4-3-2017

Evaluating School Vouchers: Evidence from a Within-Study Comparison

Kaitlin P. Anderson
Michigan State University, ande2018@msu.edu

Patrick J. Wolf
University of Arkansas, Fayetteville, pwolf@uark.edu

Citation

Anderson, K. P., & Wolf, P. J. (2017). Evaluating School Vouchers: Evidence from a Within-Study Comparison. *Education Reform Faculty and Graduate Students Publications.* Retrieved from https://scholarworks.uark.edu/edrepub/10

UNIVERSITY OF ARKANSAS
College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

Evaluating School Vouchers:
Evidence from a Within-Study Comparison

Kaitlin P. Anderson

Patrick J. Wolf, Ph.D.

April 3, 2017

EDRE Working Paper 2017-10

Evaluating School Vouchers: Evidence from a Within-Study Comparison

Kaitlin P. Anderson
University of Arkansas

Patrick J. Wolf
University of Arkansas

Abstract

Randomized Controlled Trials (RCTs) are the "gold-standard" for estimating causal impacts of educational programs. Students subject to lotteries, however, often are not representative of the broader population of students experiencing the educational treatment. With few exceptions, researchers are not able to determine how much selection bias exists when various quasi-experimental approaches are used in place of experimental ones within a school choice context. We are left wondering about the magnitude of the internal-for-external validity tradeoff that education researchers often face.

This study assesses the extent to which methods such as propensity score matching or observational models with control variables can replicate the "benchmark" experimental results of the District of Columbia Opportunity Scholarship (DC OSP) school voucher evaluation. The federal private school voucher program is an exemplar subject for study because self-selection is assumed to be a major influence on whether or not a low-income urban student attends a private school. We treat Instrumental Variables Analysis (IV) estimates of the impact of private schooling on student outcomes, some of which are being presented for the first time in this study, as the causal "benchmark" estimate. While our data are fairly limited, and the results relatively imprecise, we find preliminary evidence that covariate choice matters, and that method choice matters, but perhaps only when comparing to a broader sample that includes students who did not apply to the program.

Interestingly, we find that the direction of the estimation bias that we detect from some of the quasi-experimental approaches is positive when the sample is limited to program applicants, but negative when it is expanded to include non-applicants. This finding suggests that the applicants to means-tested school voucher programs are negatively selected, but the subgroup of applicants who actually use a voucher if offered tend to be positively selected.

Keywords: school vouchers, school choice, within-study comparison, randomized controlled trial, quasi-experimental design, internal validity, external validity, selection bias

Introduction

Randomized Controlled Trials (RCTs) are commonly used to estimate causal impacts of educational programs, and have been called the "gold-standard" of evaluation (Mosteller & Boruch, 2002; Rossi, Lipsey & Freeman, 2004). School districts or programs often use lotteries to determine access to oversubscribed programs, permitting rigorous RCT evaluations of program impacts to be conducted. Lottery-based RCTs eliminate the potential self-selection bias associated with participation in voucher programs or charter schools, as mere chance replaces parental motivation as the factor that determines whether a child gains access to school choice. The randomized control group (a.k.a. lottery losers) becomes the ideal counterfactual, representing what would have happened to the randomized treatment group (a.k.a. lottery winners) absent the intervention. Due to the elimination of self-selection bias, properly implemented lottery-based RCTs have strong internal validity.

A major limitation of education RCTs, however, is their external validity. The students subject to lotteries often are not representative of the broader population of students experiencing the educational treatment (e.g. Abdulkadiroglu et al., 2009). With few exceptions (Bifulco 2012; Forston et al., 2012), researchers are not able to determine how much selection bias is introduced when various quasi-experimental approaches are used in place of experimental ones within a school choice setting. We are left wondering about the magnitude of the internal-for-external validity tradeoff that education researchers often face.

This study contributes to our understanding of this key methodological concern by assessing the extent to which quasi-experimental methods such as propensity score matching or observational models with control variables can replicate the "benchmark" experimental results of the original District of Columbia Opportunity Scholarship (DC OSP) evaluation conducted from 2004-2009 (Wolf et al., 2010). The federal private school voucher program is an especially appropriate subject for such a methodological study because self-selection is assumed to be a major influence on whether or not a low-income urban student attends a private school. Less than 22 percent of OSP voucher recipients in 2004 were lottery winners (Wolf et al., 2005, p. 24), however, leaving us wondering if the unbiased experimental estimates of the program's impacts drawn from students who faced lotteries similarly apply to students who were not subject to lotteries. Should we prioritize internal or external validity in this case? Can we have both?

Literature Review

Within-study comparisons are motivated by the internal-external validity divide regarding evaluation methodologies. Internal validity is the great virtue of experimental approaches. Because the offer of the experimental treatment and mere chance are the only factors that distinguish a randomized treatment group from a randomized control group, the two comparison groups are similar in all relevant respects in expectation (Cook & Campbell, 1979, p. 56; Cook & Payne, 2002). The key consequence of randomizing a large number of study participants is that the control group becomes the ideal counterfactual, demonstrating the fate that would have befallen the treatment group if not for the treatment offer. Any differences observed between the treatment and control groups, post-randomization, can be assumed to have been actually caused by the intervention of the treatment, and not any confounding factor, so long as the experiment was implemented successfully.

Internal validity "refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause" (Cook & Campbell, 1979, p. 37). Causality is central to the consideration of internal validity. Because experiments provide a sound foundation for identifying where causal relationships apparently do and do not exist, we speak of randomized experiments as having strong internal validity (Egalite & Wolf, 2016; Sadoff, 2014).

Few randomized experiments also have strong external validity, however. External validity "refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized…across different types of persons, settings, and times." (Cook & Campbell, 1979, p. 37) Just as internal validity is the ability to know, with confidence, that a causal relationship actually exists, external validity signals the size of the population for whom the relationship likely exists. Experiments often involve special populations eligible for a targeted program intervention piloted in a particular place. Since context often mediates the experimental effects of programs (e.g. Gleason et al., 2010), and population characteristics often moderate those effects (e.g. Howell et al., 2002), the strong internal validity of experiments with weak external validity can represent a pyrrhic victory. Ideally, we do not simply want to know what impact a program has on distinctive populations in particular places. We want to know how it will affect lots of different people at scale. We do not want to have to choose our validity (Foreman, Anderson, Ritter & Wolf, 2017). We want it all.

Quasi-experimental designs (QED) would seem to be the solution to our conundrum. Since QEDs do not rely on the randomization of distinctive populations of participants, they tend to have strong external validity. Moreover, QEDs are called "quasi" experimental because they employ one or more techniques that promise to approximate experimental impact estimates at least under certain conditions or subject to key assumptions. A QED can have strong internal validity as well as strong external validity; but its' internal validity is by no means assured. Potential confounds lead many education researchers to question the internal validity of QEDs, especially in school choice evaluations, where unmeasured self-selection bias is a particular concern (Shakeel, Anderson & Wolf, 2016; Barrow & Rouse, 2008; Levin 1998). We cannot assume causality from QEDs. Causality must be inferred and such causal inferences are subject to challenge. Enter the within-study comparison.

Within-study comparisons (WSGs) use both experimental and quasi-experimental or observational methods to evaluate the same intervention. Impact estimates from a successfully implemented randomized experiment are presented as the "true" or "benchmark" causal effects, and alternative non-experimental methodologies are evaluated based on their ability to generate effect estimates that are similar to the experimental results. If a non-experimental methodological approach with strong external validity largely replicates the findings of an experimental analysis in evaluating the same intervention we can have at least some confidence that the quasi-experimental results are both causal and broadly generalizable.

A variety of WSG methods have been used to assess the extent to which QED methods replicate experimental findings (e.g. Jaciw, 2016; Cook, Shadish, & Wong, 2008; Bloom, Michalopoulos, & Hill, 2005; Glazerman, Levy, & Myers, 2003). The first such studies were by Lalonde (1986) and Fraker and Maynard (1987), with both WSGs focused on job training and employment interventions. Although the complete literature on WSGs is too large to review here, it does point to certain common findings that apply to evaluations of education interventions such as the private school choice program we focus on here.

First, non-experimental estimates are less biased when comparison groups are geographically "nearer" to the experimental sample (Jaciw, 2016; Cook, Shadish, & Wong, 2008; Shadish, Clark, & Steiner, 2008; Aiken et al., 1998; Heckman et al., 1998; Heckman, Ichimura, and Todd, 1997). Forming comparison groups from national data sets, which violates this maxim, produced results very dissimilar from experimental results of job training programs

(Fraker & Maynard, 1987; Lalonde, 1986). Geographic context matters in education as well, as unmeasured self-selection factors tend to cluster in neighborhoods such that non-experimental methods fail to approximate experimental results when student school district or census tract is not used to construct comparison groups (Bifulco 2012; Witte et al. 2014).

Second, the selection of covariates is vital for reducing bias in QEDs (Shadish, Clark, & Steiner, 2008; Smith & Todd, 2005). For example, in many cases, propensity score matching using a standard set of demographics (but not baseline test scores) performs poorly in reproducing the experimental effects of education programs (e.g., Bifulco, 2012; Shadish, Clark, & Steiner, 2008; Wilde and Hollister, 2007). The point of control variables is to address selection bias. Selection bias occurs when one or more characteristic of program participants simultaneously influences both program access and the outcome used to evaluate the program. Variables that are related to outcomes or that predict self-selection into a program are the most important for approaching experimental estimates with QED methods (Cook et al., 2008; Glazerman et al., 2003). Cook et al. (2008) further argue that pretreatment (i.e. baseline) outcome measures are generally stronger predictors of posttreatment outcomes in educational interventions focused on academic achievement than on job training programs, and that in general we would expect pretreatment outcomes to better control for selection on unobservables within an education context.

Third, a variety of circumstances and modeling choices appear to matter. Pirog et al. (2009) conclude that propensity score matching and difference-in-differences approaches do not consistently reproduce experimental results because they are sensitive to modeling and sampling frame choices. Two WSC studies that compared propensity score analysis to experimental estimates found that propensity score matching performed poorly (Agodini & Dynarski, 2004;

Wilde & Hollister, 2007). The treatment and comparison groups in those WSGs were drawn from different geographic settings, however, and did not include pretreatment measures. It is possible that those two violations of WSG best-practices, more so than propensity score matching itself, were the culprits in preventing the replication of experimental results. Other studies that pull treatment and comparison group members from similar local settings and use pretreatment outcomes measures find that nonexperimental methods can closely match experimental estimates (Bifulco, 2012; Shadish, Clark, & Steiner, 2008; Aiken et al., 1998).

Fourth, findings from WSGs do not always generalize from one field to another, or even across subjects within education. Even well-executed WSGs can lack their own external validity. Steiner, Cook, Shadish, & Clark (2010) find that selection bias in QEDs of mathematics interventions can be removed using baseline characteristics that reflect topic preference, but that both topic preference and a proxy for pretest score are required to reduce the bias in QEDs of vocabulary outcomes.

While there is a variety of literature already available on this topic, more work is needed, particularly in the area of school choice interventions. Disputes regarding the appropriateness of experimental versus QED analytic approaches have raged throughout the nearly 30-year history of school choice research in the U.S. Witte's (1995) evaluation of the first school voucher program, in Milwaukee, was criticized by subsequent researchers for using quasi-experimental analytic samples and methods when experimental ones were available (Greene, Peterson & Du, 1999; Rouse, 1998). Witte (2000) responded that the experimental samples were too small to yield internally valid results and too particular to produce externally valid ones. Hoxby (2009) objected to the QED methods used in the National Charter School Study (CREDO, 2009), with outside scholars continuing to weigh-in on the dispute (e.g. Ackerman & Egalite, 2016).

Education researchers continue to divide over the claim that randomized experiments are the "gold standard" for evaluating private school choice programs (e.g. Egalite & Wolf, 2016; Lubienski, 2016).

Although debates over experimental versus quasi-experimental methods have been so heated surrounding school choice research, we are aware of only two WSCs of a school choice intervention (Bifulco, 2012; Fortson et al., 2012). Bifulco (2012) used data from two interdistrict magnet middle schools near Hartford, Connecticut and compared estimated impacts on reading scores in grade 8 using nonexperimental and experimental methods. His findings support some of the lessons learned from WSCs on other topics. For two of three comparisons used, the nonexperimental analyses yielded bias as high as 56 percent of the causal effect estimated under random assignment. Including pretreatment outcome measures reduced the bias by as much as 96 percent. In addition, Bifulco (2012) found that the information used to match students or adjust samples is more important than the particular QED method employed, because propensity score methods, regression analyses, and difference-in-differences estimators provided similar results. He also concluded that comparison groups from the same or local settings perform well in helping QEDs to approximate experimental results.

Fortson et al. (2012) compared four approaches – OLS regression, exact matching, propensity score matching, and fixed effects – to experimental results within an evaluation of charter schools. Comparison group members for the QEDs were drawn from the same school districts as the charter school students in the baseline period. While they also found that pretreatment outcome measures greatly reduced the bias, they reported more differences in bias levels across methods than Bifulco (2012). Fortson et al.'s regression-based nonexperimental impact estimates were significantly different from the experimental impact estimates, while their matching estimators performed slightly better.

The results of WSCs overall are not comforting to those hoping to bridge the internal-external validity divide by relying on quasi-experimental evidence for causal inference. A meta-analysis of WSCs published almost 15 years ago found that even in analyses with a rich set of covariates and pretreatment outcome measures, quasi-experimental and experimental methods often produced estimates that differ by policy-relevant magnitudes (Glazerman, Levy, & Myers, 2003). For now, if we want evidence to inform the experimental versus non-experimental methodological debate in a salient policy field such as school choice, we and other researchers will need to produce it on a study-by-study basis. That is our purpose here.

Data and Methods

This study compares the performance of quasi-experimental methods such as propensity score matching or observational models with control variables to "benchmark" experimental results from an evaluation of the District of Columbia Opportunity Scholarship Program (DC OSP) conducted from 2004-2009 (see Wolf et al., 2010).

According to Cook et al. (2008, 728-729), WSCs are most instructive when the following conditions obtain:

1. Various counterfactual groups to the treatment group are possible with one being the result of random assignment (i.e. "control") and one or more selected through a non-random process ("comparison").

2. All models use the same general effect estimator, such as the Local Average Treatment Effect (LATE).

3. The control and comparison groups are consistent in key factors such as the conditions under which variables are measured and geographic location.

4. The analysts estimating the experimental and quasi-experimental effects of the program are different and blind to each other's results.

5. The experiment is implemented successfully, with "no differential attrition or treatment crossovers."

6. The quasi-experiment(s) is implemented successfully, with appropriate matching algorithms, matching variables, and control variables.

7. The results of the various estimations are compared regarding their pattern of statistical significance, effect sizes, and proportional difference in effect sizes.

We admit that criterion 4 from this list is violated in our case, as the same analyst produced both the experimental estimates, which attempt to replicate those from a prior analysis, and the quasi-experimental effects from the data. The experiment was not perfectly implemented, as required of criterion 5, as there is differential sample attrition. Still, those experimental performance parameters fell within the range of what the Institute for Education Science's What Works Clearinghouse judges to be acceptable for generating causal estimates without qualification. Our WSC analysis fully satisfied the remaining five criteria.

We utilize student-level data contained in the restricted-use file associated with the original DC OSP evaluation. Student assessment data come from two sources: Stanford Achievement Test- version 9 (SAT-9) scores from DC Public Schools (DCPS) for the baseline year (2003-04) and the first outcome year (2004-05), as well as SAT-9 test scores for the two cohorts of the DC OSP subsample (Cohort 1: 2003-04 through 2007-08, Cohort 2: 2004-05 through 2008-09). All test-scores are standardized by grade level and subject to z-scores with a mean of zero and standard deviation of one. Currently, we only have DCPS data for two years, so we conduct two sets of analyses: 1) a comparison over four outcome years using the restricted, experimental sample, which includes both a 2003-04 and 2004-05 cohort and 2) a comparison for one outcome year for an unrestricted sample which includes in its comparison group DCPS students who never applied to the DC OSP.

DCPS and DC OSP data were merged and cleaned to have comparable student demographic information variables. For example, while DCPS reports free-and reduced-price lunch (FRL) status for students, the DC OSP dataset included a measure of household income. We converted this household income to FRL status for comparability, using federal guidelines.[1] If lunch status, gender, or special needs status (special education or limited English proficiency) was missing in the baseline year (2003-04), it was backfilled based on 2004-05 data. Missing grade level indicators for any year were deductively imputed based on the grade information available for other years. Further, in certain year-sub-sample combinations, scale scores were not reported, so raw scores were translated into scale scores for this analysis, and then translated into z-scores.

Samples differ by analysis. In our restricted, experimental samples (eight in total representing math and reading samples over four outcome years), we have lottery indicators and are able to conduct instrumental variables (IV) analysis as our “benchmark” treatment-on-treated effect. We refer to these eight samples interchangeably as the restricted samples, experimental samples, or IV samples. Larger samples, not restricted to those for whom IV analysis was possible, are referred to as our “unrestricted sample,” but the sample sizes differ by analysis type.

[1] Families with household income less than 130% of the federal poverty line are eligible for free lunch, and families with household income of 130% to 185% of the federal poverty line are eligible for reduced-price lunch (https://aspe.hhs.gov/prior-hhs-poverty-guidelines-and-federal-register-references).

Thus we have two components to the WSC, one that holds the sample constant across analytic methods and another that permits the quasi-experimental methods to draw from a more expansive sample, consistent with the goal of enhancing external validity.

In Table 1 we report observable characteristics for the lottery winners and losers, separately for the first and second cohorts. Lottery winners and losers differed by grade level in both cohorts 1 and 2, and there were some differences in baseline test scores in cohort 2.

Table 1: Lottery covariate balance, by cohort

	Lottery Statistics Cohort 1				**Lottery Statistics Cohort 2**			
	Lottery Winners	**Lottery Losers**	**Difference**	**P-value**	**Lottery Winners**	**Lottery Losers**	**Difference**	**P-value**
N Students	300	190			1,090	590		
Male	53%	52%	1%	0.7670	50%	49%	1%	
Black	89%	90%	-2%	0.5940	86%	87%	-1%	0.5150
FRL-Eligible	100%	100%	0%	1.0000	100%	100%	0%	1.0000
Special Needs	19%	23%	-4%	0.2560	10%	11%	-1%	0.7240
Baseline Math Z-score	0.250	0.349	-0.100	0.2915	-0.084	-0.041	-0.043	0.3829
Grades K-5					-0.003	0.022	-0.025	0.6908
Grades 6-8					-0.311	-0.041	-0.269 **	0.0144
Grades 9-12					-0.139	-0.209	0.071	0.5632
Baseline Reading Z-score	0.146	0.243	-0.098	0.2952	-0.104	0.067	-0.171 ***	0.0007
Grades K-5					-0.065	0.163	-0.228 ***	0.0003
Grades 6-10					-0.217	-0.020	-0.198 *	0.0734
Grades 9-12					-0.116	-0.120	0.004	0.9765
Mean Grade Level (04-05)	7.355	8.513	-1.158 ***	<0.0001	3.91452	4.62818	-0.714 ***	<0.0001

**Special needs indicates special education and/or limited English proficiency.*

In Tables 2 and 3, we report descriptive statistics for the IV samples in math and reading for the first outcome year (Table 2) and for the largest unrestricted samples (Table 3).[2] Within these restricted (IV) samples, as expected, there is a statistically significant difference between DCPS and private school students in terms of the percent of students who won the DC OSP

[2] Descriptive tables for additional samples are available by request.

lottery. In addition, the private school students were less likely to have special needs, and tended to be in lower grades, on average. In the second cohort only, private school students were also more likely to be Black.

Table 2: Restricted (IV) Samples, by Cohort (Outcome Year 1)

Panel A: Math Samples	**Outcome Year 1 Cohort 1 (N=340)**				**Outcome Year 1 Cohort 2 (N=1,370)**			
	Private Students Year 1	**DCPS Students Year 1**	**Difference**	**P-value**	**Private Students Year 1**	**DCPS Students Year 1**	**Difference**	**P-value**
N Students	190	150			780	590		
Won Lottery	90%	40%	50% ***	<0.0001	91%	26%	65% ***	<0.0001
Male	48%	53%	-5%	0.3740	49%	50%	-1%	0.7840
Black	85%	91%	-5%	0.1200	89%	86%	3% *	0.0970
FRL-Eligible	100%	100%	0%	1.0000	100%	100%	0%	1.0000
Special Needs	14%	29%	-15% ***	0.0010	8%	13%	-5% ***	0.0050
Baseline Math Z-score	0.264	0.370	-0.106	0.3282	0.001	-0.053	0.055	0.2790
Baseline Reading Z-score	0.184	0.159	0.025	0.8129	0.026	-0.030	0.056	0.2807
Mean Grade Level (04-05)	7.3	8.1	-0.8 ***	<0.0001	3.7	5.1	-1.3 ***	<0.0001

Panel B: Reading Samples	**Outcome Year 1 Cohort 1 (N=340)**				**Outcome Year 1 Cohort 2 (N=1,300)**			
	Private Students Year 1	**DCPS Students Year 1**	**Difference**	**P-value**	**Private Students Year 1**	**DCPS Students Year 1**	**Difference**	**P-value**
N Students	190	150			740	560		
Won Lottery	90%	40%	50% ***	<0.0001	92%	26%	67% ***	<0.0001
Male	48%	53%	-5%	0.374	49%	50%	-1%	0.816
Black	85%	91%	-5%	0.120	88%	85%	3% *	0.067
FRL-Eligible	100%	100%	0%	1.000	100%	100%	0%	1.000
Special Needs	14%	29%	-15% ***	0.001	8%	13%	-4% **	0.011
Baseline Math Z-score	0.264	0.370	-0.106	0.328	0.013	-0.052	0.065	0.203
Baseline Reading Z-score	0.184	0.159	0.025	0.813	0.027	-0.028	0.055	0.299
Mean Grade Level (04-05)	7.3	8.1	-0.8 ***	<0.0001	3.9	5.4	-1.5 ***	<0.0001

**Special needs indicates special education and/or limited English proficiency.*

Table 3 includes descriptive statistics for our observational Ordinary Least Squares (OLS) analyses, not restricted to the IV sample. For these unrestricted sample analyses, we only include the first cohort of private school students observed in the DC OSP data to hold constant the baseline year (2003-04) and outcome year (2004-05). The math and reading samples were very similar. Private school students were more likely to be FRL-eligible (FRL-eligibility was a requirement for DC OSP eligibility), and were in higher grades than the DCPS students.

Table 3: Unrestricted Samples (Outcome Year 1, Cohort One Only)

	Math (N=17,730)				**Reading (N=17,850)**			
	Private Students 04-05	**DCPS Students 04-05**	**Difference**	**P-value**	**Private Students 04-05**	**DCPS Students 04-05**	**Difference**	**P-value**
N Students	190	17,540			190	17,660		
Won Lottery	90%	0%	90% ***	<0.0001	90%	0%	90% ***	<0.0001
Male	48%	47%	1%	0.939	48%	47%	1%	0.931
Black	85%	87%	-2%	0.803	85%	87%	-2%	0.772
FRL-Eligible	100%	72%	28% ***	<0.0001	100%	72%	28% ***	<0.0001
Special Needs	14%	17%	-3%	0.220	14%	17%	-3%	0.235
Baseline Math Z-score	0.029	0.087	-0.058	0.421	0.029	0.088	-0.060	0.406
Baseline Reading Z-score	0.074	0.058	0.016	0.819	0.074	0.062	0.012	0.864
Mean Grade Level (04-05)	7.271	6.712	0.558 ***	0.002	7.271	6.756	0.515 ***	0.004

**Special needs indicates special education and/or limited English proficiency.*

Random assignment is the best method for assessing the causal impact of a program, particularly when there are concerns about selection bias. The federal private school voucher program is an especially appropriate subject for such a methodological study because self-selection is assumed to influence whether or not a low-income urban student attends a private school. The specific factors assumed to drive self-selection into private schools, such as parental value of education and motivation to overcome financial and logistical challenges, are difficult to measure and control for absent random assignment. Still, some quasi-experimental methods likely will do better or worse at approximating the experimental results.

In order to compare the quasi-experimental methods, we first select a benchmark. We treat Instrumental Variables Analysis (IV) estimates of the impact of private schooling on student outcomes as the "benchmark" estimate of causal impact. A validated random lottery is the ideal instrumental variable with which to recover unbiased estimates of the effect of an intervention like private schooling in the face of substantial non-compliance with the original assignment of students to the treatment of private schooling through the mechanism of a voucher or the control condition (Murray, 2006; Howell & Peterson, 2006).

The lotteries used to create the experimental analysis sample for the DC OSP evaluation produced treatment and control groups with approximately similar baseline conditions (Wolf, Gutmann, Puma & Silverberg, 2006). For cohort 1, about 10% of control group students (20 out of 190) attended private school during the first outcome year. Similarly, for cohort 2, about 9% of control group students attended private school during their first outcome year. With the asset of a validated lottery and the problem of substantial experimental non-compliance, we argue that IV estimates are the most defensible benchmark to use in this case. In addition, the treatment-on-treated (TOT) effect is considered by some to be the parameter of interest in school choice studies and is the estimand typically used in WSCs (Bifulco, 2012). The experimental TOT based on an IV is the local average treatment effect (LATE), the estimated impact for compliers (who attend a private school if offered a voucher but not otherwise) (Cook et al., 2008), while the nonexperimental TOT provides the estimated impact for everyone who attended a private school regardless of voucher application or receipt. We attempt to use the same variables reported for the experimental impacts in Wolf et al. (2013).

Specifically, our IV model is given by the two-stage least squares estimation below:

First Stage:

$$private_{it} = \hat{\pi}_0 + \hat{\pi}_1 lottery_{it} + X_i\theta + \varepsilon_{it} \quad (1)$$

Second Stage:

$$score_{it} = \beta_0 + \beta_1 \widehat{private}_{it} + \beta_2 score_{it-1} + X_i\gamma + \varepsilon_{it} \quad (2)$$

where

$private_{it}$ = 1 if the student attended a private school in the outcome year and 0 otherwise

$lottery_{it}$ = 1 if the student won the DC OSP lottery, and 0 if the student did not win

X_i is a vector of student and family observable characteristics measured in the baseline year including reading and math test scores, grade, age, household income, number of children in the household, the number of months at current residence (as a measure of stability), the number of days between September 1st until the date of testing, and indicator variables for gender, Black, in special education, mother's education, mother's employment status, and ever having attended a school in need of improvement (SINI).

Observations are weighted using the weights from the original evaluation study which take into account the probability of winning a scholarship based on grade-band, nonresponse, and subsampling for nonresponse conversion.

Some sample attrition is inevitable in longitudinal studies. Sample attrition occurs when a student in the study in the baseline year does not provide outcome data in a subsequent year. Sample attrition is not program attrition. Program attrition occurs when a student awarded a scholarship either never uses it or uses it for less than the full amount of time allowed. Some students who were in the program attrition group nevertheless provided subsequent outcome data, and therefore were not part of the sample attrition group. Other students who dropped out of the program also stopped providing test score data and therefore were in both the program and sample attrition groups. Finally, some students who remained in the program and therefore were not part of the program attrition group did not provide outcome data and therefore were part of the sample attrition group. The sample attrition for the first outcome year used in the experimental analysis as the benchmark for this study was 21 percent for the treatment group and 26 percent for the control group (Wolf et al., 2007, pp. F-4). The differential non-response (i.e. sample attrition) rate for the two groups in the experiment was 5 percentage points, which is within the range permitted by the What Works Clearinghouse (n.d.) for an experiment to be considered causal without qualification.

Some control group students crossed over to the treatment condition of private schooling. For control group students who provided Year 1 outcome data in math, 15 percent of them were attending private school without the direct assistance of an Opportunity Scholarship (Wolf et al., 2007, p. 38). While these control group "non-compliers" violate the Cook et al. (2008) criterion for a successfully implemented experiment, some control crossovers are inevitable in a school choice experiment where families are not only allowed but encouraged to choose an alternative school for their child. The IV procedure employed in our analysis factors the size of the control group crossover rate into its unbiased estimation of the impact of the Treatment on the Treated (Howell et al., 2006, pp. 49-52).

We compare the IV-generated benchmarks to the results from three types of alternative research designs for determining the effect of private schooling on student outcomes, in order (theoretically) from most- to least-biased: observational without controls (i.e. comparing simple group averages for private school students to public school students), observational with controls for baseline test scores and demographics, and propensity score matching. We interpret the extent to which the results from the alternative methods deviate from the results from the benchmarks as the degree of self-selection bias from employing that particular quasi-experimental method (e.g. Bifulco 2012).

Observational without controls (mean-comparison)

Simple mean comparisons serve as the comparison of outcomes for private and public school outcomes with the most potential bias. Unfortunately, policy analysts and advocates continue to judge the relative effectiveness of private versus public schools using such simple comparisons of average outcomes with no adjustments for student background. This approach serves primarily as a "negative baseline" to establish the upper-bound of the range of bias possible from non-experimental methods.

Observational with controls

Two main assumptions are required for a regression-based analysis to produce unbiased estimates. First, regression assumes the absence of confounds, that all factors confounding the relationship between treatment group status (in our case, attending a private school or not) and test scores are observable, measurable, and included in the model (Rosenbaum & Rubin, 1983; Little & Rubin, 2000). This assumption is untestable in practice, however, the use of pretreatment outcome measures as controls makes this assumption more reasonable.

The second assumption of the regression approach is that the relationships between confounding factors and outcome measures (test scores) are specified correctly, accounting for possible nonlinearity or interactions among two or more variables (see Fortson, 2012, pp. 19-21).

The OLS models differ by type of analysis and sample (restricted or unrestricted), and some models use only a subset of these explanatory variables. The model for the restricted sample (equation 3) contains a robust set of controls collected during the DC OSP evaluation.

$$score_{it} = \beta_0 + \beta_1 private_{it} + \beta_2 score_{it-1} + X_i\gamma + \varepsilon_{it} \tag{3}$$

where

$private_{it}$ = 1 if the student attended a private school in the outcome year and 0 otherwise

X_i is a vector of student and family observable characteristics measured in the baseline year including reading and math test scores, grade, age, household income, number of children in the household, the number of months at current residence (as a measure of stability), the number of days between September 1st until the date of testing, and indicator variables for gender, Black, in special education, mother's education, mother's employment status, and ever having attended a SINI school.

The model for the unrestricted sample (which includes DCPS data and therefore is more limited in covariate choices is:

$$score_{it} = \beta_0 + \beta_1 private_{it} + \beta_2 score_{it-1} + \beta_3 FRL_{it-1} + \beta_4 special\_needs_{it-1} + \\ \beta_5 black_{it-2} + \beta_6 grade_{it} + \beta_7 grade_{it-1} + \varepsilon_{it} \tag{4}$$

These models are estimated with heteroskedastic-robust standard errors. If the twin assumptions of included confounding variables and correct specification are satisfied, the estimation of equations (3) and (4) could yield unbiased results that approximate our experimental benchmark. Most evaluators view that as a big "if".

Matching strategy

Under the assumption that potential outcomes are independent of treatment, conditional on a set of covariates *X*, it also can be assumed that potential outcomes are independent of treatment, conditional on propensity score, $\pi(X)$. A propensity score is defined as the likelihood of being in the treatment group given a subject's measured baseline characteristics (Rosenbaum & Rubin, 1984). In addition, for propensity score matching to produce consistent estimates, the distribution of covariates must be the same for treatment and comparison groups, conditional on estimated propensity score. This requirement is known as common support.

We conduct matching by first requiring exact matches in terms of outcome year grade, prior year grade, standardized test score (z-score) rounded to .01 standard deviations, free-and reduced-price lunch status, and special needs status, and then a nearest neighbor match based on the propensity to be in a private school in the outcome year. This propensity score is based on all other covariates that are available and used in the OLS models, equations (3) and (4). Then, among the matched sample, we conduct OLS regression as in equations (3) and (4), depending on whether we are comparing within the restricted or unrestricted samples. Multivariate regression is particularly important if there is not baseline equivalence on all characteristics, which is possible when using nearest neighbor matching as opposed to exact matching. The propensity score approach is less restrictive than exact matching on all characteristics, allowing for a larger number of matches.

Bifulco (2012) finds that comparison groups constructed using propensity scores including baseline test scores and a measure of geography match non-treatment students to treatment students sufficiently well so that program effect estimates from matching methods differ from experimental benchmarks by less than 10 percent. Therefore, we also conduct some matching analyses assuming that baseline test scores are unknown and assess whether matching performs better when baseline test scores are included. This element of our WSC is an attempt to replicate Bifulco's results for Connecticut public charter schools in the case of private schooling in Washington, DC. Finally, we compare the results of these four methods across two achievement outcomes: math achievement and reading achievement, permitting us to test the robustness of our bias calculations to various outcomes.

Non-response weights for each of the non-experimental methods are calculated as the inverse probability of response (having a test score in the outcome year). These weights are calculated for math and reading separately. We define $\pi_{imath}$and $\pi_{ireading}$ as the predicted probability of having an observed math or reading test score in the outcome year, based on the following probit model:

$$has\_test\_score_{it} = \delta_0 + \delta_1 private_{it} + \delta_2 math\_score_{it-1} + \delta_3 reading\_score_{it-1} \\ + \delta_4 FRL_{it-1} + \delta_5 special\_needs_{it-1} + \delta_6 black_{it-1} + \delta_7 grade_{it} + \delta_8 grade_{it-1} + \varepsilon_{it} \quad (5)$$

Results

Restricted (Experimental IV) Sample

Results for our four outcome years are in Tables 4-11, in reading (Tables 4-7) and math (Tables 8-11). Our ability to make meaningful comparisons to the experimental "benchmark" in these cases is limited by a noisy IV result in all four math outcome years and in the first reading outcome year. Nonetheless, we can compare our other methods to these noisy zeros, to assess whether non-experimental methods would lead us to substantively different conclusions.

Some of these LATE estimates of the impact of the DC voucher program on student achievement are being presented for the first time. During the original valuation, IV estimates of LATE were presented only when intent-to-treat calculations indicated that the experimental impact of being offered an Opportunity Scholarship was statistically significant. Thus, none of the IV math estimates and only the Year 3 IV reading estimates were included in the prior evaluation reports.

Table 4: Restricted (IV) Sample First Year Reading Results (Both Cohorts)

	Benchmark IV	**Matching w/ Baseline Scores**		**Matching w/ Demographics Only**		**OLS with Controls**			**Full IV Sample**
	IV Regression Results	**Regression**	**Mean Comparison Only**	**Regression**	**Mean Comparison Only**	**Regression**			**Mean Comparison Only**
Private Schooling	-0.0333	0.0224	0.0363	0.0570	0.1077*	-0.0205	0.0141	-0.00473	0.0594
	(0.0833)	(0.0513)	(0.0659)	(0.0545)	(0.0582)	(0.0568)	(0.0593)	(0.0598)	(0.0496)
PY Reading Z-score	0.386***	0.528***				0.386***	0.402***		
	(0.0602)	(0.0474)				(0.0611)	(0.0624)		
PY Math Z-score	0.0705*	0.138***				0.0706*	0.120***		
	(0.0406)	(0.0406)				(0.0410)	(0.0441)		
Household Income (000s)	-0.0013	-0.0005		0.0008		-0.0013		-0.0021	
	(0.0030)	(0.0029)		(0.0030)		(0.0030)		(0.0033)	
Male	-0.0495	0.0213		-0.0890		-0.0496		-0.119**	
	(0.0537)	(0.0523)		(0.0545)		(0.0543)		(0.0563)	
Special Needs	-0.644***	-0.492***		-0.840***		-0.643***		-0.811***	
	(0.0846)	(0.0855)		(0.0766)		(0.0849)		(0.0888)	
Black	-0.0519	-0.124		-0.187**		-0.0514		-0.144*	
	(0.0737)	(0.0805)		(0.0868)		(0.0747)		(0.0864)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	0.415	0.752**		1.235***		0.405	0.135	0.846**	
	(0.339)	(0.362)		(0.352)		(0.340)	(0.138)	(0.387)	
Observations	1,650	880	880	1,180	1,180	1,650	1650	1650	1,650
R-squared	0.290	0.412		0.153		0.290	0.233	0.126	
Adjusted R-squared	0.276	0.395		0.135		0.277	0.224	0.111	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to the nearest 10.

Table 5: Restricted (IV) Sample Second Year Reading Results (Both Cohorts)

	Benchmark IV	Matching w/ Baseline Scores		Matching w/ Demographics Only		OLS with Controls			Full IV Sample
	IV Regression Results	Regression	Mean Comparison Only	Regression	Mean Comparison Only		Regression		Mean Comparison Only
Private Schooling	0.183**	0.0693	0.0806	0.108*	0.1701***	0.141***	0.167***	0.166***	0.1466***
	(0.0859)	(0.0585)	(0.0694)	(0.0599)	(0.0618)	(0.0516)	(0.0550)	(0.0594)	(0.0527)
PY Reading Z-score	0.0776*	0.472***				0.417***	0.428***		
	(0.0456)	(0.0510)				(0.0413)	(0.0410)		
PY Math Z-score	0.416***	0.118***				0.0772*	0.105**		
	(0.0408)	(0.0446)				(0.0464)	(0.0474)		
Household Income	-0.0009	-0.0021		-0.0025		-0.0007		0.0001	
	(0.0029)	(0.0036)		(0.0037)		(0.0029)		(0.0036)	
Male	-0.135***	-0.0729		-0.180***		-0.135***		-0.176***	
	(0.0501)	(0.0602)		(0.0609)		(0.0507)		(0.0582)	
Special Needs	-0.440***	-0.154		-0.630***		-0.447***		-0.629***	
	(0.113)	(0.104)		(0.0978)		(0.114)		(0.106)	
Black	-0.147*	-0.0936		-0.222**		-0.147*		-0.196*	
	(0.0838)	(0.101)		(0.101)		(0.0846)		(0.0998)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	-0.0126	0.407		1.238***		0.0280	-0.136*	0.449	
	(0.395)	(0.376)		(0.353)		(0.404)	(0.0734)	(0.462)	
Observations	1,460	710	710	1,030	1,030	1460	1460	1460	1,460
R-squared	0.295	0.311		0.104		0.296	0.260	0.097	
Adjusted R-squared	0.281	0.286		0.0834		0.281	0.251	0.0789	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to the nearest 10.

Table 6: Restricted (IV) Sample Third Year Reading Results (Both Cohorts)

	Benchmark IV	Matching w/ Baseline Scores		Matching w/ Demographics Only		OLS with Controls			Full IV Sample
	IV Regression Results	Regression	Mean Comparison Only	Regression	Mean Comparison Only	Regression			Mean Comparison Only
Private Schooling	0.189*	0.127**	0.1735**	0.297***	0.3409***	0.209***	0.259***	0.269***	0.3000***
	(0.0966)	(0.0615)	(0.0703)	(0.0571)	(0.0584)	(0.0564)	(0.0575)	(0.0609)	(0.0536)
PY Reading Z-score	0.0816*	0.426***				0.412***	0.416***		
	(0.0493)	(0.0529)				(0.0448)	(0.0443)		
PY Math Z-score	0.412***	0.161***				0.0814	0.120**		
	(0.0446)	(0.0466)				(0.0498)	(0.0511)		
Household Income	0.0033	0.0099***		0.0030		0.0032		0.0030	
	(0.0029)	(0.0036)		(0.0032)		(0.0029)		(0.0034)	
Male	-0.0707	-0.0306		-0.0932		-0.0707		-0.127**	
	(0.0485)	(0.0610)		(0.0572)		(0.0490)		(0.0566)	
Special Needs	-0.362***	-0.178*		-0.541***		-0.361***		-0.558***	
	(0.0888)	(0.104)		(0.0867)		(0.0899)		(0.0866)	
Black	-0.119	-0.138		-0.198**		-0.119		-0.155	
	(0.0854)	(0.0953)		(0.0857)		(0.0863)		(0.0998)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	0.442	0.495		0.717**		0.413	-0.170**	0.777	
	(0.452)	(0.358)		(0.361)		(0.462)	(0.0792)	(0.554)	
Observations	1,370	740	740	1,140	1,140	1,370	1370	1370	1,370
R-squared	0.306	0.299		0.108		0.306	0.277	0.109	
Adjusted R-squared	0.292	0.276		0.091		0.292	0.268	0.092	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to the nearest 10.

Table 7: Restricted (IV) Sample Fourth Year Reading Results (Both Cohorts)

	Benchmark IV	**Matching w/ Baseline Scores**		**Matching w/ Demographics Only**		**OLS with Controls**			**Full IV Sample**
	IV Regression Results	**Regression**	**Mean Comparison Only**	**Regression**	**Mean Comparison Only**	**Regression**			**Mean Comparison Only**
Private Schooling	0.237*	0.210***	0.2167***	0.278***	0.3071***	0.163***	0.151**	0.220***	0.3138***
	(0.137)	(0.0604)	(0.0707)	(0.0578)	(0.0610)	(0.0567)	(0.0628)	(0.0633)	(0.0554)
PY Reading Z-score	0.102**	0.481***				0.328***	0.334***		
	(0.0456)	(0.0517)				(0.0487)	(0.0518)		
PY Math Z-score	0.324***	0.0248				0.101**	0.168***		
	(0.0497)	(0.0466)				(0.0450)	(0.0477)		
Household Income	0.0026	0.0060*		0.0077**		0.0026		0.0032	
	(0.0034)	(0.0035)		(0.0034)		(0.0035)		(0.0039)	
Male	-0.0566	-0.152**		-0.135**		-0.0552		-0.116*	
	(0.0565)	(0.0630)		(0.0588)		(0.0564)		(0.0595)	
Special Needs	-0.568***	-0.364***		-0.714***		-0.566***		-0.767***	
	(0.0975)	(0.102)		(0.0927)		(0.0968)		(0.104)	
Black	-0.119	-0.144		-0.166*		-0.125*		-0.177*	
	(0.0729)	(0.0952)		(0.0891)		(0.0745)		(0.0906)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	0.745**	1.098***		1.268***		0.808**	-0.0554	1.070***	
	(0.340)	(0.340)		(0.345)		(0.322)	(0.0733)	(0.352)	
Observations	1,330	710	710	1,040	1,040	1330	1330	1330	1,330
R-squared	0.287	0.302		0.152		0.289	0.218	0.145	
Adjusted R-squared	0.273	0.278		0.134		0.274	0.209	0.129	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to the nearest 10.

Table 8: Restricted (IV) Sample First Year Math Results (Both Cohorts)

	Benchmark IV	Matching IV Sample w/ Baseline Scores		Matching IV Sample w/ Demographics Only		OLS with Controls			Full IV Sample
	IV Regression Results	Regression	Mean Comparison Only	Regression	Mean Comparison Only	Regression			Mean Comparison Only
Private Schooling	-0.00752	-0.0171	0.0112	0.0313	0.0628	-0.00434	0.0122	0.0009	0.0626
	(0.0794)	(0.0538)	(0.0615)	(0.0556)	(0.0570)	(0.0547)	(0.0555)	(0.0610)	(0.0484)
PY Math Z-score	0.397***	0.444***				0.397***	0.422***		
	(0.0488)	(0.0477)				(0.0492)	(0.0514)		
PY Reading Z-score	0.1000**	0.0732*				0.0999**	0.104**		
	(0.0482)	(0.0422)				(0.0487)	(0.0481)		
Household Income (000s)	-0.0019	-0.0045		0.0003		-0.0019		-0.0017	
	(0.0029)	(0.0030)		(0.0031)		(0.0029)		(0.0034)	
Male	0.0525	0.0437		0.0328		0.0524		0.00418	
	(0.0526)	(0.0545)		(0.0562)		(0.0531)		(0.0578)	
Special Needs	-0.376***	-0.457***		-0.675***		-0.376***		-0.604***	
	(0.0722)	(0.0848)		(0.0805)		(0.0723)		(0.0808)	
Black	-0.145*	-0.183*		-0.278***		-0.145*		-0.257***	
	(0.0752)	(0.0939)		(0.0923)		(0.0760)		(0.0913)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	-0.122	0.214		0.510		-0.124	0.0680	0.345	
	(0.398)	(0.277)		(0.338)		(0.400)	(0.101)	(0.403)	
Observations	1,720	910	910	1,250	1,250	1720	1720	1720	1,720
R-squared	0.260	0.263		0.086		0.260	0.239	0.069	
Adjusted R-squared	0.247	0.243		0.067		0.247	0.230	0.054	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to nearest 10.

Table 9: Restricted (IV) Sample Second Year Math Results (Both Cohorts)

	Benchmark IV	**Matching IV Sample w/ Baseline Scores**		**Matching IV Sample w/ Demographics Only**		**OLS with Controls**			**Full IV Sample**
	IV Regression Results	**Regression**	**Mean Comparison Only**	**Regression**	**Mean Comparison Only**	**Regression**			**Mean Comparison Only**
Private Schooling	0.0453	0.00951	0.0411	0.0764	0.1257**	0.0695	0.0740	0.0879	0.0864
	(0.0847)	(0.0638)	(0.0714)	(0.0605)	(0.0613)	(0.0555)	(0.0562)	(0.0638)	(0.0527)
PY Math Z-score	0.381***	0.416***				0.381***	0.413***		
	(0.0547)	(0.0546)				(0.0550)	(0.0576)		
PY Reading Z-score	0.124***	0.104**				0.124***	0.124***		
	(0.0438)	(0.0490)				(0.0435)	(0.0418)		
Household Income (000s)	0.00296	0.00142		0.00157		0.00291		0.00343	
	(0.0031)	(0.0038)		(0.0036)		(0.0032)		(0.0039)	
Male	0.0571	0.0649		0.0221		0.0573		0.0492	
	(0.0511)	(0.0664)		(0.0613)		(0.0517)		(0.0604)	
Special Needs	-0.282***	-0.230**		-0.455***		-0.278***		-0.507***	
	(0.0989)	(0.103)		(0.0885)		(0.100)		(0.0872)	
Black	-0.182**	-0.164		-0.283***		-0.181**		-0.259**	
	(0.0876)	(0.109)		(0.0990)		(0.0888)		(0.110)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	-0.678	0.664*		1.356***		-0.702	-0.0483	-0.282	
	(0.531)	(0.359)		(0.353)		(0.552)	(0.0781)	(0.538)	
Observations	1,460	730	730	1,030	1,030	1460	1460	1460	1,460
R-squared	0.282	0.239		0.086		0.282	0.256	0.082	
Adjusted R-squared	0.267	0.212		0.0653		0.267	0.246	0.0638	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to nearest 10.

Table 10: Restricted (IV) Sample Third Year Math Results (Both Cohorts)

	Benchmark IV	**Matching IV Sample w/ Baseline Scores**		**Matching IV Sample w/ Demographics Only**		**OLS with Controls**			**Full IV Sample**
	IV Regression Results	**Regression**	**Mean Comparison Only**	**Regression**	**Mean Comparison Only**	**Regression**			**Mean Comparison Only**
Private Schooling	-0.00847	0.0424	0.0602	0.194***	0.2226***	0.108*	0.121**	0.163**	0.1845***
	(0.103)	(0.0623)	(0.0695)	(0.0571)	(0.0580)	(0.0575)	(0.0579)	(0.0641)	(0.0535)
PY Math Z-score	0.266***	0.343***				0.264***	0.279***		
	(0.0494)	(0.0502)				(0.0494)	(0.0528)		
PY Reading Z-score	0.240***	0.190***				0.237***	0.232***		
	(0.0503)	(0.0472)				(0.0509)	(0.0505)		
Household Income (000s)	0.0006	-0.0012		0.0019		0.0003		0.0008	
	(0.0031)	(0.0031)		(0.0030)		(0.0031)		(0.0035)	
Male	0.0452	0.0828		0.0172		0.0458		0.00523	
	(0.0558)	(0.0627)		(0.0577)		(0.0562)		(0.0627)	
Special Needs	-0.171**	-0.345***		-0.455***		-0.165**		-0.396***	
	(0.0833)	(0.0926)		(0.0725)		(0.0833)		(0.0826)	
Black	-0.204**	-0.306***		-0.260***		-0.203**		-0.253**	
	(0.0855)	(0.102)		(0.0920)		(0.0864)		(0.104)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	0.161	0.822**		0.723**		-0.00529	-0.0203	0.428	
	(0.504)	(0.323)		(0.350)		(0.509)	(0.0861)	(0.557)	
Observations	1,370	760	760	1,150	1,150	1370	1370	1370	1,370
R-squared	0.235	0.255		0.072		0.238	0.223	0.056	
Adjusted R-squared	0.219	0.230		0.0534		0.223	0.214	0.0385	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to nearest 10.

Table 11: Restricted (IV) Sample Fourth Year Math Results (Both Cohorts)

	Benchmark IV	Matching IV Sample w/ Baseline Scores		Matching IV Sample w/ Demographics Only		OLS with Controls			Full IV Sample
	IV Regression Results	Regression	Mean Comparison Only	Regression	Mean Comparison Only	Regression			Mean Comparison Only
Private Schooling	0.0063	0.0026	0.0312	0.141**	0.1592**	0.0770	0.0666	0.118*	0.1704***
	(0.135)	(0.0672)	(0.0726)	(0.0607)	(0.0618)	(0.0575)	(0.0582)	(0.0624)	(0.0558)
PY Math Z-score	0.284***	0.343***				0.285***	0.327***		
	(0.0456)	(0.0549)				(0.0462)	(0.0465)		
PY Reading Z-score	0.123***	0.147***				0.119***	0.118***		
	(0.0409)	(0.0477)				(0.0412)	(0.0420)		
Household Income (000s)	0.0007	0.0033		0.0063*		0.0007		0.0019	
	(0.0036)	(0.0037)		(0.0035)		(0.0037)		(0.0039)	
Male	0.0214	-0.0509		-0.0663		0.0199		-0.0138	
	(0.0575)	(0.0671)		(0.0608)		(0.0582)		(0.0615)	
Special Needs	-0.322***	-0.215**		-0.541***		-0.324***		-0.562***	
	(0.0830)	(0.0972)		(0.0892)		(0.0835)		(0.0859)	
Black	-0.280***	-0.322***		-0.283***		-0.274***		-0.336***	
	(0.0888)	(0.107)		(0.0924)		(0.0890)		(0.100)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls	Y	Y		Y		Y		Y	
Constant	0.337	1.181***		0.948**		0.280	0.00469	0.526	
	(0.399)	(0.439)		(0.375)		(0.393)	(0.0748)	(0.408)	
Observations	1,330	740	740	1,040	1040	1,330	1,330	1,330	1,330
R-squared	0.198	0.204		0.076		0.199	0.170	0.076	
Adjusted R-squared	0.181	0.178		0.0561		0.182	0.161	0.0579	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to nearest 10.

We also display 90% confidence intervals around these various estimates in Figures 1-4 (reading estimates) and Figures 5-8 (math estimates). In each figure, the far left point estimate with confidence interval is the experimental benchmark, and each additional point estimate with confidence interval as we move from left to right is hypothesized to have more selection bias. Matching models are hypothesized to have the least bias, followed by OLS with controls, and finally simple mean comparisons. In addition, models including baseline test scores are hypothesized to be less biased than models with demographics only. In these figures, we exclude the simple mean comparisons of our matching analyses, favoring the regression results within the matched samples, because our matching samples are not statistically equivalent on all baseline characteristics. See Appendix A for Baseline Equivalency Tables for the first year outcomes.[3]

In general, the reading estimates across these various model types are quite similar, holding sample constant. For example, Figure 1 indicates that all models estimate null effects of the program on reading in Year 1, so we have no false positives. Similarly, in Figures 3 and 4, all estimates are that the program had a positive and statistically significant effect on reading outcomes in both Year 3 and Year 4, indicating no false negatives. In Figure 2, however, we have evidence of one false negative, and surprisingly, this is the matching model that was hypothesized to be the least biased. This indicates that holding constant certain factors about the comparison group is important for removing bias (Aiken et al., 1998; Bifulco, 2012; Heckman, Ichimura, and Todd, 1997; Heckman et al., 1998; Shadish, Clark, & Steiner, 2008). However, after holding sample constant, there is not consistent evidence that there is an additional reduction in bias when including pretreatment outcome measures as covariates.

[3] Baseline equivalency tables for additional samples (year 2-4 outcomes for the restricted samples) are available upon request.

Figure 1:

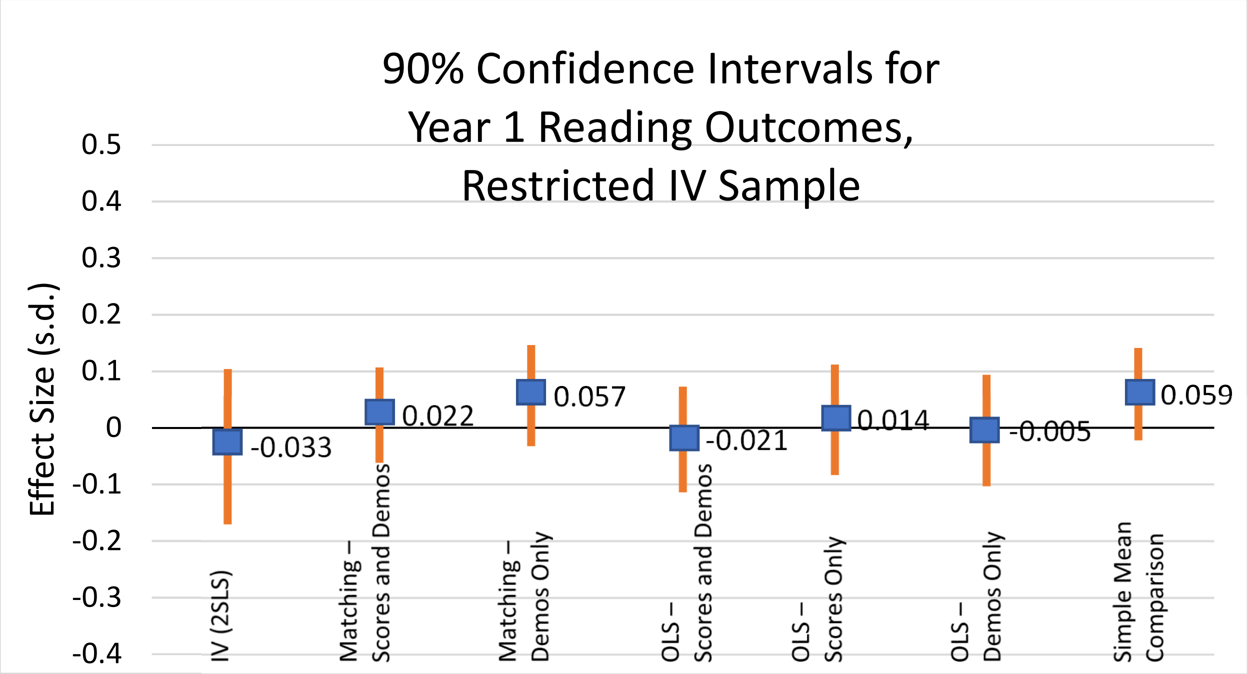


Figure 2:

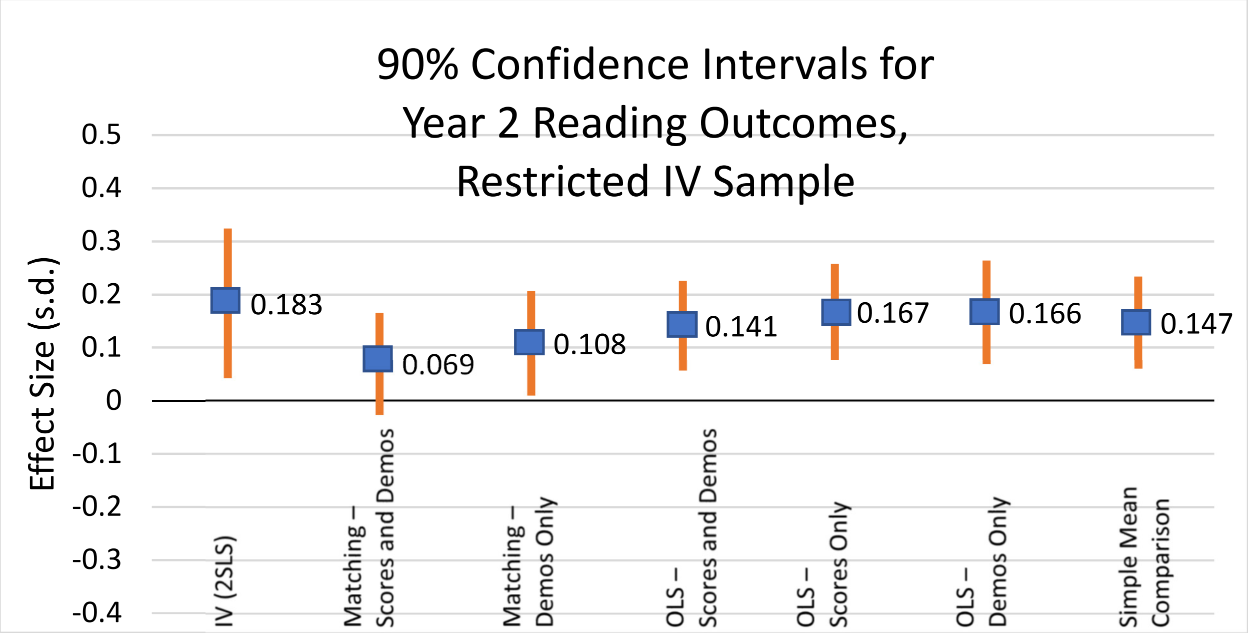


Figure 3:

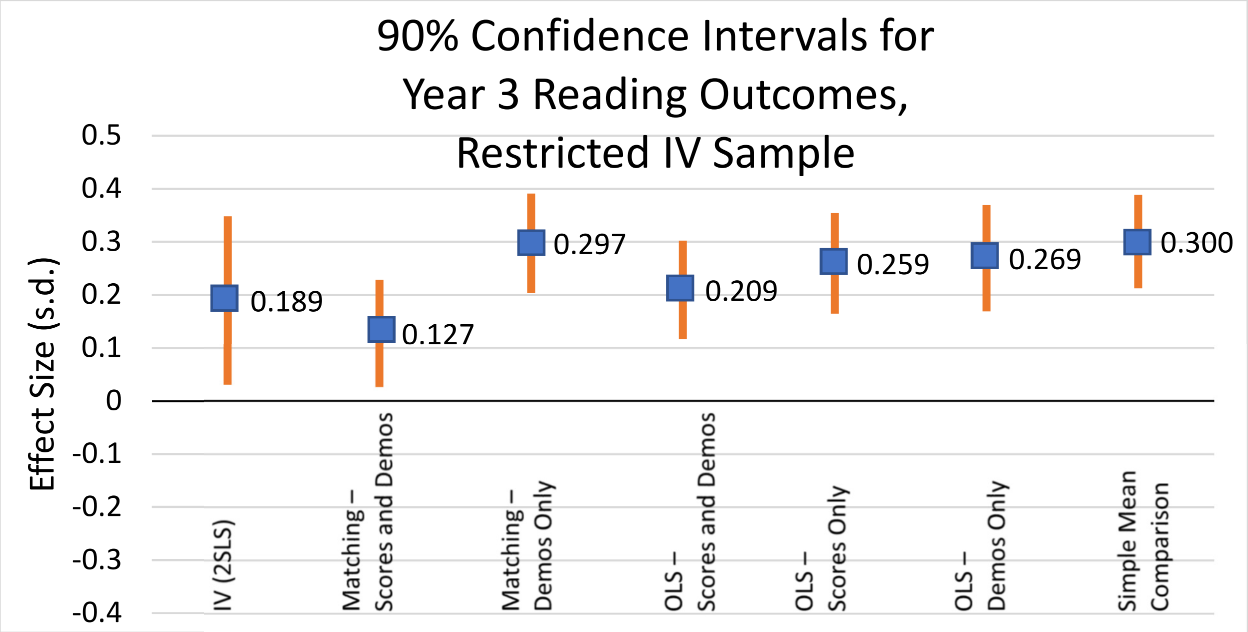


Figure 4:

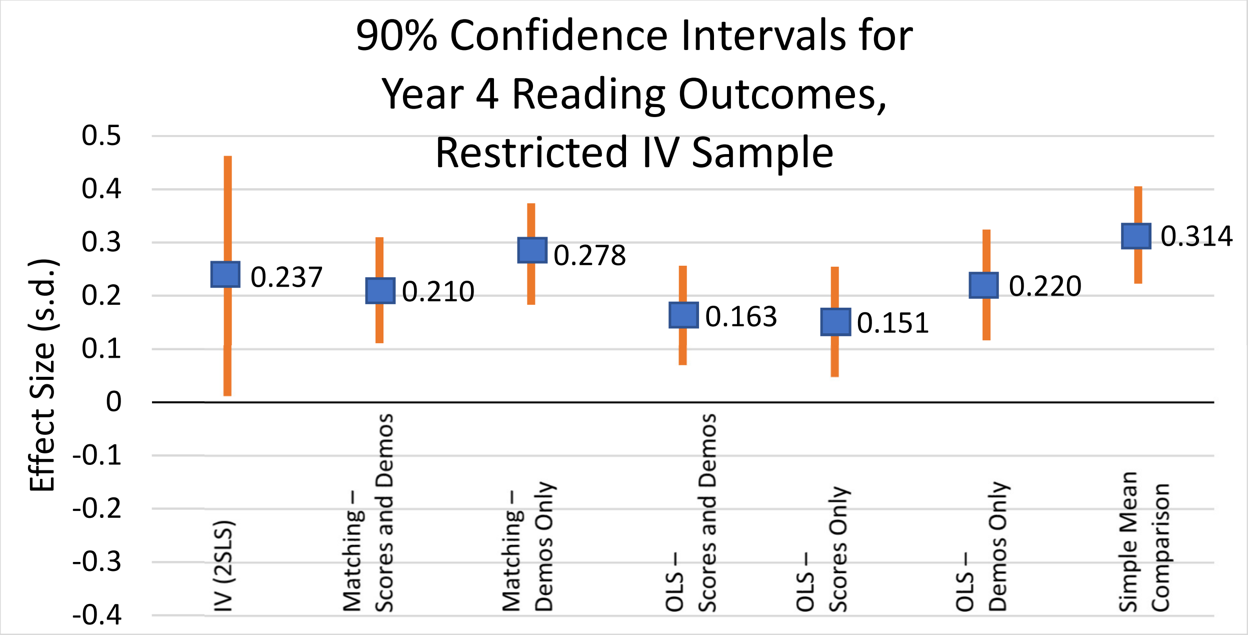


Figure 5:

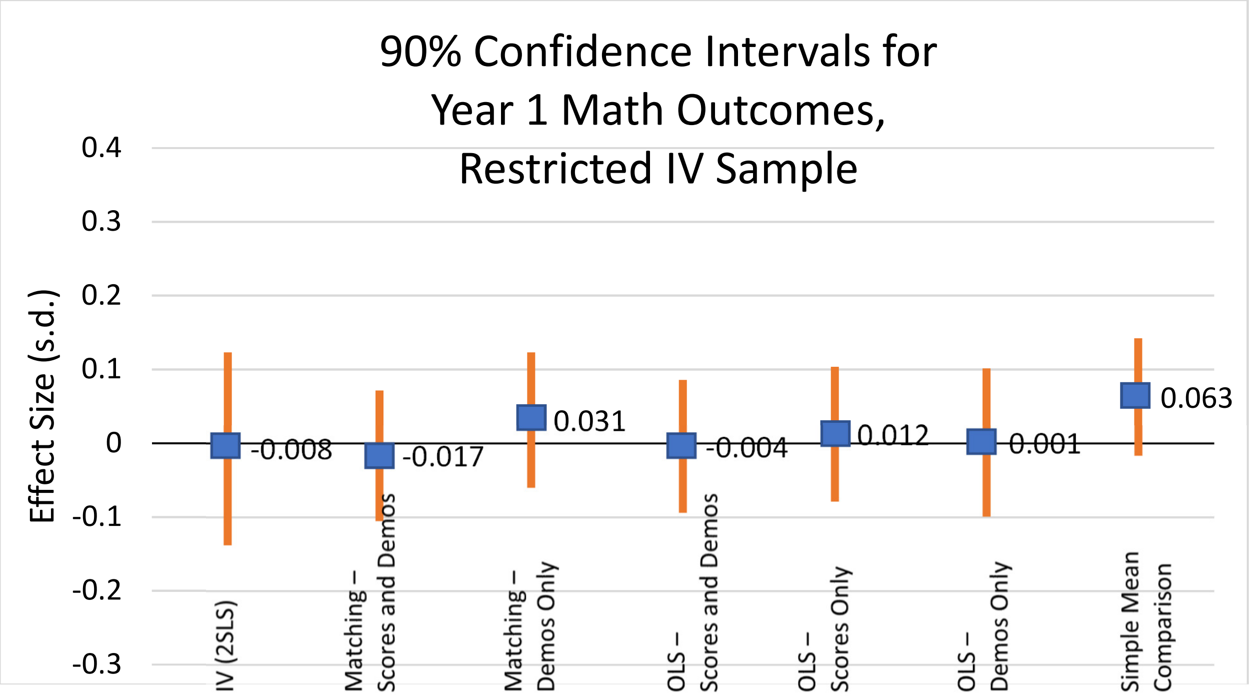


Figure 6:

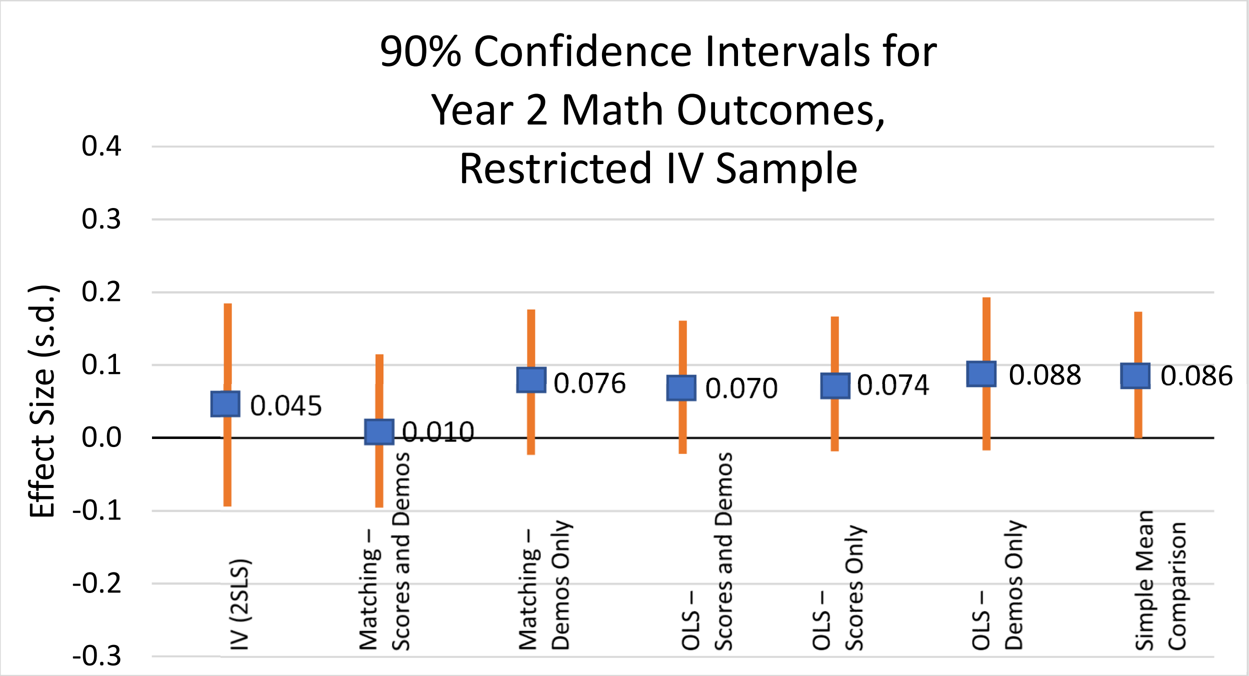


Figure 7:

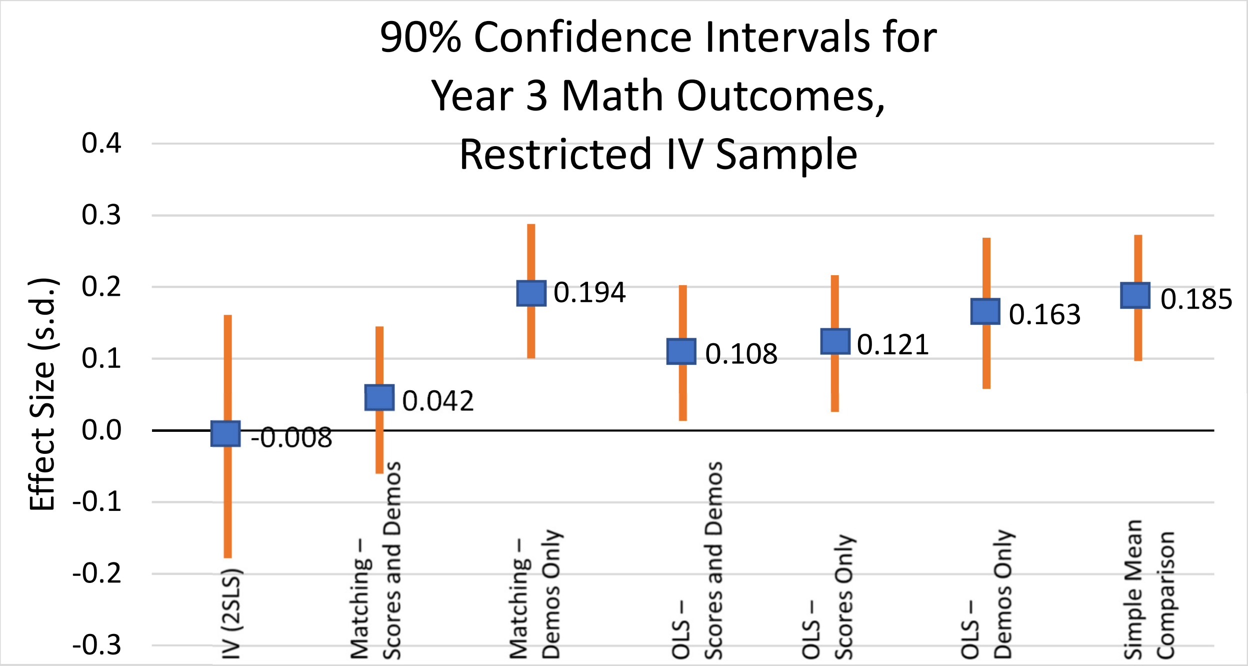
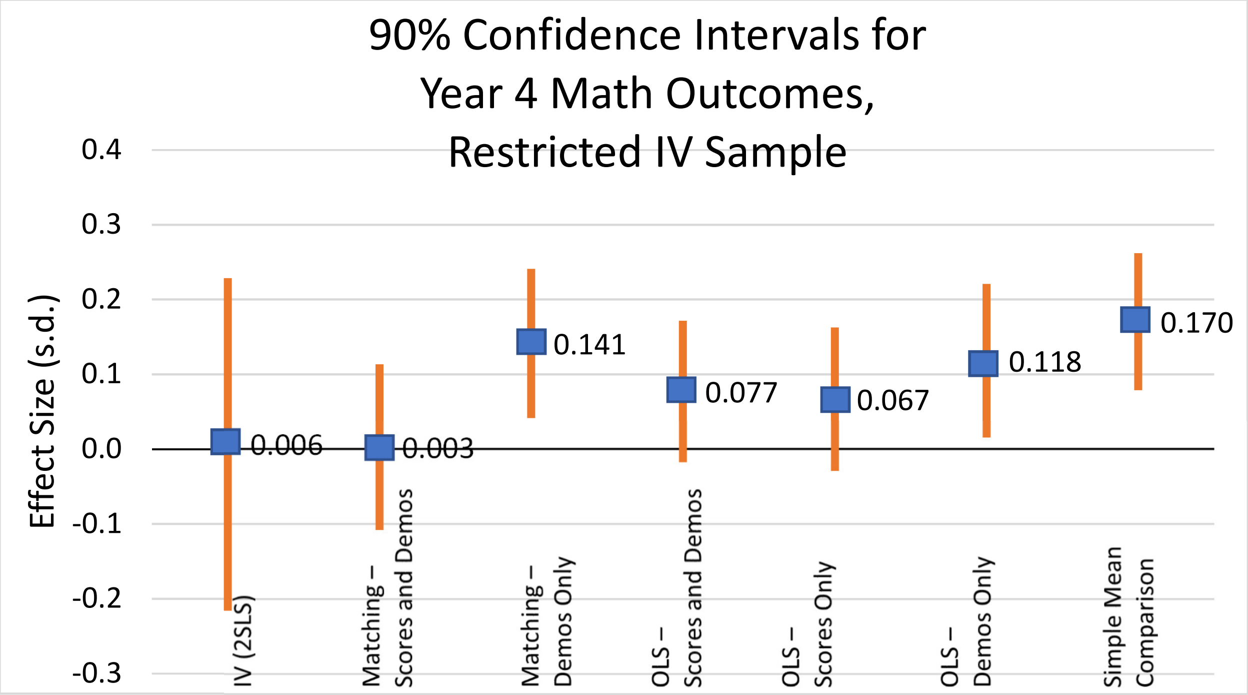


Figure 8:



The math experimental benchmarks in all four outcome years were null, so it is more difficult to draw many conclusions from these results, but again we see that in the first two outcome years, all estimates, regardless of model, were null, indicating at least a lack of false positives from using potentially biased quasi-experimental analytic methods. In Figures 3 and 4, we see that the matching model with full controls (hypothesized to be least biased), was at least closest to the experimental benchmark, whereas the matching method without baseline test scores, all the approaches that relied on control variables, and the simple mean comparisons consistently yielded a false positive estimate that the program increased student achievement in math in Year 3. In Year 4, all non-experimental methods that accounted for baseline test scores produced the correct substantive result of null effects in math; whereas, all three non-experimental estimates that ignored baseline test scores yielded false positive results.

Unrestricted Sample Results (for Additional External Validity)

The next set of results we present come from our unrestricted sample analyses that compare the first cohort of students that applied to the DC OSP and were enrolled in private schools in the first outcome year, compared to students who attended DCPS in the first outcome year (regardless of whether they applied to DC OSP or not). We compare these unrestricted analyses to experimental benchmarks using IV regression, but removing the second cohort of DC OSP students. We are limited to one year of analysis for the unrestricted sample because DC changed its accountability exam from the SAT-9 to a criterion-referenced test during the second year of the OSP. The students in the OSP, and the members of the randomized control group, continued to take the SAT-9 for the entire four outcome years of the program evaluation, but the test change precludes us from using DC non-applicants for our WSC after Year 1.

Table 12 and 13 present the reading and math results, respectively. Figures 9 and 10 illustrate the corresponding 90% confidence intervals. The most notable result from Tables 12-13 and Figures 9-10, is that when seeking external validity by increasing the pool of potential comparison students to the DCPS students, we have negatively biased impact estimates in all instances. For the math result in particular, every non-experimental estimate in Figure 10 would lead to a conclusion that private schooling has a negative impact, relative to an experimental estimate that was null. For the reading result in Figure 10, which is noisy by positive in the experimental benchmark, we would have made an incorrect conclusion (either null or negative) in every model using non-experimental methods.[4]

[4] The careful reader may note the large difference between the year one IV regression reading impact estimates in Table 4 (null) and Table 12 (0.384 s.d., significant at the 90% confidence level). These differences are driven primarily by differences in the cohorts included, not by covariates chosen. See Appendix B for a table of IV regression estimates with differing cohorts included (both or cohort 1 only) and differing covariates included in the model (full RCT controls or restricted controls that would have been available in the DCPS data).

Table 12 Comparison of Methods in Unrestricted Sample (Outcome Year One Reading)

	Benchmark (Restricted Sample)	**Matching Unrestricted Sample w/ Baseline Scores**		**Matching Unrestricted Sample w/ Demographics Only**		**OLS w/ Unrestricted Sample**			**Full Unrestricted Sample**
	IV Regression Results	**Regression**	**Mean Comparison Only**	**Regression**	**Mean Comparison Only**	**Regression**			**Mean Comparison Only**
Private Schooling	0.384*	0.0185	0.0267	-0.0529	0.0335	-0.0624	-0.0700*	-0.0677	-0.1104
	(0.219)	(0.0721)	(0.0790)	(0.0589)	(0.08090	(0.0426)	(0.0402)	(0.0493)	(0.0727)
PY Math Z-score	0.287***	0.139***				0.178***	0.191***		
	(0.0780)	(0.0406)				(0.0152)	(0.0145)		
PY Reading Z-score	0.259***	0.366***				0.333***	0.341***		
	(0.0999)	(0.0490)				(0.0214)	(0.0250)		
FRL Status	N/A	N/A		0.573		-0.0561		-0.183***	
				(0.900)		(0.0557)		(0.0563)	
Male	-0.00161	0.0414		-0.0827		-0.0983***		-0.123***	
	(0.104)	(0.0700)		(0.0752)		(0.0298)		(0.0302)	
Special Needs	-0.249**	-0.162**		-0.377***		-0.0388		-0.466***	
	(0.114)	(0.0774)		(0.0716)		(0.0486)		(0.0498)	
Black	-0.0176	-0.144**		-0.185		-0.137**		-0.327***	
	(0.142)	(0.0688)		(0.115)		(0.0610)		(0.0605)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls									
Constant	-0.174	-0.242		-0.0544		-1.947***	-2.355***	0.483***	
	(0.361)	(0.205)		(0.920)		(0.0955)	(0.0491)	(0.0630)	
Observations	350	370	370	380	380	17,850	17850	17850	17,850
R-squared	0.322	0.200		0.069		0.189	0.185	0.060	
Adjusted R-squared	0.285	0.168		0.0305		0.188	0.184	0.0582	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

All sample sizes rounded to nearest 10.

Table 13: Comparison of Methods in Unrestricted Sample (Outcome Year One Math)

	Benchmark (Restricted Sample)	**Matching Unrestricted Sample w/ Baseline Scores**		**Matching Unrestricted Sample w/ Demographics Only**		**OLS w/ Unrestricted Sample**			**Full Unrestricted Sample**
	IV Regression Results	**Regression**	**Mean Comparison Only**	**Regression**	**Mean Comparison Only**	**Regression**			**Mean Comparison Only**
Private Schooling	0.105	-0.245***	-0.2316***	-0.159***	-0.1615**	-0.251***	-0.248***	-0.278***	-.3015***
	(0.211)	(0.0486)	(0.0624)	(0.0606)	(0.0644)	(0.0477)	(0.0452)	(0.0467)	(0.0723)
PY Math Z-score	0.665***	0.414***				0.360***	0.368***		
	(0.0819)	(0.0477)				(0.0212)	(0.0199)		
PY Reading Z-score	-0.149**	0.0555				0.0829***	0.0883***		
	(0.0728)	(0.0413)				(0.0216)	(0.0266)		
FRL Status	N/A	N/A		-0.554*		-0.00701		-0.112*	
				(0.299)		(0.0557)		(0.0576)	
Male	0.180*	0.00973		0.0327		-0.0610**		-0.0776**	
	(0.106)	(0.0496)		(0.0609)		(0.0304)		(0.0308)	
Special Needs	-0.101	-0.111*		-0.357***		-0.0385		-0.397***	
	(0.102)	(0.0609)		(0.0692)		(0.0457)		(0.0485)	
Black	0.0346	-0.0267		-0.0123		-0.0832		-0.292***	
	(0.145)	(0.0627)		(0.0880)		(0.0606)		(0.0605)	
Baseline Grade Indicators	Y	Y		Y		Y	Y	Y	
Outcome Grade Indicators	Y	Y		Y		Y	Y	Y	
Additional RCT Controls									
Constant	-0.647*	0.217**		0.703		-0.649***	-0.889***	1.856***	
	(0.376)	(0.0841)		(0.430)		(0.0984)	(0.0617)	(0.0546)	
Observations	350	360	360	380	380	17,730	17,730	17,730	17,730
R-squared	0.373	0.429		0.182		0.166	0.165	0.052	
Adjusted R-squared	0.338	0.408		0.149		0.165	0.164	0.0510	

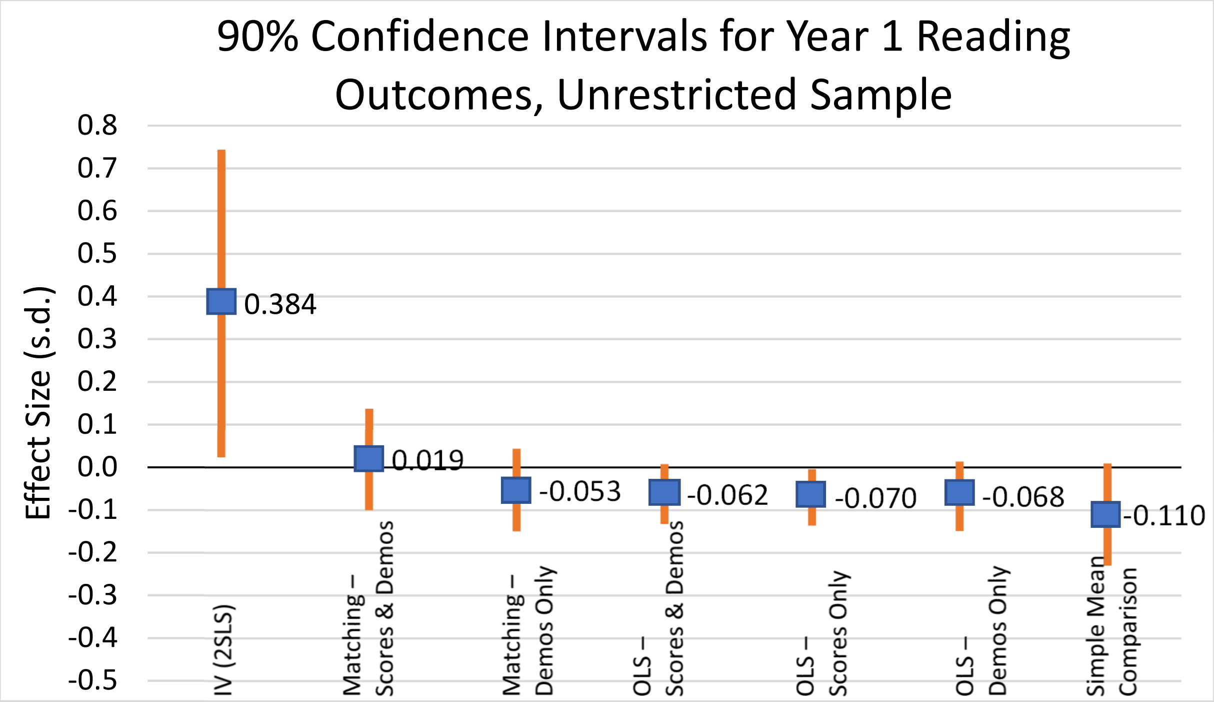
Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Additional RCT Controls include ever attended a SINI school, age, stability (number of months at current residence), number of children in household, mother's education, mother's empoloyement status, and number of days from September 1 until the date of testing.

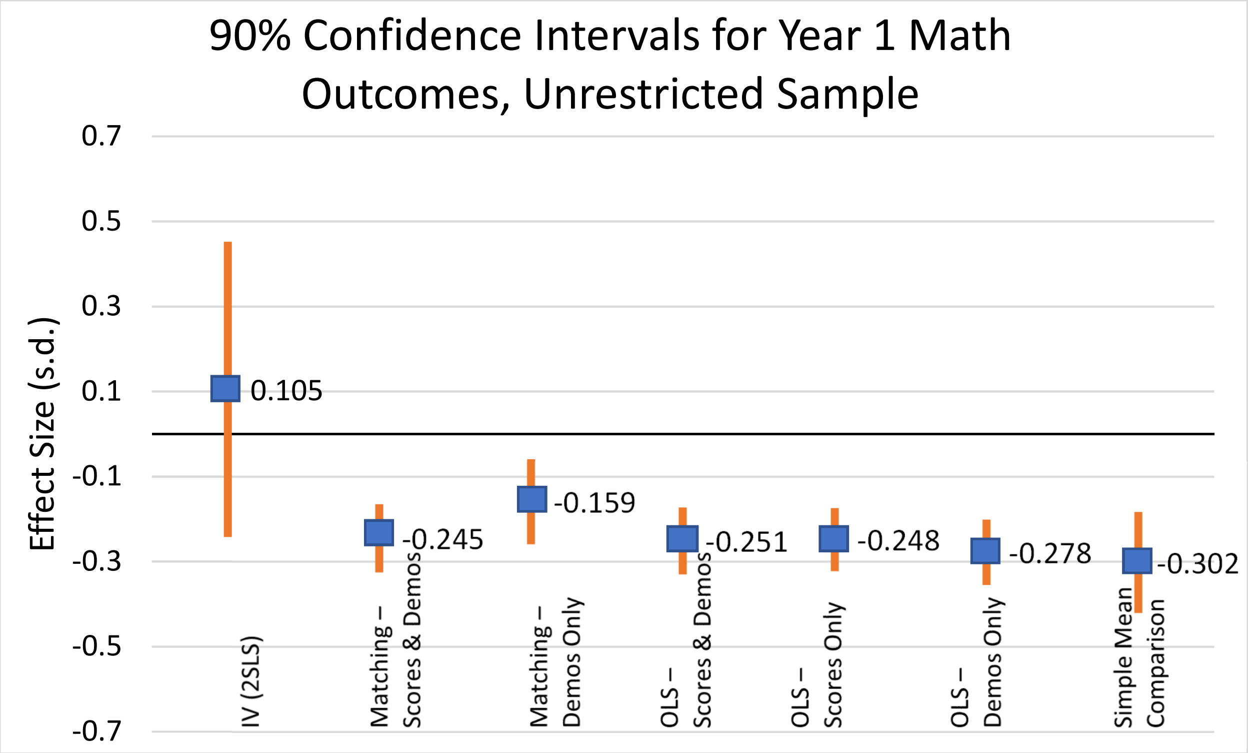
All sample sizes rounded to nearest 10.

Figure 9:



**IV(2SLS) results are for the restricted sample (cohort 1 only), all other results from the unrestricted sample (cohort 1 only).*

Figure 10:



**IV(2SLS) results are for the restricted sample (cohort 1 only), all other results from the unrestricted sample (cohort 1 only).*

Discussion and Conclusions

Most of the clearest results of our study confirm some prior knowledge. A comparison of the restricted and unrestricted sample results reiterates previous findings that sampling frame is important and that estimates comparing to similar, local settings are less biased than estimates from comparison to the broader population (Aiken et al., 1998; Bifulco, 2012; Heckman, Ichimura, and Todd, 1997; Heckman et al., 1998; Shadish, Clark, & Steiner, 2008. While the reading restricted results, and the first two years of the math restricted results show that estimates are somewhat similar across various model types, the two figures comparing the estimates from the unrestricted samples indicate that the non-experimental estimates are much further away from the point-estimate of the experimental estimate (although these experimental estimates are rather noisy).

This study also provides some support for the importance of pre-treatment outcomes as covariates (Bifulco, 2012; Cook et al., 2008; Fortson, et al., 2012; Glazerman et al., 2003; Shadish, Clark, & Steiner, 2008; Wilde and Hollister, 2007). For example in the year 3 and 4 restricted math results, we see that matching methods including baseline test scores in addition to student demographic variables were the only methods that somewhat approximate the point estimate of the experimental estimate. In year 4, in particular, quasi-experimental approaches informed by baseline test score variables all got the findings right while approaches that did not use pre-program test scores all got them wrong. However, in other case, such as the unrestricted math results, some models including baseline test scores produce estimates further away from the experimental estimates, suggesting that the importance of accounting for baseline test scores to approximate experimental estimates might be context-dependent.

However, we do not necessarily have support for a prior finding that choice of covariates is even more important than model choice (Bifulco, 2010), particularly within our unrestricted sample. Rather, we tend to confirm the findings of Fortson et al. (2012) who find that matching generally performs better than descriptive models with controls (at least within our unrestricted samples). For example, in Figures 9 and 10, the estimates from the matching models are all closer in magnitude to the experimental estimates than the estimates from the OLS models. This evidence is only suggestive, however, as our experimental benchmarks in these cases are particularly noisy. In contrast, in our restricted samples, model choice appears to matter less, suggesting that there is less benefit from propensity score matching when the IV sample is already “matching” on desire to apply to a program, and similar demographic characteristics as well.

Finally, our WSC identifies conflicting directions of bias across the restricted and unrestricted sample, especially regarding math outcomes. When quasi-experimental methods were used on data restricted to program applicants, the results tended to indicate that the program had significant positive effects when the experimental LATE estimates suggested that the true effects were null. When those same quasi-experimental methods were used on data that included program non-applicants, the results tended to indicate that the program had significant negative effects when, again, the true experimental effects were null. What might explain this interesting pattern of results?

Applicants to school voucher programs may be negatively selective on unobservable characteristics. This claim flies in the face of most assumptions that voucher programs cream the best and most motivated students (e.g. Levin 1998), but it may be that parents are attracted to school choice programs when their child struggles to fit in at school due to unobservable conditions that are not fully captured by baseline test scores. Quasi-experimental analyses of the effect of voucher programs that include voucher applicants and non-applicants (e.g. Metcalf et al., 2003) might consistently under-estimate the positive effects of vouchers because they cannot control for this apparent negative selection. Among the pool of applicants to voucher programs (i.e. students in our restricted sample), in contrast, those that actually used their voucher appear to have been positively selected, consistent with prior research (e.g. Fleming et al., 2016; Campbell, West & Peterson, 2005; Howell, 2004). Quasi-experimental approaches to estimating voucher effects on applicant samples might consistently over-estimate the positive effects of vouchers because they cannot control for this apparent positive selection. In both cases many of the quasi-experimental estimates were wrong but in one case they were wrong low and in the other case they were wrong high.

While we find these results highly suggestive, our work is limited in important ways. First, our analytic samples and study period are limited by data availability. We only can make use of DCPS test score data for the unrestricted sample for the baseline year and a single outcome year. Unfortunately, the benchmark experimental estimates in both reading and math that first year were noisy zeros, rendering less-than-ideal the comparisons with the quasi-experimental estimates that actually would have increased the external validity of the study.

Second, the current study frame limits the generalizability of these results to different contexts. While selection in to school choice programs is often hypothesized to be positive (meaning that more advantaged families are more likely to opt-in), in our context the private school students were all applicants to a means-tested voucher program (DC OSP), and therefore are not representative of private school students generally. Therefore, the revelation of possible negative selection into the OSP revealed by our analysis may not apply to other school choice programs with different eligibility criteria or to private schools students in general. That particular finding may lack external validity.

Despite these limitations, researchers and policy makers attempting to evaluate educational programs should consider the importance of particular covariates, model choice, and sampling frame when pondering the internal-external validity tradeoff in school choice evaluations.   Especially in the hot-house of school choice research, getting the answer wrong is a risk researchers should do their best to avoid.

References

Abdulkadiroglu, A., Angrist, J., Cohodes, S., Dynarksi, S., Fullerton, J., Kane, T., & Pathak, P. (2009). *Informing the debate: Comparing Boston's charter, pilot, and traditional schools.* The Boston Foundation.

Ackerman, M.A. & Egalite, A.J. (2016). Assessing tradeoffs between observational and experimental designs for charter school research. Presented December 19, 2016 at the Annual Conference of the Association for Education Finance and Policy in Denver, Colorado.

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

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

Barrow, L., & Rouse, C. E. (2008). School vouchers: Recent findings and unanswered questions. *Economic Perspectives*, 32, 2-16.

Bifulco, 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.

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

Busso, M., DiNardo, J., McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics* 96(5): 885-897.

Campbell, D. E., West, M. R., Peterson, P. E. (2005). Participation in a national, means-tested school voucher program. *Journal of Policy Analysis and Management, 24(3),* 523-541.

Center for Research on Education Outcomes (CREDO). (2009). *Multiple choice: Charter school performance in 16 states*. Stanford, CA: CREDO.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin Company.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings.

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

Egalite, A. J., & Wolf, P. J. (2016). A review of the empirical research on private school choice. *Peabody Journal of Education*, 91(4), 441-454.

Fleming, D. J., Cowen, J. M., Witte, J. F., & Wolf, P. J. (2015). Similar students, different choices: Who uses a school voucher in an otherwise similar population of students? *Education and Urban Society, 47(7),* 785-812.

Foreman, L. M., Anderson, K. P., Ritter, G. W., & Wolf, P. J. (2017). *Choosing your validity: Analyzing the impacts of charter schools in a U.S. state using two types of matching designs.* Working Paper.

Fortson, K., Verbitsky-Savitz, N., Kopa, E., Gleason, P. (April 2012). *Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates.* U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2012-4019.

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

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earning impacts. *American Academy of Political and Social Science, 589,* 63-93.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). The evaluation of charter school impacts: Final report. NCEE 2010-4029. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Greene, J. P., Peterson, P. E., & Du, J. (1999). Effectiveness of school choice: The Milwaukee experiment. *Education & Urban Society*, 31(2), 190-213.

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

Heckman, J. J., Ichimura, H., Smith, J. & Todd, P. E. (1998). Characterizing selection bias using experimental data. *Econometrica, 66,* 1017–1098.

Hirano, K., Imbens, G.W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4): 1161-1189.

Howell, W. G. (2004). Dynamic selection effects in means-tested, urban school voucher programs. *Journal of Policy Analysis and Management, 23(2)*, 225-250.

Howell, W. G. & Peterson, P. E. *with* Wolf, P .J. & Campbell, D. E. *The Education Gap: Vouchers and Urban Schools*, (Revised Edition), Washington: Brookings, 2006.

Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management*, 21(2), 191-217.

Hoxby, C. M. (2009). A statistical mistake in the CREDO study of charter schools. Unpublished memorandum. Available at http://users.nber.org/~schools/charterschoolseval/memo_on_the_credo_study.pdf.

Jaciw, A. P. (2016). Assessing the accuracy of generalized inferences from comparison group studies using a within-study comparison approach: The methodology. *Evaluation Review, 40*(3), 199-240.

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

Levin, H. M. (1998). Educational vouchers: Effectiveness, choice, and costs. *Journal of Policy Analysis and Management*, 17, 373-392.

Lubienski, C., & Brewer, T. J. (2016). An analysis of voucher advocacy: Taking a closer look at the uses and limitations of 'gold standard' research. *Peabody Journal of Education*, 91(4), 455-472.

Metcalf, K. K., West, S. D., Legan, N. A., Paul, K. M., & Boone, W. J. (2003). *Evaluation of the Cleveland Scholarship and Tutoring Program: Summary report.* Bloomington, IN: Indiana University School of Education, Indiana Center for Evaluation.

Mosteller, F. & Boruch, R. (2002). *Evidence matters: Randomized trials in education research.* Washington, D.C.: The Brookings Institutions.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*, 20(4), 111-132.

Pirog, M. A., Buffardi, A. L., Chrisinger, C. K., Singh, P., & Briney, J. (2009). Are alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations. *Journal of Policy Analysis and Management, 28,* 169–172.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-24.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach.* Thousand Oaks: Sage Publications.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics*, 113 (2), 553-602.

Sadoff, S. (2014). The role of experimentation in education policy. *Oxford Review of Economic Policy* 30(4), 597-620.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103,* 1334–1343.

Shakeel, M. D., Anderson, K. P., & Wolf, P. J. (2016). *The participant effects of private school vouchers across the globe: A meta-analytic and systematic review*. Economics Research Network, EDRE Working Paper 2016-07.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome Lalonde's critique of nonexperimental estimators? *Journal of Econometrics, 125,* 305–353.

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

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,* 455–477.

Witte, J. F. (2000). *The market approach to education: An analysis of America's first voucher program.* Princeton, NJ: Princeton University Press.

Witte, J. F., Thorn, C., & Steer, T. (1995). *Fifth year report: Milwaukee Parent Choice Program.* Report to the Wisconsin State Legislature. Madison, WI: Department of Public Instruction.

Witte, J. F., Wolf, P. J., Cowen, J. M., Carlson, D., & Fleming, D. F. (2014). High stakes choice: Achievement and accountability in the nation's oldest urban voucher program. *Education Evaluation and Policy Analysis*, 36(4), 437-456.

Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., Eissa, N., & Carr, M. (2010). *Evaluation of the DC Opportunity Scholarship Program.* U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2006-4003.

Wolf, P. Gutmann, B., Puma, M., Rizzo, L., Eissa, N.O. & Silverberg, M. Evaluation of the DC Opportunity Scholarship Program: Impacts After One Year, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Washington, DC: U.S. Government Printing Office, 2007. Retrieved from: https://ies.ed.gov/ncee/pubs/20074009/

Wolf, P., Gutmann, B., Puma, M., & Silverberg, M. (2006). *Evaluation of the DC Opportunity Scholarship Program: Second Year Report on Participation.* U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2006-4003.

Wolf, P., Kisida, B., Gutmann, B., Puma, M., Eissa, N., & Rizzo, L. (2013). School vouchers and student outcomes: Experimental evidence from Washington, DC. *Journal of Policy Analysis and Management, 32*(2), 246-270.

Appendix A – Baseline Equivalency for Matching Samples

Restricted (IV) Sample (Match Based on Test Scores and Demographics) - Includes Both Cohorts

Baseline Equivalency in Math for Year 1 Outcomes

	Private	Public	Difference	P-Value
Number of Observations	460	460	-	
Average Grade	4.59	4.59	-	1.000
Prior Year Math Z-Score	0.024	0.023	0.002	0.977
Prior Year Reading Z-Score	0.067	-0.001	0.068	0.263
% Male	0.49	0.51	(0.02)	0.643
% FRL	1.00	1.00	-	1.000
% Black	0.89	0.88	0.01	0.601
% Special Needs	0.09	0.15	(0.06) ***	0.005

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Baseline Equivalency in Reading for Year 1 Outcomes

	Private	Public	Difference	P-Value
Number of Observations	440	440	-	
Average Grade	5.08	5.08	0.00	1.000
Prior Year Reading Z-Score	0.04	0.03	0.00	0.980
Prior Year Math Z-Score	0.03	0.05	(0.03)	0.662
% Male	0.49	0.50	(0.01)	0.735
% FRL	1.00	1.00	-	1.000
% Black	0.90	0.86	0.03	0.147
% Special Needs	0.11	0.14	(0.04)	0.103

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Unrestricted Sample (Match Based on Test Scores and Demographics) - Cohort 1 Only

Baseline Equivalency in Math for Year 1 Outcomes

	Private	Public	Difference	P-Value
Number of Observations	180	180	-	
Average Grade	7.17	7.17	-	1.000
Prior Year Math Z-Score	-0.04	-0.05	0.00	0.976
Prior Year Reading Z-Score	0.03	0.02	0.01	0.935
% Male	0.49	0.51	(0.02)	0.675
% FRL	1.00	1.00	-	1.000
% Black	0.85	0.88	(0.03)	0.443
% Special Needs	0.14	0.86	(0.71) **	0.012

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Baseline Equivalency in Reading for Year 1 Outcomes

	Private	Public	Difference	P-Value
Number of Observations	180	180	-	
Average Grade	7.25	7.25	-	1.000
Prior Year Math Z-Score	-0.01	0.03	(0.04)	0.657
Prior Year Reading Z-Score	0.02	0.02	0.00	0.973
% Male	0.49	0.48	0.01	0.834
% FRL	1.00	1.00	-	1.000
% Black	0.85	0.90	(0.05)	0.160
% Special Needs	0.14	0.17	(0.03)	0.390

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Restricted (IV) Sample (Match Based on Demographics Only) - Includes Both Cohorts

Baseline Equivalency in Math for Year 1 Outcomes

	Private	Public	Difference	P-Value
Number of Observations	620	620	-	
Average Grade	4.88	4.88	-	1.000
Prior Year Math Z-Score	0.03	0.04	(0.01)	0.863
Prior Year Reading Z-Score	0.06	0.02	0.04	0.473
% Male	0.47	0.49	(0.02)	0.571
% FRL	1.00	1.00	-	1.000
% Black	0.89	0.87	0.01	0.486
% Special Needs	0.10	0.16	(0.06) ***	0.001

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Baseline Equivalency in Reading for Year 1 Outcomes

	Private	Public	Difference	P-Value
Number of Observations	590	590	-	
Average Grade	5.16	5.16	0.00	1.000
Prior Year Math Z-Score	0.04	0.05	(0.01)	0.850
Prior Year Reading Z-Score	0.06	0.03	0.03	0.552
% Male	0.46	0.48	(0.02)	0.561
% FRL	1.00	1.00	-	1.000
% Black	0.88	0.86	0.02	0.382
% Special Needs	0.10	0.16	(0.06) ***	0.001

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Unrestricted Sample (Match Based on Demographics Only)

Baseline Equivalency in Math for Year 1 Outcomes

	Private	Public	Difference		P-Value
Number of Observations	190	190	-		
Average Grade	7.27	7.27	-		1.000
Prior Year Math Z-Score	0.03	0.09	(0.06)		0.510
Prior Year Reading Z-Score	0.07	0.08	(0.01)		0.890
% Male	0.48	0.55	(0.07)		0.183
% FRL	1.00	0.94	0.06	***	<0.001
% Black	0.85	0.91	(0.05)		0.116
% Special Needs	0.14	0.24	(0.09)	**	0.019

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Baseline Equivalency in Reading for Year 1 Outcomes

	Private	Public	Difference		P-Value
Number of Observations	190	190	-		
Average Grade	7.27	7.27	-		1.000
Prior Year Math Z-Score	0.03	0.09	(0.07)		0.465
Prior Year Reading Z-Score	0.07	0.10	(0.03)		0.692
% Male	0.48	0.54	(0.06)		0.219
% FRL	1.00	0.94	0.06	***	<0.001
% Black	0.85	0.91	(0.05)		0.116
% Special Needs	0.14	0.24	(0.09)	**	0.019

**Significant at the 10% level, **Sig. at the 5% level, ***Sig.at the 1% level*

Appendix B - IV Regression Estimates of Year One Reading Outcomes, Differing Cohorts (Both or Cohort 1 Only) and Covariate Type

	Both Cohorts		Cohort 1 Only	
Private Schooling	-0.0333	-0.0338	0.316	0.384*
	(0.0833)	(0.0858)	(0.205)	(0.219)
PY Math Z-score	0.0705*	0.0822**	0.231***	0.287***
	(0.0406)	(0.0411)	(0.0722)	(0.0780)
PY Reading Z-score	0.386***	0.399***	0.251***	0.259***
	(0.0602)	(0.0589)	(0.0965)	(0.0999)
Household Income (000s)	-0.0013		-0.0011	
	(0.0030)		(0.0061)	
Male	-0.0495	-0.0377	-0.0542	-0.00161
	(0.0537)	(0.0539)	(0.0994)	(0.104)
Special Needs	-0.644***	-0.646***	-0.264**	-0.249**
	(0.0846)	(0.0833)	(0.105)	(0.114)
Black	-0.0519	-0.0112	-0.0615	-0.0176
	(0.0737)	(0.0721)	(0.124)	(0.142)
Baseline Grade Indicators	Y	Y	Y	Y
Outcome Grade Indicators	Y	Y	Y	Y
Additional RCT Controls	Y		Y	
Constant	0.415	0.208	2.533**	-0.174
	(0.339)	(0.152)	(1.175)	(0.361)
Observations	1,650	1,650	350	350
R-squared	0.290	0.276	0.397	0.322
Adjusted R-squared	0.276	0.267	0.348	0.285

Note: Entire restricted (IV) sample would have been FRL-eligible.