstantopoulos, Miller, & Van der Ploeg, 2013) that examined the effects of Indiana's formative assessment system on student achievement in mathematics and English Language Arts (ELA). The observational arm was constructed using school information from the experimental treatment group (N=32), as well as state achievement data from nearly all other schools that served 4th through 8th graders in Indiana, specifically schools that did not participate in the RCT evaluation and had not implemented a formative assessment system (N=976). St. Clair et al. (2013) found that a single school-level pretest

[3] Percent bias remaining is calculated by: $\frac{\hat{\tau}_{ne}-\hat{\tau}_{re}}{|\hat{\tau}_i-\hat{\tau}_{re}|} \times 100\%$, where $\hat{\tau}_{ne}$ and $\hat{\tau}_{re}$ are the non-experimental and experimental treatment effect estimates, and $\hat{\tau}_i$ is the initial unadjusted estimate from the non-experimental study.

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

measure substantially improved performance of the observational approach from the unadjusted model. They found that when a single school-level pretest measure was included in the regression model, the observational bias was -0.02 sds for math and it was -0.04 sds for ELA (compared to bias estimates of -0.55 sds (math) and -0.60 sds (ELA) for the unadjusted state comparison). Here, inclusion of a single pretest measure resulted in 3% remaining bias for math, and 7% for reading. The authors concluded that a single pretest measure was enough to eliminate almost all of the bias in the observational study.

Steiner, Cook, Shadish, and Clark (2010) used a WSC design to examine the role of covariate selection in addressing bias in observational studies. The authors reanalyzed WSC data from Shadish, Clark, and Steiner (2008), who implemented a WSC design with independent benchmark and non-experimental arms (Marcus, Stuart, Shadish, & Steiner, 2012; Wong & Steiner, under review). In this approach, volunteer undergraduate students in a psychology class were randomly assigned into an experimental or observational condition. Students in the RCT were randomly assigned again into a short vocabulary (N = 116) or math (N = 119) intervention, while students in the observational condition were allowed to select between a short math (N = 79) or vocabulary (N=131) intervention. Observational effects were estimated using regression or PSM approaches, and compared to the RCT effect. Because of concerns with possible testing effects, "proxy-pretests" were used to assess students' baseline vocabulary and math skills. These measures tested students on vocabulary and math skills, but were different from posttest measures in terms of content and scale. Student proxy-pretests were included as baseline covariates in the regression analyses, and as matching covariates in PS models. Treatment effects for the PS approaches were estimated using stratification, weighting, and regression-adjusted methods.

For the naïve comparison without adjustment, the observational study produced an effect size difference of 0.22 sds for vocabulary and for 0.30 sds for math. When proxy-pretest was included for the vocabulary outcome, non-experimental bias was 0.04 sds for the regression model, and 0.05 to 0.08 sds when PSM was used. For the math outcome, the adjusted effect size difference was larger – 0.20 sds when treatment effects were estimated using regression and between 0.18 and 0.21 sds when matching methods were used. Thus, using the pretest measure alone, the remaining bias in the observational study ranged from 11% to 29% for the vocabulary outcome, and for math, it was more than double – ranging from 61% to 71%.

Although the proxy-pretest reduced bias in Steiner et al. (2010), the math pretest did not perform as well as the pretest measures in Fortson et al. (2012; 2015) or in St. Clair et al. (2013).

Why? Steiner et al. conducted follow-up analyses and observed that proxy pretests were only weakly correlated with treatment selection, and moderately correlated with math and vocabulary outcomes. However, students' topic preference was strongly correlated with their treatment selection. In fact, the two most highly correlated constructs with treatment selection were students' preference for math over literature ($r = 0.38$) and enjoyment of mathematics ($r = -0.36$). Moreover, they found that when the observational study was constructed based on students' topic preference alone, much of the bias was eliminated. Here, bias for regression and PSM approaches ranged from 0.05 to 0.11 sds for vocabulary, and from -0.12 to -0.04 sds for math. These results highlight that although pretest measures often reduce bias in the observational study, its actual performance depends on its correlation with treatment selection and the outcome. In the case of math, topic preference was more highly correlated with treatment selection and the outcome, and was critical for addressing bias. Hallberg, Cook, and Steiner (under review) reanalyzed Shadish et al. (2008) and St. Clair et al. (2013) to compare correlations between pretest and treatment selection with the size of non-experimental bias. Their results replicated conclusions suggested by Steiner et al. (See Table 2 for summary of results).

Multiple Pretest Measures. Researchers also have examined the performance of *multiple* pretests in observational contexts (Shadish, Cook, & Campbell, 2002). Although it is common for education researchers to have only one or two pretest time points in an observational study, the emergence of state longitudinal data systems suggest that researchers will have access to many more time points of student achievement data. In these cases, researchers may use pretest information as baseline covariates in regression or matching approaches, or through DID or CITS approaches. We begin by focusing on the performance of multiple pretest measures when trend effects are not adjusted for in the model. We then look at the case when trend effects are included in the observational model, as in a CITS design.

Bifulco (2012) examined the performance of two years of pretest information using lottery data of New Haven magnet schools (Bifulco, Cobb, & Bell, 2009). The outcomes of interest were students' standardized reading achievement scores. In the RCT benchmark, student applicants were randomly assigned admittance into a charter school (treatment N=192; control N=347), and the observational arm consisted of comparison students from school districts within the same metropolitan area, or districts that were geographically farther away but demographically similar to treatment districts. For each of these comparison groups, Bifulco (2012) examined results using cross-sectional estimators, such as regression and propensity score methods (nearest neighbor,

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

caliper, and kernel density), as well as differences-in-differences approaches. For all cross-sectional estimators, Bifulco adjusted for students' demographic characteristics, including age, race/ethnicity, free-reduced lunch status, special education status, and gender. To assess pretest performance, he compared treatment effects from regression and matching approaches, with and without two years of student pretest information. For the DID models, Bifulco estimated treatment effects using matched and unmatched samples. Here, he used propensity score methods (nearest neighbor, caliper, and kernel density) to match treatment and comparison students, and examined differences in change scores for the subsample of matched students. He also estimated DID effects for the full unmatched sample.

For the DID models, when treatment and comparison units were matched using propensity scores, non-experimental bias ranged from -0.07 sds to 0.01 sds compared to -0.04 sds to 0.03 sds for the full, unmatched sample. For the cross-sectional models, Bifulco concluded that the pretest almost always substantially reduced bias from models that did not include pretest information. When multiple pretests were included in the matching model, the non-experimental bias ranged from -0.10 sds to 0.04 sds, depending on the choice of comparison samples and propensity score methods used. For the regression models, bias ranged from -0.05 sds to 0.01 sds – again, depending on the comparison sample used. However, in matching models where pretest scores *were not included*, the range of the non-experimental bias estimates was larger – from -0.19 sds to 0.15 sds. For regression models without the pretest, the effect sizes differences were from -0.08 sds to 0.14 sds. Overall, Bifulco notes that pretests improved the performance of the non-experimental methods, often by more than 80%. And, he notes that the magnitude of the bias estimates are similar, regardless of whether pretests are used as control covariates in cross-sectional methods, or to compute DID estimates.

However, are multiple pretest measures needed? Fortson et al. (2012; 2015) and St. Clair et al. (2013) examined the performance of a single pretest measure compared to multiple pretest scores in a series of cross-sectional regression models. In some cases, multiple pretest measures reduced bias over a single pretest measure, but in general, multiple pretest measures did not make much of a difference, except in a case where a pre-intervention trend was present. Fortson et al. found that compared to regression models with only one student-level pretest measure, the second year of pretest data reduced bias from 0.15 sds (single pretest) to 0.09 sds (two pretests) for the reading outcome. For the math outcome, the effect size difference was 0.09 sds regardless of whether one or two years of pretest information was included in the model. St. Clair et al. (2013) compared the

performance of up to six school-level pretest scores that were included as covariates in regression models. Overall, St. Clair et al. found that when the pre-intervention trend was linear and stable across time – as it was for the math achievement outcome – the inclusion of multiple pretest time points did not reduce observed bias beyond what was achieved by a single pretest alone (-0.02 sds for a single pretest versus -0.05 sds for two years of pretest data to -0.09 sds for all 6 years of pretests). However, when selection-maturation effects were evident, as was the case for their ELA outcome, additional pretest measures in the cross-sectional approaches increased non-experimental bias from -0.04 sds with a single pretest measure to -0.24 sds with all six pretest time points included in the model (and no adjustment for trend effects).

Because six pre-intervention time points were available in St. Clair et al. (2013), the authors also examined the performance of non-experimental methods when baseline trends were accounted for in the analysis. In their CITS analyses, they estimated treatment effects for the full sample, and for sub-samples of treatment and comparison schools matched on six pre-intervention time points. For the math outcome, the CITS had biases similar in magnitude to the cross-sectional models without trend adjustment. Here, the effect size difference for the CITS models were small, ranging from -0.08 sds to -0.01 sds. For the ELA outcome, however, the CITS models performed considerably better than the cross-sectional models, with bias estimates ranging from -0.04 sds to 0.07 sds for the CITS analyses. Finally, the authors noted that matching observations prior to CITS analyses made little difference in reducing bias. The unmatched CITS analyses produced an effect size difference of 0.07 sds for ELA and -0.01 sds for math. For the matched samples[4], non-experimental bias estimates ranged from -0.04 to -0.03 sds for ELA and -0.08 to -0.04 sds for math.

Finally, Somers, Zhu, Jacob, and Bloom (2013) examined the performance of pretests when six pre-intervention time points are available. They compared bias estimates for DID models that adjusted for pre-intervention school means, and CITS models that adjusted for both pre-intervention means and slopes. They also assessed the benefits of using PSM (nearest neighbor and radius) in conjunction with their DID and CITS analyses. The causal benchmark in this study was not a randomized experiment, but a regression-discontinuity design evaluating the effects of Reading First (Gamse, Jacob, Horst, Boulay, & Unlu, 2008) (N=69 treatment schools, N=69 control schools). Comparisons in the observational arm of the WSC were drawn from multiple sources: schools within the same state (N=611), schools within districts that were eligible for Reading First

[4] The authors matched treatment and comparison schools based on: (1) 6 years of pretest achievement data; and (2) 6 years of pretest achievement data, but with pretest scores closer to the intervention receiving greater weight.

(N=419), and schools that applied for Reading First grants but did not receive it (N=99). The authors looked at school-level outcome measures for both reading and math standardized test scores.

Overall, Somers et al. (2013) concluded that the observational approaches succeeded in replicating RD benchmark results, regardless of whether DID or CITS approaches were used. For the DID models across all comparison samples, effect size biases ranged from -0.07 to 0.03 sds for the math outcome, and -0.08 sds to 0.05 sds for the reading outcome. For the CITS models, non-experimental bias was comparable to the DID analyses. Here, the effect size differences ranged from -0.07 to 0.02 sds for the math outcome, and -0.09 to 0.03 sds to the reading outcome. This was likely because there was no evidence of baseline trend effects in either treatment or comparison group. Finally, the authors found that although matching units on observed baseline covariates did not reduce bias in treatment effect estimates, it helped improve precision of the observational effect estimates.

These findings suggest that although education researchers have reasons to be optimistic about the use of pretest measures for matching and covariate adjustment, the pretest alone does not guarantee that the observational study will produce unbiased results. When a single pretest measure is included, bias in the observational study ranged from -0.10 to 0.21 sds. Steiner et al. (2010) noted that pretest measures performed poorly in cases where its correlation with treatment selection was weak, as it was for selection into math and vocabulary training. When no trend effects were present, multiple pretest measures did not improve non-experimental bias (Fortson et al., 2012, 2015; math outcome for St. Clair et al., 2013), but in the case where trend effects were evident (ELA outcome for St. Clair et al., 2013), the CITS model was needed to replicate benchmark results. The magnitude of the bias for cross-sectional estimators (regression, PSM) was comparable to bias observed for DID and CITS approaches. Moreover, matching treatment and comparison cases in DID and CITS analyses did not appear to reduce bias, though they improved the precision of observational estimates (Somers et al., 2013). Taken together, one reason why education pretest covariates may perform well in observational studies is that they tend to have linear and stable functional forms that are relatively easy to model. However, this may not generalize to cases that have other less stable outcomes. For now, researchers should investigate the strength of the correlation of the pretest with the treatment selection and the outcome variables, allowing for flexibility in the regression model when enough pre-intervention time points are available.

Performance of Local versus Non-Local Comparison Groups

Results from the early WSC literature in the job training literature found that observational comparisons drawn from the same labor markets produced less biased treatment effects (Heckman, Ichimura, Smith, & Todd, 1998; Smith & Todd, 2005). In theory, local matching can address bias from both observed and unobserved omitted variables related to treatment status and the outcome. The assumption here is that these confounders are differenced out when comparisons are made within the same region, school district, or labor market. Local comparisons are a special case of within-unit matching approaches, such as difference-in-difference methods that are common in the econometrics literature. Recent WSCs in education have extended this work by examining the performance of local comparisons from the same school district, and from neighboring metro areas, and states.

In addition to examining the performance of the pretest, Fortson et al. (2012) and Bifulco (2012) examined bias for local and non-local comparisons using data from charter school lotteries. Fortson et al. (2012) looked at the performance of regression results with baseline controls when observational comparisons were drawn from the same feeder schools as treatment students, compared to when comparisons were drawn from different feeder schools but within the same school district. One concern with "local" comparisons in the same feeder schools is that these students may be fundamentally different from treatment students given that they had the option to participate in the lottery but did not. Comparison students from different feeder schools may have opted out of the lottery for reasons not related to their outcome (e.g. charter school was too far to attend), but may be different from treatment students on other dimensions. Overall, the authors concluded that the observational results between the two comparison groups were not qualitatively or statistically different from each other. Using the same regression model for both local and non-local comparisons, the authors estimated non-experimental bias of 0.07 for reading and 0.07 for math; for within-district comparisons, the estimated bias was 0.08 for reading and 0.09 for math. Thus, in the case where both comparison pools were already geographically local, drawing comparisons from within the same school versus different schools but within the same district did not seem to make much difference.

Bifulco (2012) compared observational results from comparisons that were: (1) from the same school district (comparison group 1: N = 4126), (2) from a different metropolitan area but in demographically similar school districts (comparison group 2: N = 5446), and (3) from the same metropolitan area but from school districts that were materially more advantaged than treatment

schools (comparison group 3: N = 9938). For each comparison group, he estimated treatment effects using the same series of cross-sectional and DID estimators. Because the method of analyses did not produce substantively different results, we report their ranges here. Across all cross-sectional and DID estimators that accounted for pretest scores, Bifulco found that bias ranged from -0.04 to 0.01 sds for comparisons from group 1; from -0.05 to 0.04 sds for comparisons drawn from group 2, and from -0.10 to -0.04 sds for comparisons in group 3. Thus, the non-local comparison with similar observable characteristics performed as well as the within-district comparison, and better than the comparison district in the same geographic area but with observably different demographic characteristics. These results underscore the fact that local comparisons are often preferred because they are assumed to be more similar to treatment cases than more distal comparisons. However, this may not always be the case, so researchers should also assess covariate balance on observable characteristics.

In a reanalysis of the experimental Indiana Formative Assessment data used by St. Clair et al. (2013), Hallberg, Wong, and Cook (under review) tested the performance of "local" comparisons within the same school district versus comparison schools that were matched from across the state. Here, "local" consists of all schools within the same school district without additional statistical adjustment or matching of units (N=51 schools). "Non-local" comparisons were drawn from across the state, but were matched using propensity score methods (nearest neighbor) on multiple pretest covariates at the student and school levels, as well as on demographic information about students, teachers, and schools (N=128). Overall, Hallberg et al. (under review) found that the *unadjusted local* comparisons performed as well as *non-local matched* comparisons from across the state. For the local within-district comparison, the effect size difference between the RCT and observational study was 0.02 standard deviations for ELA and -0.04 standard deviations for the math outcome. For the matched non-local state comparison, bias was 0.00 and -0.03 for ELA and Math respectively.

Results from Bifulco (2012) suggest that geographically local comparisons performed well as long as units were observably similar to treatment cases. When local comparisons are sufficiently different, researchers should consider comparisons that are geographically farther, but are similar to treatment cases on observable characteristics. Hallberg et al. (under review) examined this hypothesis by looking at the performance of a "hybrid" approach, which seeks to optimize the mix of local and non-local comparison units. In the hybrid approach (Stuart & Rubin, 2007), geographically local matches are chosen first, but if the local matches appear too different from treatment schools, they are discarded in favor of non-local matches that are matched using

propensity score methods. To implement this method, Hallberg et al. (under review) estimated propensity scores for all treatment and comparison schools in the observational sample. If the distance in the estimated propensity scores between treatment and comparison schools within the same district exceeded an optimal caliper threshold, then a non-local comparison school was selected using nearest neighbor PSM to replace the local comparison school. Hallberg et al. found that the hybrid approach eliminated most of the remaining bias for math and reading (effect size of .01 sds for both outcomes), though there was not much bias remaining from either local or focal matching from across the state.

Finally, Dong and Lipsey (under review) conducted a WSC using data from the RCT evaluation of the Building Blocks (BB) Pre-K math curriculum in Tennessee (treatment N=210; control N=195). Here, the outcome measures were the Research-based Elementary Math Assessment (REMA) and Woodcock Johnson Quantitative Concepts scale. Observational comparison groups for the WSC design were constructed from four sources. First, they used control group information from three *other* RCT evaluations of pre-K interventions with similar samples, timeframe, and measures: (1) "The Tennessee PCER study of the Bright Beginnings and Creative Curriculum pre-k programs" (Preschool Curriculum Evaluation Research Consortium, 2008), (2) "Experimental Evaluation of the Tools of the Mind Pre-K Curriculum" (Wilson & Farran, 2013), and (3) "Building Blocks (BB) Pre-K Math Curriculum" (Clements & Sarama, 2006) in Massachusetts and New York. To construct the observational comparisons, they deleted the RCT treatment students and used controls as non-equivalent comparisons for treatment cases in the Tennessee Building Blocks evaluation. The fourth non-equivalent comparison was constructed from a measurement study of pre-K students in Tennessee with the same baseline and outcome measures (Lipsey & Meador, 2013).

In the WSC observational arm, "local" comparisons consisted of Tennessee control students in the RCT evaluations of the PCER Study (N=195), Tools of the Mind (N=130), as well as Tennessee pre-K students in the measurement study (Sample 1: N=535; Sample 2: N=361). A fourth within-state group combined all TN students in the same comparison pool (N=860). For all local comparisons, the authors examined bias for the WJ-Quantitative Concepts outcome. Non-local comparisons were drawn from RCT control groups of Tools of the Mind evaluation in North Carolina (N=328), as well as control groups from the Building Blocks (BB) evaluations in Massachusetts (N=92) and New York (N=286), and a combined sample of NY and MA students. Here, they used the NC sample to examine bias for the WJ-Quantitative Concepts outcome, and the

MA and NY samples to examine bias for the REMA outcome. Observational treatment effects were estimated using regression and PSM approaches (weighting, optimal matching, and stratification) that adjusted for students' pretest information, as well demographic characteristics including age, test interval, sex, race, and English language status.

Overall, the authors found that matching within-state and out-of-state did not make a large difference in the performance of the observational method, and that the in-state estimates were more biased than the out-of-state results. In-state effect sizes ranged from -0.64 to 0.25 sds for the WJ-Quantitative Concepts outcome, while effect sizes ranged from 0.03 to 0.08 sds for the same outcome when comparisons were drawn from North Carolina. However, for the REMA outcome, bias ranged from -0.28 to -0.14 sds using other out-of-state samples. Still, given the small sample sizes in the RCT benchmark and non-experimental comparison groups, the range of the effect size differences are challenging to interpret due to the large standard errors.

In follow-up exploratory analyses, the authors considered why non-experimental methods performed so poorly for the sample of in-state comparisons. They found that bias was largest in cases where comparison samples were small (as was the case in TN Tools of the Mind, TN PCER, and MA BB), and when comparable matches on observable characteristics could not be found (even for within-state comparisons). The authors note that when balance and overlap were achieved on observed covariates, bias from the observational comparisons rarely exceeded 0.15 sds in either the in-state or out-of-state comparisons. This is an important finding about the implementation of PSM approaches, and one that is not systematically addressed in other WSCs. Future WSCs should examine the relationship between balance and overlap, and non-experimental bias.

Local comparisons are often prioritized as observational comparisons because units within the same school, school district or state are believed to be more similar than geographically distal units. However, Dong and Lipsey (under review) suggest the limits of local comparisons when there are not sufficient cases for matching to treatment units. When there are few comparisons available for matching, and when matched cases lack balance and overlap with treatment units on observed covariates, even local comparisons can produce badly biased estimates. In these situations, researchers may consider hybrid approaches that attempt to combine both local and matched comparisons (Hallberg et al., under review). Moreover, our review has defined "local" broadly, including treatment and comparison units within the same state (Dong & Lipsey, under review), metropolitan area (Bifulco, 2012), school district (Bifulco, 2012; Fortson et al., 2012; Hallberg et al.,

under review), or even within the same school (Fortson, et al., 2012). Further empirical evidence is needed for examining what it means to be "local" in education settings.

Performance of Theory of Selection

Observational approaches such as PSM and regression succeed only to the extent that the researcher has considered and identified all covariates related to treatment selection and the outcome, and has measured reliably all relevant constructs. To implement this approach, researchers should have a conceptual understanding of how and why units select into treatment conditions. This may be based on prior empirical research, substantive theory, or the researcher's own knowledge of the selection process. In addition, the researcher should have access to high quality observational data that measures all relevant constructs and outcomes of interest. In practice, however, education researchers rely on observational datasets that vary in terms of the type and quality of covariate information available. For example, many observational studies use administrative datasets that allow researchers to draw from large comparison pools of students and schools. However, these datasets include information only about students' and schools' demographic characteristics, such as their race/ethnicity, free-reduced price lunch status, and gender – although more recent administrative datasets also contain information about students' prior achievement scores. Researchers may also utilize information from national surveys, such as the Early Childhood Longitudinal Survey (ECLS-K/-B) or the National Longitudinal Study of Adolescent to Adult Health (Add Health). These datasets have rich covariate information representing multiple domains of interest, including students' prior academic achievement and experiences, personality characteristics, physical and social-emotional health, and home parenting practices. Still, even in these cases, there is no guarantee that all covariates related to treatment selection and outcome are included in the dataset and correctly identified and modeled by the researcher. Finally, researchers may create their own observational datasets. This has the advantage of allowing the researcher to consider all covariates that are theoretically related to treatment selection and the outcome prior to data collection, providing an opportunity for the researcher to gather information on key factors not found in most observational datasets.

In this section, we examine bias for the case where the researcher has a well-articulated theory of treatment selection, and construct measures that are plausibly related to treatment selection and the outcome, as well as for the case where the researcher has weak or no theory of selection, relying on generic "covariates of convenience" (e.g. student demographic characteristics) for constructing observational comparisons.

Rich Covariate Set. As discussed above, Shadish, Clark, and Steiner (2008) examined whether PSM methods could overcome participants' selection into a math or reading intervention when the selection process was carefully considered in advance. The WSC design had independent benchmark and non-experimental arms, where college students were randomly assigned into the RCT or observational conditions, ensuring that participants in both conditions were treated identically except for their assignments into treatment conditions. Students in the observational arm were allowed to select participation in a short vocabulary or math training. Hypothesizing that students would select a training based on their desire to participate in or avoid the math intervention, the authors reviewed the education and psychology literature to identify all plausible factors that might be related to students' math preference and achievement. Based on their theory of selection, they administered and collected baseline measures of student demographic characteristics, vocabulary and math aptitudes (proxy pretests), prior experience and achievement in math and literature courses, personality traits (extroversion, emotional stability, agreeableness, openness, and conscientiousness), math anxiety, and depression. To estimate treatment effects for the observational arm, they used PSM (stratification, weighting, and ANCOVA) and ANCOVA without matching. Prior to any statistical adjustment in the observational arm, the estimated non-experimental bias was 0.22 sds for vocabulary and 0.32 sds for math, suggesting that students positively selected into vocabulary and math trainings. Once the matching or ANCOVA models were applied using all covariates, bias ranged from -0.06 to 0.02 sds for vocabulary and -0.13 to 0.01 sds for math. These results led the authors to conclude that when a rich set of covariates plausibly related to treatment selection and the outcome are available, observational methods can approximate the results of an experimental benchmark.

Although the WSC design in Shadish et al. (2008) addresses most internal validity concerns about estimating non-experimental bias, its generalization may be limited given the laboratory-like setting, and short intervention period. Pohl, Steiner, Eisermann, Soellner, and Cook (2009) conducted a replication of the WSC approach used by Shadish et al. (2008), in which units were randomly assigned into benchmark and observational arms. As in the Shadish et al. study, observational cases were allowed to self-select into treatment, while benchmark units were randomly assigned. However, the replication study differed from Shadish et al. in several key ways. First, WSC participants were university students in Berlin, Germany instead of college students at an American university. Second, although the math intervention was the same in both Shadish et al. and Pohl et al. (2009), the vocabulary training in Pohl et al. was designed to improve English for German

students, instead of an English vocabulary lesson for American students. Third, the selection processes for the two studies differed. In Shadish et al., students volunteered to participate in the WSC to receive credit for an undergraduate psychology class. In turn, these students selected a topic (vocabulary versus math) in which they were more proficient. In Pohl et al., however, students could receive additional training in an academic area in which they needed to demonstrate competency (English vocabulary). Here, students with weaker English skills selected into vocabulary training, perhaps with the intention to receive additional tutelage. For the observational study, Pohl et al. used ANCOVA and PSM (stratification, and regression-adjusted) to estimate treatment effects based on 25 covariates that were hypothesized to be related to the selection mechanism. Covariates included demographic factors, proxy pretests, prior achievement variables, topic preference, and other psychological indices. The outcomes of interest were English and math outcomes. The authors found that for the math outcome, there was not much evidence of initial bias. Here, the naïve comparison produced an effect size difference of 0.06 sds. The covariate adjusted and matching approaches produced bias estimates that ranged from 0.00 to .11 sds. For the English vocabulary outcome, selection bias was evident. For the naïve comparison, the effect size difference was -0.38 standard deviations, and for the matched or covariate adjusted sample, it was between 0.00 and 0.07 sds. This suggests that when selection into treatment conditions was present, covariate adjustment or matching was able to eliminate nearly all of the bias. Furthermore, when there was no initial bias, covariate adjustment and PSM did not increase bias.

Despite the replication study by Pohl et al. (2009), the generalization of these results may be limited. First, both treatment and comparison samples were drawn from the same school, so students were likely quite similar at baseline. Second, treatment selection here may have been easier to model than what researchers typically face in field settings. Still, these examples suggest that when researchers prospectively consider the selection mechanism and collect rich covariate information, unbiased observational results may be achieved. Additional replications, however, are needed.

Demographic Covariates. In many observational studies, the carefully considered selection process and rich covariate sets exemplified by Shadish et al. (2008) and Pohl et al. (2009) are the exceptions rather than the rule. Observational studies often depend on simple predictors of convenience such as race/ethnicity, gender, free-reduced price lunch status, disability status, and gender. However, these approaches rarely succeed in replicating benchmark results in WSC evaluations. For example, Wilde and Hollister (2007) used RCT data from Tennessee's Student Teachers Achievement Ratio Project (Project STAR) (Word et al., 1990), which examined the effects

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

of class size on student reading and math achievement outcomes. Since students were randomly assigned into small class within schools, the authors used RCT results from 11 of the 79 schools in the original Project Star experiment as the causal benchmark for the WSC. To construct the comparison, they matched 367 experimental treatment students in the 11 WSC benchmark schools to a potential pool of 4,589 control group students in the 68 *other* Project Star schools. To estimate observational treatment effects, the authors used nearest neighbor PSM with regression-adjustment. However, the available covariates for estimating the propensity score were limited. They included: student sex, race, birth and year, free-reduced price lunch status; teacher education and experience; and urbanicity of school. No pretest information was available. Across the 11 WSC benchmark schools where Wilde and Hollister (2007) attempted to replicate experimental results, bias estimates ranged from -2.58 sds to 1.85 sds, where in 6 of the 11 cases, the difference between experimental and observational results were statistically significant. However, these individual school results were noisy due to small sample sizes. As such, the authors also examined the average non-experimental bias across all schools. Here, their mean non-experimental bias estimate was 0.17 sds, which was statistically significant. These results led Wilde and Hollister to conclude that the observational results did not succeed in replicating benchmark results.

Steiner et al. (2010), Fortson et al. (2012, 2015), and Bifulco (2012) also examined bias when the observational study was constructed using demographic characteristics alone. Steiner et al. found that student demographic information did not perform well on their own, removing at most 47% of bias and in one instance increasing bias 18% over the unadjusted model[5]. For vocabulary, bias ranged from 0.11 to 0.15, but for math, the bias ranged from 0.28 to 0.36 (depending on the analytic method used). This is compared to the initial bias of 0.24 sds for vocabulary and 0.26 sds for math. Fortson et al.'s regression analyses with only demographic characteristics produced non-experimental bias of 0.45 sds for the reading outcome, and 0.46 sds for the math outcome. And across all comparison pools, Bifulco's models produced effect size differences of between -0.08 and 0.14 sds for regression methods, and -0.19 and 0.15 sds for propensity score methods (for a reading outcome). Even here, the magnitude of the bias may have been mitigated because comparison pools were geographically local (within the same institution for Steiner et al., 2010, and within the same

[5] Percent change in non-experimental and unadjusted estimates is calculated as: $\frac{\hat{\tau}_{ne} - \hat{\tau}_i}{\left|\hat{\tau}_i\right|} \times 100\%,$ where $\hat{\tau}_{ne}$ and $\hat{\tau}_i$ are the same as before.

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

school district for Forston et al., 2012), or demographically similar to treatment units (as was the case in Bifulco) before any adjustment was needed. Wilde and Hollister's (2007) example suggests that even when comparisons are within state, observation methods that rely on demographic characteristics for covariate adjustment or matching can produce badly biased results.

Discussion

This paper reviews results from the most recent WSCs to uncover best practices for covariate selection in education observational studies. We have examined the performance of observational approaches using pretest measures on the outcome, geographically local units, and rich covariate sets that were plausibly related to treatment selection.

This review suggests the following findings for covariate selection in observational studies. First, matching units via the pretest reduces bias in observational treatment effect estimates in education settings. When the pretest is highly correlated with both treatment selection and the outcome, and has a stable and linear baseline trend, it is enough to eliminate bias completely. Second, the single pretest does not seem to increase bias in observational estimates, suggesting that while it is possible to increase bias by controlling for the pretest (Pearl, 2009; Steiner & Kim, 2014), we do not observe empirical evidence that it occurs in field settings. Third, although the pretest often reduces bias in observational approaches, it does not always eliminate it. Bias remains in cases where there are baseline trend effects, or when important selection covariates are omitted from the model. Steiner et al. (2010) demonstrate that for an observational method to produce unbiased treatment effects, researchers must also consider other critical covariates that are strongly related to treatment selection and the outcome. This finding highlights the importance of sensitivity tests (Rosenbaum, 1987) in observational studies to assess the robustness of treatment effects to possible omitted variables. Fourth, we find that local comparisons perform as well as matching of non-local units on observable characteristics. However, if the local comparison pool is not sufficiently large to produce similar matches, local comparisons can produce badly biased results. In these cases, it may be useful to consider "hybrid" matching approaches, where geographically local matches are chosen first, but if the local matches appear too different from treatment schools, they are discarded in favor of non-local matches that are closely matched on observed characteristics (Stuart & Rubin, 2007; Hallberg et al., under review). Fifth, when there is a theory of selection and a rich covariate set, observational methods perform well in replicating benchmark results. However, these results were demonstrated in only two WSC cases, and more replications are needed. Finally, these results confirmed earlier WSC findings from the job training literature that matching or adjustment based

demographic characteristics alone, such as age, sex, race/ethnicity, are insufficient for addressing bias in observational studies (Fortson et al., 2012, 2015; Steiner et al., 2010; Wilde & Hollister, 2007).

Because of space constraints, this paper does not address directly the quality of WSC designs for evaluating covariate choices. Cook et al. (2008) highlighted criteria for causally interpretable WSCs, and Wong and Steiner (under review) extend this work by presenting assumptions for WSC design approaches. Overall, we note that WSCs conducted post-2008 attended to many of the common validity threats that had challenged interpretations from earlier empirical evaluations. This included making attempts to demonstrate the validity of the benchmark for evaluating the non-experimental method, comparing outcomes that were measured at the same times and on the same metrics, and estimating the same causal estimand for both the benchmark and non-experimental conditions.

However, one challenge with comparing WSC results across studies is that analysts still lack consensus on methods for assessing correspondence between experimental and observational results. Moreover, the field does not have standard criteria for reporting WSC results, nor for interpreting the magnitude of bias from observational studies. In this paper, we addressed this issue by converting reported bias estimates into effect size differences between observational and benchmark results, and reporting standard errors whenever possible. Across the board, observational methods appear to produce biases that ranged from 0.00 to 0.20 standard deviations. Although these effect size differences appear small, they are comparable to what we have defined as meaningful effect sizes for successful interventions in education contexts (Hill, Bloom, Black, & Lipsey, 2008). Still, it is a limitation of this review that many of the bias estimates presented here do not have standard errors. Future syntheses of WSCs would be facilitated by more uniform standards for reporting and interpreting results. To this end, Steiner and Wong (under review) examine the performance of multiple measures for assessing correspondence in WSC results, and Rindskopf and Shadish (under review) propose a Baysian correspondence criteria for determining whether non-experimental and RCT results replicate.

This review demonstrates that although we are beginning to accumulate evidence on better observational design practice, for now, none of these design elements alone can guarantee unbiased results in field settings. As such, we recommend that researchers: 1) theoretically consider the selection process into treatment assignment and, if possible conduct a pilot study identifying all relevant covariates that are related to both treatment assignment and the outcome; 2) collect rich covariate information that includes multiple pretests and other measures believed to be related to

treatment assignment and the outcome; 3) conduct diagnostic checks to assess baseline equivalence and overlap in treatment and comparison cases once the matching procedure has been conducted; and 4) consider the robustness of their observational results given that biases as large as 0.20 standard deviations may remain.

This paper focuses on WSC results in education contexts. However, it is useful to consider WSC results from other fields of study to assess how well these results generalize to other types of outcomes and populations. Thus far, results have been less sanguine than what we have observed in education settings. In the areas of political science (Arceneaux, Gerber, & Green, 2010) environmental policy (Ferraro and Miranda, 2014), and in job training (LaLonde, 1986; Fraker & Maynard, 1987), the pretest has not performed well in replicating benchmark results. It remains unclear why in non-educational contexts, the pretest fails to produce comparable results to the RCT benchmark. It is possible that reading and math achievement scores in education contexts produce linear functions that are relatively easy to model, while outcomes such as voting behaviors in political science, water usage in environmental policy, and earning and employment outcomes in job training have more complex functional forms that are not well captured by DID approaches that are commonly used today.

Although researchers have used WSCs for evaluating the performance of observational approaches for nearly thirty years, the synthesis of these results for informing better observational practice remains a relatively new endeavor. We have shown that synthesis of WSC results can provide descriptive information about the contexts and conditions when non-experimental methods perform well in field settings, as well as demonstrate the limitations of observational methods. Moreover, this review highlights areas in which more empirical validation studies are needed. For example, while we observe seven studies examining the role of the pretest, and four examining the performance of local geographic matching, we have far less information about the performance of rich covariate sets, and none that use national observational data such as the ECLS-K. Furthermore, Dong and Lipsey (under review) showed that PSM approaches performed poorly when matched treatment and comparison groups have poor covariate balance and lack overlap in estimated propensity scores. The education observational literature would benefit from more WSCs that examine the performance of matching methods using different criteria for assessing balance and overlap in education observational settings. Moreover, questions about covariate selection in multi-level data (Hallberg, Cook, & Figlio, 2013; Steiner & Kim, 2015), as well as measurement error in covariates for propensity score matching (Steiner, Cook, & Shadish, 2011), are only beginning to be

addressed in the WSC context. As observational methods continue to develop, empirical validation of these methods are needed as well as the continued synthesis of these results for informing practice.

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

References

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation efficacy of a university-level remedial writing program. *Evaluation Review, 22*(2), 207-244. http://doi.org//10.1177/0193841X9802200203

Angrist, D. & Pishke, J. (2009) Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press.

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. http://doi.org/10.1177/0049124110378098

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics, 60*(1), 47-57. http://doi.org/10.2307/1924332

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology, 6*(2), 191-208. http://doi.org//10.1007/s11292-010-9095-3

Bifulco, R., Cobb, C. D., & Bell, C. (2009). Can interdistrict choice boost student achievement? The case of Connecticut's interdistrict magnet school program. *Educational Evaluation and Policy Analysis, 31*(4), 323-345. http://doi.org// 10.3102/0162373709340917

Bilfulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management, 31*(3), 729-751. http://doi.org/10.1002/pam.20637

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

Clements, D. & Sarama, J. (June 2006). *Scaling Up TRIAD: Teaching Early Mathematics for Understanding with Trajectories and Technology*, proposal funded by the Institute for Education Sciences.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-

study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750. http://doi.org/10.1002/pam.20375

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053-1062. http://doi.org//10.1080/01621459.1999.10473858

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151-161. http://doi.org//10.1162/003465302317331982

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932-945. http://doi.org//10.1162/REST_a_00318

Diaz & Handa (2006). An assessment of propensity score matching as a nonexperimental impact estimator. *The Journal of Human Resources, XLI*(2), 319-345. http://doi.org/10.3368/jhr.XLI.2.319

Dong, D., & Lipsey, M.W. (under review). How well propensity score methods approximate experiments using pretest and demographic Information in educational research?

Duncan, G.J., Coe, R., Corcoran, M., Hill, M., Hoffman, S., & Morgan, J.N. (1984). *Years of poverty, years of plenty: The changing economic fortunes of American workers and families*. Ann Arbor, MI: Institute for Social Research.

Ferraro, P.J., & Miranda, J.J. (2014). The performance of non-experimental designs in the evaluation of environmental policy: A design-replication study using a large-scale randomized experiment as a benchmark. *Journal of Economic Behavior and Organization, 107*, 344-365. http://doi.org/10.1016/j.jebo.2014.03.008

Ferraro, P.J., & Wichan, C. (Under Review). A cautionary tale on using panel data estimators to measure program impacts.

Fortson, K., Verbitsky-Savitz, N., Kopa, E. & Gleason, P. (2012). *Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates*. (NCEE Technical Methods Report No. 2012-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading first impact study*. Final Report. NCEE 2009-4038. National Center for Education Evaluation and Regional Assistance.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy, 589*, 63-93. http://doi.org/10.1177/0002716203254879

Gleason, P., M. Clark, C. Tuttle, and E. Dwoyer. *The Evaluation of Charter School Impacts*. National Center for Education Evaluation and Regional Assistance 2010-4029. Washington, DC: NCEE, Institute of Education Sciences, U.S. Department of Education, 2010.

Hallberg, K., Cook, T., & Figlio, D. (2013). Empirically examining the performance of approaches to multi-level matching to study the effect of school-level interventions. Presented at Society for Research on Educational Effectiveness Fall 2013 Conference, Washington DC.

Hallberg, K., Cook, T.D., Steiner, P., and Clark (Under Review). The role of pretests among other covariates in reducing selection bias in quasi-experiments: Evidence from between- and within- study contrasts.

Hallberg, K., Wong, V., & Cook T.D. (Under Review). Evaluating methods for selecting school level comparisons in quasi-experimental designs: Results from a within-study comparison.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*(5), 1017-1098. http://doi.org/10.3386/w6699

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177. http://doi.org//10.1111/j.1750-8606.2008.00061.x

Institute of Education Sciences. (2015). What Works Clearinghouse Data Query (Neil Seftor Personal Communication, 9/16/2015).

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis, 35*(4), 481-499. http://doi.org/10.3102/0162373713498930

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *The American Economic Review, 76*(4), 604-620. Retrieved from http://www.jstor.org/stable/1806062

Lipsey, M. W. & Meador, D. N. (2013). *Learning-related cognitive self-regulation school readiness measures for preschool children*. IES PI Meeting. Washington, DC.

Marcus, S. M., Stuart, E. A., Wang, P., Shadish, W. R., & Steiner, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods, 17*(2), 244. http://dx.doi.org/10.1037/a0028031

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics, 86*(1), 156-179. http://doi.org//10.1162/003465304323023732

Morgan, J.N., Dickinson, K., Dickinson, J., Benus, J., & Duncan, G. (1974). *Five thousand American families - Patterns of economic progress, Vol 1*. Ann Arbor, MI: Institute for Social Research.

National Science and Technology Council. (2011). *The federal science, technology, engineering, and mathematics (STEM) education portfolio, Pub. L. No. 111-358*. Washington, D.C.: Executive Office of the President.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. New York, NY: Cambridge University Press.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal Inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31*(4), 463-479. http://doi.org/10.3102/0162373709343964

Preschool Curriculum Evaluation Research Consortium (2008). *Effects of preschool curriculum A-4 programs on school readiness (NCER 2008-2009)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Rindskopf, D. & Shadish, W.R. (under review). Using Bayesian Correspondence Criteria to Compare Results from a Randomized Experiment and a Quasi-Experiment Allowing Self-Selection.

Rosenbaum, P.R. (1987). Model-Based direct adjustment. *Journal of the American Statistical Association, 82*, 387-394. http://doi.org/10.1080/01621459.1987.10478441

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. http://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P.R., & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38. http://doi.org//10.1080/00031305.1985.10479383

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), Oct 1974, 688-701. http://dx.doi.org/10.1037/h0037350

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(2), 757-763. http://doi.org//10.7326/0003-4819-127-8_Part_2-199710151-00064

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*, 1334-1356. http://doi.org/10.1198/016214508000000733

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

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16*(2), 179. http://dx.doi.org/10.1037/a0023345

Shafer, J.L. & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279-313. http://dx.doi.org/10.1037/a0014268

Smith, J. C., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305-353. http://doi.org/10.1016/j.jeconom.2004.04.011

Somers, M., Zhu, P., Jacob, R., & Bloom, H. (2013). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation (MDRC working paper in research methodology).* New York, NY: MDRC.

St Clair, T., Cook, T.D., & Hallberg, K. (2013). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation,* 1-17. http://doi.org/10.1177/1098214014527337

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. http://doi.org//10.3102/1076998610375835

Steiner, P.M., & Kim, Y. (2014, March). *On the bias-amplifying effect of near instruments in observational studies*. Paper presented at Spring Conference of the Society for Research on Educational Effectiveness, Washington, D.C.

Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2008). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267. http://dx.doi.org/10.1037/a0018719

Steiner, P.M. & Wong, V.C. (under review). Assessing Correspondence between Experimental and Non-Experimental Results in Within-Study Comparisons

Stuart, E.A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1), 1-21

Stuart, E. A., & Rubin, D. B. (2007). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics, 33*(3), 279-306. http://doi.org/1076998607306078v1

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management, 26*(3), 455-477. http://doi.org//10.1002/pam.20262

Wilson, S. J. & Farran, D. C. (2013). *Experimental evaluation of the Tools of the Mind preschool curriculum.* Paper presented at the SREE Spring 2012 Conference.

Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management, 32*(4), 853-877. http://doi.org//10.1002/pam.21721

Wong, V. C. & Steiner, P.M., (under review). Designs of Empirical Evaluations of Non-Experimental Methods in Field Settings.

Word, E., Johnston, J., Bain, H., Fulton, B., Zaharias, J., Achilles, C., et al. (1990). *The state of Tennessee Student/Teacher Achievement Ratio (STAR) Project final summary report (1985–1990).* Tennessee State Department of Education.

Table 1. Summary of Eligible WSC Studies for Review

Study	Benchmark	Design	Level of Analysis	Education Level	Outcome	Benchmark Sample Size
Aiken, West, Schwalm, Carroll, & Hsiung (1998)	Remedial Writing Course (Aiken et al., 1998)	RCT	Student	Post-Secondary	TSWE Writing Sample	107
Bifulco (2012)	Magnet School Lottery (Bifulco, Cobb, & Bell, 2009)	RCT	Student	Middle School	Reading	506
Dong & Lipsey (2014)	Building Blocks in TN (Clements & Sarama, 2006)	Cluster RCT	Student	Preschool	REMA WJ-QC	409
Fortson, Verbitsky-Savitz, Kopa, & Gleason (2012, 2015)	Charter Lottery (Gleason, Clark, Tuttle, & Dwoyer, 2010)	RCT	Student	Elementary/ Middle School	Math Reading	924 926
Hallberg, Wong, & Cook (under review)	Indiana Diagnostic Assessment Intervention (Konstantopoulos, Miller, & Van der Ploeg, 2013)	Cluster RCT	School	Elementary/ Middle School	ELA Math	63
Hallberg, Cook, Steiner, & Clark (under review)	Indiana Diagnostic Assessment Intervention (Konstantopoulos, Miller, & Van der Ploeg, 2013)	Cluster RCT	School	Elementary/ Middle School	ELA Math	63
	Vocabulary and Math Intervention (Shadish et al., 2008)	RCT	Student	Post-Secondary	Math Vocabulary	235
Pohl, Steiner, Eisermann, Soellner, & Cook (2009)	English and Math Intervention (Pohl et al., 2009)	RCT	Student	Post-Secondary	ELA Math	99
Shadish, Clark, & Steiner (2008)	Vocabulary and Math Intervention (Shadish et al., 2009)	RCT	Student	Post-Secondary	Math Vocabulary	235

Table 1 Continued. Summary of Eligible WSC Studies for Review

Study	Benchmark	Design	Level of Analysis	Education Level	Outcome	Benchmark Sample Size	
Somers, Zhu, Jacob, & Bloom (2013)	Reading First Impact Study (Gamse, Jacob, Horst, Boulay, & Unlu, 2008)	RD	School	Elementary	Math Reading	168	
St. Clair, Cook, & Hallberg (2013)	Indiana Diagnostic Assessment Intervention (Konstantopoulos, Miller, & Van der Ploeg, 2013)	Cluster RCT	School	Elementary/ Middle School	ELA Math	63	
Steiner, Cook, Shadish, & Clark (2010)	Vocabulary and Math Intervention (Shadish et al., 2009)	RCT	Student	Post-Secondary	Math Vocabulary	235	
Wilde & Hollister (2007)	Project STAR (Word, Johnston, Bain, Fulton, Zaharias, Achilles, et al., 1990)	Multisite RCT	Student	Elementary	Combined Math and Reading	117	School 1
						120	School 2
						106	School 3
						131	School 4
						136	School 5
						131	School 6
						118	School 7
						103	School 8
						138	School 9
						112	School 10
						108	School 11
						1320	Combined

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 2: WSC Results with a Single Pretest

	WSC Study	Benchmark Study	Outcome	Analytic Method	*n*[a]	Bias ES Difference[b]	
						Lower range	*Upper range*
1	Aiken, West, Schwalm, Carroll, and Hsiung (1998)	RCT of Remedial English Class among College Freshmen	Writing Assessment	Regression	99	-0.10	
						--	
			TSWE	Regression	99	-0.02	
						--	
2	Fortson, Verbitsky-Savitz, Kopa, & Gleason (2012, 2015)	Evaluation of Charter School Impacts	Reading	Regression	20,099	0.15	0.16
						--	--
			Math	Regression	20,335	0.09	0.10
						--	--
3	Hallberg, Cook, Steiner, & Clark (Under Review)	Evaluation of Indiana Formative Assessment	ELA	PSM	1040	0.03	
						--	
			Math	PSM	1040	0.02	
						--	
		RCT of Vocabulary or Math Training	Vocabulary	PSM	247	0.08	
						--	
			Math	PSM	198	0.21	
						--	
4	St. Clair, Cook, & Hallberg (2013)	Evaluation of Indiana Formative Assessment	ELA	Regression	1040	-0.04	
						--	
			Math	Regression	1040	-0.02	
						--	
5	Steiner, Cook, Shadish, and Clark (2010)	RCT of Vocabulary and Math Training	Vocabulary	Regression	247	0.04	
						(0.17)	
				PSM	247	0.05	0.08
						(0.17)	(0.17)
			Math	Regression	198	0.20	
						(0.19)	
				PSM	198	0.18	0.21
						(0.19)	(0.19)

[a] Sample size includes the experimental treatment group and the non-experimental comparison group used in the analysis. The smallest and largest samples are reported when different comparison groups were used to estimate quasi-experimental effect sizes.

[b] Bias ES difference is equivalent to the estimated quasi-experimental treatment effect minus the experimental benchmark result, divided by the standard deviation of the experimental control group mean. Both lower and upper bias effect size estimates are provided in instances when multiple applicable treatment estimates are reported in study. Standard errors are reported in parentheses below the effect size. "--" indicates the information was not reported.

Table 3: WSC Results with Multiple Pretests

	WSC Study	Benchmark Study	Outcome	Analytic Method[a]	*n*[b]	Bias ES Difference[c]	
						Lower range	*Upper range*
1	Bifulco (2012)	Connecticut Magnet School Lottery	Reading	PSM	4,284 - 10,096	-0.10	0.04
						(0.10)	(0.09)
				Regression	4,284 - 10,096	-0.05	0.01
						(0.04)	(0.04)
				DID+PSM	4,284 - 10,096	-0.07	0.01
						(0.04)	(0.05)
				DID	4,284 - 10,096	-0.04	0.03
						(0.04)	(0.04)
2	Fortson, Verbitsky-Savitz, Kopa, & Gleason (2012, 2015)	Evaluation of Charter School Impacts	Reading	Regression	12,716	0.09	
						--	
			Math	Regression	12,731	0.09	
						--	
3	Somers, Zhu, Jacob, & Bloom (2013)	Reading First Impact Study	Reading	DID	168 – 680	-0.08	0.05
						(0.06)	(0.07)
				CITS	168 – 680	-0.09	0.03
						(0.07)	(0.07)
				DID+PSM	127 - 432	-0.09	0.03
						(0.07)	(0.08)
				CITS+PSM	138 – 418	-0.11	0.01
						(0.07)	(0.08)
			Math	DID	168 – 680	-0.07	0.03
						(0.07)	(0.09)
				CITS	168 – 680	-0.07	0.02
						(0.07)	(0.09)
				DID+PSM	127 - 432	-0.12	0.00
						(0.08)	(0.09)
				CITS+PSM	138 – 418	-0.11	0.05
						(0.14)	(0.13)

Table 3 Continued: WSC Results with Multiple Pretests

	WSC Study	Benchmark Study	Outcome	Analytic Method[a]	*n*[b]	Bias ES Difference[c]	
						Lower range	*Upper range*
4	St. Clair, Cook, & Hallberg (2013)	Indiana Formative Assessment RCT	ELA	Regression (2 years)	1040	-0.07	
						--	
				Regression (3 years)	1040	-0.12	
						--	
				Regression (4 years)	1040	-0.15	
						--	
				Regression (5 years)	1040	-0.20	
						--	
				Regression (6 years)	1040	-0.24	
						--	
				CITS	1040	0.07	
						(0.14)	
				CITS+PSM	192	-0.04	-0.03
						--	--
			Math	Regression (2 years)	1040	-0.05	
						--	
				Regression (3 years)	1040	-0.07	
						--	
				Regression (4 years)	1040	-0.06	
						--	
				Regression (5 years)	1040	-0.08	
						--	
				Regression (6 years)	1040	-0.09	
						--	
				CITS	1040	-0.01	
						(0.17)	

[a] Indicates way in which pretest was incorporated into model. PSM=Propensity Score Matching; DID=Difference in Differences; CITS=Comparative Interrupted Time Series

[b] Sample size includes the experimental treatment group and the non-experimental comparison group used in the analysis. The smallest and largest samples are reported when different comparison groups were used to estimate quasi-experimental effect sizes.

[c] Bias ES difference is equivalent to the estimated quasi-experimental treatment effect minus the experimental benchmark result, divided by the standard deviation of the experimental control group mean. Both lower and upper bias effect size estimates are provided in instances when multiple applicable treatment estimates are reported in study. Standard errors are reported in parentheses below the effect size. "--" indicates the information was not reported.

	CITS+PSM	192	-0.08	-0.04
			--	--

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 4: WSC Results for Local vs. Non-Local Comparison Groups

	WSC Study	Benchmark Study	Outcome	Local[a]					Non-Local[a]				
				Locale	Comparison group[a]	*n*[b]	Bias ES Difference[c]		Locale	Comparison group	*n*[b]	Bias ES Difference[c]	
							Lower range	*Upper range*				*Lower range*	*Upper range*
1	Bifulco (2012)	Magnet School Lottery	Reading	Within-district		4,284	-0.04	0.01	Within-metro area		10,096	-0.10	-0.04
							(0.07)	(0.05)				(0.10)	(0.04)
									Within-State		5,604	-0.05	0.04
												(0.05)	(0.09)
2	Dong & Lipsey (2014)	Bulding Blocks Pre-K Math Curriculum	WJ-Quant Concepts	Within-state	TN Tools of the Mind	340	-0.64	-0.48	Within-country	NC Tools of the Mind	538	0.03	0.08
							(0.31)	(0.27)				(0.12)	(0.11)
					TN PCER	405	0.23	0.25					
							(0.14)	(0.13)					
					TN Measurement Study	1,106	0.11	0.15					
							(0.11)	(0.11)					
					TN All	1,431	0.04	0.12					
							(0.12)	(0.11)					
			REMA						Within-country	MA BB	303	-0.28	
												(0.21)	
										NY BB	497	-0.14	
												(0.21)	
										MA & NY BB	589	-0.24	-0.20
												(0.18)	(0.20)

[a] Dong & Lipsey provide separate treatment effects for samples of different size and composition within the local and non-local categories. These comparison groups are noted in the table.

[b] Sample size includes the experimental treatment group and the non-experimental comparison group used in the analysis. The smallest and largest samples are reported when different comparison groups were used to estimate quasi-experimental effect sizes.

[c] Bias ES difference is equivalent to the estimated quasi-experimental treatment effect minus the experimental benchmark result, divided by the standard deviation of the experimental control group mean. Both lower and upper bias effect size estimates are provided in instances when multiple applicable treatment estimates are reported in study. Standard errors are reported in parentheses below the effect size. "--" indicates the information was not reported.

Table 4 Continued: WSC Results for Local vs. Non-Local Comparison Groups

	WSC Study	Benchmark Study	Outcome	Local: Locale	Local: Comparison group	Local: n^a	Local: Bias ES Difference[b] Lower range	Local: Bias ES Difference[b] Upper range	Non-Local: Locale	Non-Local: Comparison group	Non-Local: n^a	Non-Local: Bias ES Difference[b] Lower range	Non-Local: Bias ES Difference[b] Upper range
3	Fortson, Verbitsky-Savitz, Kopa, & Gleason (2012, 2015)	Evaluation of Charter School Impacts	Reading	Within-school		20,729	0.07	--	Within-district		143,070	0.08	--
			Math			20,964	0.07	--			143,826	0.09	--
4	Hallberg, Wong, & Cook (under review)	Indiana Formative Assessment RCT	ELA	Within-district		83	0.02 (0.09)		Within-state		160	0.00 (0.08)	
			Math				-0.04 (0.11)					-0.03 (0.09)	
			ELA						Hybrid approach (district and state)		990	0.01 (0.07)	
			Math									0.01 (0.09)	

[a] Sample size includes the experimental treatment group and the non-experimental comparison group used in the analysis. The smallest and largest samples are reported when different comparison groups were used to estimate quasi-experimental effect sizes.

[b] Bias ES difference is equivalent to the estimated quasi-experimental treatment effect minus the experimental benchmark result, divided by the standard deviation of the experimental control group mean. Both lower and upper bias effect size estimates are provided in instances when multiple applicable treatment estimates are reported in study. Standard errors are reported in parentheses below the effect size. "--" indicates the information was not reported.

Table 5: WSC Results with a Rich Covariate Set and a Theory of Selection

	WSC Study	Benchmark Study	Outcome	Analytic Method	*n*[a]	Bias ES Difference[b]	
						Lower range	***Upper range***
1	Pohl, Steiner, Eisermann, Soellner, & Cook (2009)	RCT of English and Math Training	English	PSM or Regression	99	0.00	0.07
						(0.22)	(0.28)
			Math		103	0.00	0.11
						(0.30)	(0.27)
2	Shadish, Clark, & Steiner (2008)	RCT of Vocabulary and Math Training	Vocabulary	PSM or Regression	247	-0.06	0.02
						(0.18)	(0.18)
			Math		198	-0.13	0.01
						(0.18)	(0.22)

[a] Sample size includes the experimental treatment group and the non-experimental comparison group used in the analysis. The smallest and largest samples are reported when different comparison groups were used to estimate quasi-experimental effect sizes.

[b] Bias ES difference is equivalent to the estimated quasi-experimental treatment effect minus the experimental benchmark result, divided by the standard deviation of the experimental control group mean. Both lower and upper bias effect size estimates are provided in instances when multiple applicable treatment estimates are reported in study. Standard errors are reported in parentheses below the effect size. "--" indicates the information was not reported.

Table 6: WSC Results with Demographic Covariates Only

	Study	Data Source	Outcome	Analytic Method	n[18]		Bias ES Difference[19]	
							Lower range	*Upper range*
1	Bifulco (2012)	Connecticut Magnet School Lottery	Reading	Regression	4284 – 10,096		-0.08 (0.05)	0.14 (0.04)
				PSM			-0.19 (0.09)	0.15 (0.09)
2	Fortson, Verbitsky-Savitz, Kopa, & Gleason (2012, 2015)	Evaluation of Charter School Impacts	Reading	Regression	20,729		0.45	--
			Math	Regression	20,964		0.46	--
3	Steiner, Cook, Shadish, and Clark (2010)	RCT of Vocabulary and Math Training	Vocabulary	Regression or PSM	247		0.11 (0.19	0.15 (0.18)
			Math	Regression or PSM	198		0.28 (0.19)	0.36 (0.20)
4	Wilde & Hollister (2007)	Tennessee Class Size Experiment (Project Star)	Combined Math & Reading	PSM	896	School 1		0.59 --
					909	School 2		0.41 --
					913	School 3		-2.58 --
					878	School 4		-0.42 --
					865	School 5		1.85 --
					934	School 6		-0.28 --
					891	School 7		-1.77 --
					906	School 8		-1.54 --
					913	School 9		0.05 --

[18] Sample size includes the experimental treatment group and the non-experimental comparison group. The smallest and largest samples are reported when different comparison groups were used to estimate quasi-experimental effect sizes.

[19] Bias ES difference is equivalent to the estimated quasi-experimental treatment effect minus the experimental benchmark result, divided by the standard deviation of the experimental control group mean. Both lower and upper bias effect size estimates are provided in instances when multiple applicable treatment estimates are reported in study. Standard errors are reported in parentheses below the effect size. "--" indicates the information was not reported.

899	School 10	0.56
		--
893	School 11	0.55
		--
--	Average	0.17
		--

EdPolicyWorks Working Paper Series No. 45. April 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia