Comparing Performance of Methods to Deal with Differential Attrition in Lottery Based Evaluations

Kaitlin Anderson, (kpa319@lehigh.edu), Lehigh University

Gema Zamarro, (gzamarro@uark.edu), University of Arkansas

Jennifer Steele, (steele@american.edu), American University

Trey Miller, (tmiller@utdallas.edu), University of Texas at Dallas

January 2020

ABSTRACT:

Background. In randomized controlled trials, attrition rates often differ by treatment status, jeopardizing causal inference. Inverse probability weighting (Hirano et al, 2003; Busso et al., 2014) and estimation of treatment effect bounds (e.g. Lee, 2009; Angrist et al., 2006) have been used to adjust for this bias. **Objectives.** We compare the performance of various methods within two samples, both generated through lottery-based randomization: one with considerable differential attrition and an augmented dataset with less problematic attrition. **Research Design.** We assess the performance of various correction methods within the dataset with problematic attrition. In addition, we conduct simulation analyses. **Results.** Within the more problematic dataset, we find the correction methods often performed poorly. Simulation analyses indicate that deviations from the underlying assumptions for bounding approaches (Angrist et al., 2006) damages the performance of estimated bounds. **Conclusions.** We recommend the verification of the underlying assumptions in attrition correction methods whenever possible and, when verification is not possible, using these methods with caution.

1. INTRODUCTION

Since its introduction by Angrist (1990) to evaluate the impact of military service on earnings, a growing literature has made use of lottery-based randomization to estimate causal effects of educational and other social programs (Abdulkadiroglu et al., 2009; Angrist, Bettinger, Bloom, Kremer, & King, 2002; Cullen, Jacob, & Levitt, 2006; Deming, Hastings, Kane, & Staiger, 2014; Dobbie & Fryer, 2009; Engberg, Epple, Imbrogno, Sieg, & Zimmer, 2014; Hoxby & Rockoff, 2004; Rouse, 1998).

Lottery-based analyses are a subset of Randomized Controlled Trials (RCTs). By comparing average outcomes of those placed and not placed, researchers hope to estimate causal effects of programs and interventions, unaffected by selection bias. However, it is common for participants not assigned to treatment to seek outside options. For example, in educational interventions, control students often seek opportunities outside the district, by moving or attending a charter or private school. If attrition mechanisms differ based on random assignment, this creates a differential attrition problem, jeopardizing the identification of causal effects.

Differential attrition and selection bias are common among social experiments[1] (Greenberg & Barnow, 2014). For example, in a review of development economics RCTs published between 2009 and early 2015, Molina and Macours (2015) find that 19 percent had differential attrition, and in many cases, authors simply restrict the analysis to a subsample in which attrition was balanced. Unfortunately, WWC reports do not allow us to precisely estimate the share of RCTs that did not meet their attrition standards,[2] but the What Works Clearinghouse (2014) suggests that bias from empirical studies in education has been fairly modest, generally

[1] The authors include "social experiments" related to health, education, employment, job training, welfare, and housing.

[2] The study design (RCT, QED, RD, etc.) is only reported for studies that did meet the standards, preventing us from knowing the denominator for such a proportion.

not exceeding 0.11 SD, though theoretically it could go much higher if attrition and baseline characteristics are highly correlated.

Removing all selective attrition bias would be possible if either all covariates determining the outcome are known (Steyer, Gabler, von Davier, & Nachtigall, 2000); or the selection process is completely known (Cook, 2008; Goldberger, 1972; Shadish, Cook, & Campbell, 2002). Often, however, researchers cannot directly observe all covariates or accurately model the selection process (Puma et al., 2009), so selection bias due to attrition remains an issue.

In this paper, we study the performance of methods aimed at correcting differential attrition, specifically inverse probability weighting (IPW) methods (Busso, DiNardo, & McCrary, 2014; Hirano, Imbens, & Ridder, 2003) and two common bounding approaches (Angrist et al., 2006; Lee, 2009). We use administrative data for seven cohorts of lottery applicants to dual-language immersion programs (DLI) in Portland Public Schools (PPS), a large urban school district. This district-level dataset, on its own, suffered from differential attrition, which may have biased the treatment effect estimate. In this context, it is not surprising that families who do not win the lottery may be less likely to enroll in the district, creating potential for selection bias. However, differential attrition rates do not necessarily indicate that selection bias will result, for example if attrition – despite occurring at different rates in the treatment and control group - is completely random. Moreover, without knowing if the assumptions for these methods are met, we could introduce more bias into our estimates. To study the risks and benefits of various correction methods, we use an augmented dataset for the full state of Oregon public schools, including charters, provided by the Oregon Department of Education (ODE), which suffers much less from both overall and differential attrition. In the absence of an augmented dataset, we may be tempted to rely on correction methods without fully

understanding whether such methods are appropriate, but a unique feature of our study is that we are able to use benchmark effects obtained from the state-level dataset to assess the performance of these correction methods within the problematic district-level dataset.

The use of a benchmark to test the relative performance of correction methods is not new (Cook, Steiner, & Pohl, 2009; Dehejia & Wahba, 1999; Garlick & Hyman, 2016; Heckman, Ichimura, Smith, & Todd, 1998; LaLonde, 1986; Robins & West, 1986; Smith & Todd, 2005; Steiner, Cook, Shadish, & Clark, 2010). In some cases, administrative data and survey data have been used to supplement each other in the presence of missing data (Barnow & Greenberg, 2015; Greenberg & Barnow, 2014; Robins & West, 1986). For example, Robins and West (1986) supplemented interview data from the Seattle and Denver Income Maintenance Experiments (SIME/DIME), which suffered from attrition, with Social Security Administration data, to assess the performance of various correction methods and assess the extent to which the estimated results under the original evaluation would have been biased. Notably, studies often lack an experimental or quasi-experimental benchmark against which correction models can be evaluated (Clark, Rothstein, & Schanzenbach, 2009; Melenberg & van Soest, 1996; Mroz, 1987; Newey, Powell, & Walker, 1990), so it is useful to understand the implications of not being able to assess the necessary assumptions for those methods. While we are not the first to use a benchmark to test the performance of various correction methods, there is little research that explicitly assesses and acknowledges the dangers of using these under often untestable assumptions. In particular, few studies have directly addressed the sensitivity of results using Angrist et al. (2006) bounds. See Barrow, Richburg-Hayes, Rouse, and Brock (2014) for one discussion of the sensitivity to the choice of artificial censoring points. In addition, we conduct simulation analyses that alter the degree to which the assumptions for various methods are met,

and then quantitatively assess at what point the methods are unable to accurately estimate the parameters of interest.

Specifically, we ask:

1. *Do various correction methods (inverse probability weighting or estimation of informative bounds) adequately compensate for differential attrition in a random assignment evaluation?*
2. *How do various assumptions within these methods affect our results?*

The rest of the paper proceeds as follows. Section 2 reviews the literature on differential attrition correction methods. Section 3 describes the data and sample for the analysis. Section 4 describes the empirical methods studied in this paper, and Section 5 presents the results of tests of testing whether the necessary assumptions for these methods are met. Section 6 presents the results of the main analysis using the non-simulated data and Section 7 describes our simulation exercise. Finally, section 8 outlines our main conclusions.

2. REVIEW OF THE LITERATURE ON DIFFERENTIAL ATTRITION CORRECTION METHODS

A common approach for minimizing differential attrition bias is IPW (e.g., Bailey, Hopkins, & Rogers, 2016; Imbens & Wooldridge, 2009; Reynolds, Temple, Ou, Arteaga, & White, 2011; Frölich & Huber, 2014; Molina & Macours, 2015; Muralidharan & Sundararaman, 2014). In this case, observations in the treatment and control group are reweighted to remain comparable to their pre-attrition samples. While the importance of the specification of the propensity score model has been studied (e.g., Austin & Stuart, 2015; Hogan & Lancaster, 2004; Wooldridge, 2007), in practice, researchers often have limited information with which to assess the appropriateness of their approach (Puma et al., 2009). As a result, Puma et al. (2009) find that

sophisticated weighting methods often do not reduce the bias to the What Works Clearinghouse threshold of 0.05 SD (What Works Clearinghouse, 2013), when data are not missing at random, as is likely to be the case when data exhibit high rates of differential attrition.

Alternatively, researchers have used bounding methods to estimate a range of possible effects under different attrition scenarios. Molina and Macours (2015) find that bounding methods were used in almost 15 percent of the 68 studies in their review, although their review may not be representative of the approaches taken within education, which may be less likely to use these types of methods. For example, Puma et al.'s (2009) assessment of correction methods does not test the performance of bounding methods, instead discussing some of the practical challenges with certain bounding approaches, and leaving out the newer Angrist et al. (2006) and Lee (2009) approaches.

In our review of the literature, focusing primarily on educational interventions, when bounding is used, the most popular approach appears to be that proposed by Lee (2009). See, for example: Aker and Ksoll (2015); Aron-Dine, Einav, and Finkelstein (2013); Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2013); Boo, Palloni, and Urzua (2014); Di Nardo, McCrary, and Sanbonmatsu (2006); Engberg et al. (2014); Glewwe, Illias, and Kremer (2010); Hastings, Neilson, and Zimmerman (2012); Karlan, Fairlie, and Zinman (2012); Kremer, Miguel, and Thornton (2009); Molina and Macours (2015); and Muralidharan and Sundararaman (2014). A similar approach is Manski's worst-case scenario bounds (Horowitz & Manski, 1998; Horowitz & Manski, 2000, Imbens & Manski, 2004; Manski, 1990; Manski, 1995). See, for example: Aron-Dine et al. (2013); Bailey et al. (2016); DiNardo et al. (2006); Holm and Jaeger (2009); Lechner and Melly (2010); Karlan et al. (2012); and Ksoll, Aker, Miller, Perez-Mendoza, and Smalley (2014). Both methods obtain bounds for extreme case scenarios under relatively weak

assumptions about nonresponse. These methods tend to provide wide and largely uninformative bounds, particularly if the rate of missing data is high (Puma et al., 2009).

Other researchers have proposed bounding the estimates based on certain assumptions about which participants who are not observed in the outcome data. These include the parametric and non-parametric bounding approaches proposed by Angrist, Bettinger, & Kremer (2006) as in Barrow, Richburg-Hayes, Rouse, and Brock (2014), as well as extensions and modifications of these bounding methods (e.g., Engberg et al., 2014; Grilli & Mealli, 2008; Huber & Mellace, 2013; Lechner & Melly, 2010; Zhang & Rubin, 2003; Zhang, Rubin, & Mealli, 2008).
In contrast with the Lee and Manski approaches, these bounding approaches make restrictive assumptions about which participants are missing from the outcome data. For example, Angrist et al. (2006)'s parametric and non-parametric bounding approaches assume that non-respondents/missing observations come from only one side of the outcome distribution. Other approaches (Huber & Mallace, 2013; Grilli & Mealli, 2008; Lechner & Melly, 2010) derive bounds under the assumption of stochastic dominance, that "the potential outcome among the always observed at any rank of the outcome distribution and in any treatment state is at least as high as that of the compliers or the defiers" (Huber & Mallace, 2013, p. 17). This assumption is not imposed by Lee (2009). Engberg et al. (2014) estimate informative bounds around the treatment effects in a magnet program using a "worst-case" scenario approach (Horowitz & Manski, 2000; Manski, 1990), assuming that "the support of the outcome variable is bounded to deal with nonrandom attrition" (p. 29). Building on this approach, they use known quantiles of the outcome distribution in constructing the bounds, similar to the Angrist et al. (2006) approach. By imposing assumptions on the attrition process, these bounding approaches can provide tighter bounds than those from more relaxed approaches like Lee (2009). However, researchers often

lack a full understanding of who the non-respondents/missing participants are, and there is little prior research on the consequences of imposing these assumptions if they turn out to be invalid.

3. DATA AND SAMPLE

This study utilizes data for seven cohorts of students who applied to attend a DLI program in Portland Public Schools (PPS) in Portland, Oregon for the school years beginning fall of 2004-2010. PPS serves about 47,000 students and is among the largest two public school districts in the Pacific Northwest.[3] Outcome data were measured through the 2013-14 academic year, so the oldest two cohorts can be tracked through eighth grade, and include grade 3-8 reading test scores on the state test, the Oregon Assessment of Knowledge and Skills (OAKS).[4]

Slots to DLI were assigned through a lottery system. In the spring prior to their child's pre-k or kindergarten year, families were able to apply for up to three school programs, including DLI. In many of the DLI programs, priority was given based on sibling, neighborhood, or native speaker preferences. Consequently, the probability of admission to a program depended on the program to which one applied and the program-specific preference category into which one fell. Randomization occurred only in school-by-year-by-preference category strata that had available slots and were oversubscribed. In other words, lotteries are considered "valid" only if there are winners and non-winners within a given lottery strata in a given year. The lottery-applicant sample included 3,457 students, 1,946 (56.3%) of which participated in valid lotteries. Of the 1,946 that participated in valid lotteries, 864 (44.4%) won DLI slots and 1,082 (55.6%) did not.

Attrition

This paper builds on the evaluation of this program presented in Authors (2017). Our

[3] For more details on the lottery process or the dual-language immersion programs in PPS see Authors (2017).
[4] Math test scores for these grades were also available, but for simplicity we focus here on the reading test scores. Full set of results are available at Authors (2017). For the simulation analyses, we focus only on grade 3 outcomes, although these models could theoretically be extended to future grades as well.

outcome of interest in this paper is students' reading assessment scores on the OAKS, as measured in grades 3 through 8. We captured these outcomes first through data provided by the district, but due to high overall attrition (27%) and an average differential attrition rate of approximately 12.5%, we supplemented these with state-wide data provided by the Oregon Department of Education (ODE). The ODE data included the PPS data but allowed us to track students who had left PPS as long as they remained in Oregon public schools—charter or traditional. As a consequence, overall attrition declined to 16.5%, and sample differential attrition declined to an average differential attrition rate of just 3%, allowing our main analysis (Authors, 2017), which used the augmented dataset, to fall within the liberal threshold (and very near the conservative threshold) for meeting the U.S. Department of Education's What Works Clearinghouse evidence standards (WWC, 2014).[5] While this does not guarantee that no differential attrition bias exists, the risk of bias is likely lower than in the district-level dataset. Thus, the ODE state-level data provides a unique opportunity, serving as the benchmark dataset in our current analysis examining the ability of IPW and bounding methods to address the higher levels of attrition in the district-provided (PPS) dataset.

Table 1 presents data on persistence in the valid lottery sample by lottery cohort in both the ODE benchmark sample and the PPS higher-attrition sample. The set of columns on the left shows the number of students in valid lotteries by cohort. The next three columns show the number of randomized students we would be able to observe in each tested grade if there were zero attrition. The number declines by grade because only the older cohorts can be tracked into the higher grades, and the only grade into which all seven cohorts are tracked is grade 3. The

[5] The What Works Clearinghouse (WWC) is an initiative of the U.S. Department of Education's Institute of Education Sciences that evaluates research studies on the effectiveness of educational interventions. The WWC produces Standards Briefs to explain the rules used to assess the quality of studies, including issues such as attrition. For more information, visit the WWC's webpage at http://whatworks.ed.gov.

next pair of columns shows the proportion of students randomized who were observed in a given grade in the ODE benchmark sample, among lottery winners and losers, respectively. The far right pair of columns reflect the proportion observed, by grade and treatment status, in the higher-attrition PPS sample. Though differential attrition is present in both samples, what is important here is that the ODE samples had substantially lower levels of overall and differential attrition, meeting the WWC liberal threshold in grades 3 through 6, and even meeting the WWC conservative threshold in grades 7 and 8. In contrast, differential attrition in the PPS sample is substantially higher, and none of the comparisons in the PPS sample fall within the WWC conservative or liberal attrition thresholds. Thus, we attempt to replicate estimated effects in the ODE sample by applying weighting and bounding methods to the higher-attrition PPS sample.

[TABLE 1 AROUND HERE]

To assess what types of students decide not to enroll in PPS, we estimate discrete choice probit models predicting whether observation in PPS in future years (enrollment and a valid test score). Separate models were used for each grade level, because attrition mechanisms may re by grade. For example, decisions to leave the district may occur differently at the transition to middle school. We estimate these models among the set of students randomized at baseline and observable in the ODE data in the outcome year. The kindergarten and third grade results are in Table 2, for the full sample, and for treatment and control units separately. The last two columns indicate whether the results differ between treatment and control. While the kindergarten and third grade models did not have baseline test scores available as possible controls, we used lagged test scores as an additional control variable in each probit model from grade 4-9.

[TABLE 2 AROUND HERE]

In the full sample, winning the lottery was the most significant predictor of enrollment in

PPS, consistent with the differential attrition described previously. Black students were less likely to be observed in PPS in grade 3, relative to white students, and students without an identified race in the data were less likely to be observed in PPS in kindergarten and third grade. Students eligible for free- and reduced-price lunch (FRPL) and special needs students were more likely to be enrolled in PPS in kindergarten. Finally, students whose first language was not English were less likely to be observed in PPS in kindergarten and third grade.

The results for the treatment and control groups indicate somewhat different attrition processes for the two groups. The last two columns of Table 2 test whether the coefficients differ between the treatment and control groups and indicates differences for Black students, FRPL students, and special needs students within the two groups. However, in both the treatment and control groups, students with a missing race variable and students whose first language was not English are less likely to enroll in PPS. Overall, Table 2 indicates that the attrition processes differ between the treatment and control groups, and we might expect they would differ in terms of unobservable characteristics as well.

We show the results of balance tests on covariates between lottery winners and non-winners for the two samples (the benchmark ODE sample and the PPS sample) in Table 3. On the left, we show covariate balance for the observable kindergarteners (no test scores available) and third graders (with observed reading test scores) in the ODE sample. Further, in Panel B, we show the covariate balance within the PPS samples (both unweighted and inverse probability weighted).[6] In addition to the unadjusted differences, we provide differences and p-values that have been adjusted for lottery strata fixed effects and therefore indicate within-strata balance. Table 3 shows that lottery winners in these samples were less likely to be of an "other race," and

[6] Given that inclusion in the PPS sample requires having an observed test score in third grade, the kindergarten covariate balance for the PPS would have been redundant to include.

for the ODE sample in third grade, slightly less likely to be female. Otherwise, the groups are observably similar. In our particular context, even relatively high rates of differential attrition from the PPS sample did not create much imbalance in terms of observable characteristics. Yet, given the relatively limited set of observable characteristics, which notably excludes baseline test scores, there could still be concerns about selection on unobservables that might bias our results.

[TABLE 3 AROUND HERE]

4. EMPIRICAL METHODS

Following the analysis in Authors (2017), our main specification for estimating the effect of winning the DLI lottery on student academic performance is the following:

$$y_{it} = \beta_0 + \left(\boldsymbol{DLI_i^{kg} G_{it}}\right)\boldsymbol{\alpha} + \boldsymbol{G_{it}\delta} + \boldsymbol{l} + \boldsymbol{X_i\gamma} + \varepsilon_{it}, \qquad (1)$$

where y_{it} represents reading test scores for student *i* at time *t*. $\boldsymbol{G_{it}}$ is a vector of grade-level fixed effects and $\boldsymbol{l}$ denotes lottery strata fixed effects. The key variables of interest, $\boldsymbol{DLI_i^{kg} G_{it}}$, denote the effect of winning a DLI slot in kindergarten for each grade level. This calculates treatment effects separately by grade, rather than one overall treatment effect, consistent with the main analysis (Authors, 2017). $\boldsymbol{X_i}$ denotes student characteristics observed in kindergarten, including race/ethnicity, gender, FRPL status, whether the child's first language is English, and whether the child is classified as needing special education services.

We obtain estimates of the model in (1) using pooled ordinary least squares and obtained clustered-robust standard errors, ε_{it}, at the student level.[7] Estimates from (1) will represent the average intent-to-treat (ITT) parameter of winning a place in the DLI program on student achievement, which is the estimated effect of winning the lottery. We consider the ITT estimate

[7] Authors (2017) estimated a student random effects model instead. Using this more efficient estimation approach lead to slightly more significant effects in several grades.

our parameter of interest when exploring the performance of different methods to correct for selective attrition. The estimates the effect of winning the lottery, regardless of whether the individual actually obtained treatment. Instrument variables approaches could be used to estimate the treatment for compliers, but such estimates are often less precise.

Our benchmark estimates are obtained by estimating model (1) on the state-sourced dataset (ODE), still restricted to the lottery sample. In the other extreme, we estimate equation (1) restricting the sample to lottery participants who eventually enrolled in PPS. These ITT effects are referred to as the "Naïve PPS" results and may be contaminated by differential attrition bias. Finally, we conduct analyses using a variety of attrition correction methods and compare to the benchmark results. In the section that follows, we describe the correction methods and test the assumptions required for each, noting any limitations in our ability to do so.

5. TESTING NECESSARY ASSUMPTIONS

Inverse probability weighting

First, we describe our IPW approach and assess whether the required assumptions appear to be met in our case. Under the strong assumptions of selection on observables, or conditional independence, and common support between treated and controls, we reweight the remaining observations of treated and controls so they remain comparable to their respective benchmark sample on observable characteristics.[8] We define $\widehat{\prod(X_\iota)}$ as the estimated probability of being observed in the PPS analytic sample. For each grade, and separately for treated and control students, we estimate the propensity of being observed as a function of the child's race/ethnicity, gender, FRPL status, whether the child's first language is English, and whether the child is

[8] Additionally, some researchers would weight observations to make treatment and comparison groups similar on observable characteristics, however in our case, given the limited differences we observe between treatment and comparisons units, even after differential attrition, we do not anticipate this will make a difference in our results.

classified in kindergarten as needing special education services. We then define weights as $\frac{1}{\widehat{\prod(X_i)}}$ and use weighted least squares to obtain estimates of the average ITT effect of winning the DLI lottery. In addition to weighting, these models also control for observable characteristics, representing a double-robust IPW method.

Although relatively easy to compute, the IPW approach relies on two strong assumptions:

1) *Common support*, meaning that the range of values of the propensity score for the propensity to be observed) has sufficient overlap among those observed in PPS and those not observed in PPS. We test this separately among treatment and comparison units, in each case comparing the distribution of propensity scores for those observed in PPS in grade 3 to those not observed. We do have some concerns about common support – in particular, while there are sufficient students who were not observed in PPS with high propensity scores, there is a lack of students observed in PPS with low propensity scores.

2) *Conditional independence* assumes we have enough information about participants to fully model selection, such that treatment status, conditional on observable characteristics, is random. This is a strong assumption given the limited information typically available in education records, and in our case, a lack of baseline test scores.

Given an inability to assume the assumptions required for IPW are fully met, we also study a variety of bounding approaches that have been proposed to relax the assumption of selection on observables.[9]

Lee (2009) bounds

[9] We do not present the results of Angrist et al.'s (2006) parametric method, as the normality assumption is not met in our case. Another option is a two-step parametric selection model correction (Heckman, 1979) which is less sensitive to the normality assumption, but requires an exclusion restriction (Puma et al., 2009). Thus, it is often impractical. Our data did not support the use of such an exclusion restriction, so we do not study that approach here.

Next, we test a bounding approach proposed by Lee (2009). The idea is to identify the "excess" number of students who are induced to enroll in the district because of winning the lottery and then "trim" the upper and lower tails of the observed test score distribution by this number. In this way, one would have bounds for the average ITT effect of DLI assuming that either the best or worst students in terms of test scores are the ones deciding not to enroll.

This approach requires two key assumptions:

1) The treatment variable is independent of the errors in the outcome and selection equation. In our case, this is guaranteed through lottery-induced randomization.

2) The selection equation can be written as a standard latent variable binary choice model, where treatment assignment only affects enrollment in the district in one direction (i.e. winning makes everybody either more probable to enroll in the district or less probable). This assumption is not possible to test, but there is no reason to suspect that winning a lottery to which a family chose to apply to would make enrollment in the district less likely. Given this, and our results in Table 2 which shows that winning the lottery was associated with higher rates of enrollment in PPS, we assume winning increases the likelihood of enrolling for all students.

Thus, for the two key assumptions required for Lee (2009) bounds, one is clearly met due to randomization, and the other, while not fully testable, is intuitively attractive.

Lee's (2009) bounding method works as follows. The observed distribution of test scores for lottery winners is a mixture of two distributions: 1) the distribution for those who would have enrolled in the district regardless and 2) the distribution of those induced to enroll because of winning the lottery. We estimate the proportion of lottery winners that were induced to enroll because of winning the lottery in the following way:

$$p = \frac{\Pr(enrolled\_PPS \mid Win = 1) - \Pr(enrolled\_PPS \mid Win = 0)}{\Pr(enrolled\_PPS \mid Win = 1)} \qquad (2)$$

Each of the probabilities in (2) is estimated from the data. As in many cases, when only district data are available, and without baseline test scores, it is impossible to know the characteristics of those induced to enroll by winning the lottery. This method proposes to construct extreme case scenarios by assuming they are either the very best students in terms of test scores or the very worst.[10] Thus, trimming the data for lottery winners by the estimated proportion of excess students (*p*), estimated following equation (2), in the top and bottom of the test score distribution, will provide us with bounds for the average ITT effect of those who would enroll in PPS irrespective of the treatment or "always enrollees" (Lee, 2009).

Lee's (2009) bounding approach requires few assumptions but in practice can lead to bounds that are wide and uninformative. Covariates can be included to help estimate tighter bounds (Tauchmann, 2014; Ksoll et al., 2014). To tighten the bounds, one would choose discrete variables that have explanatory power for the probability of enrolling in the district. Then, one would split the sample into cells defined by these variables and compute separate bounds for each cell. The average computed bound, weighted by the proportion of the sample in each cell, provides an estimate of the average ITT effect among "always enrollees."[11]

Angrist et al. (2006) parametric and non-parametric bounding approaches

The final bounding method we study in this paper is the non-parametric bounding approach proposed by Angrist et al. (2006). We tested the assumptions for the Angrist et al. (2006) parametric approach as well, which requires the uncensored latent test score distribution to be normally distributed. If this assumption holds, the treatment effect can be recovered using a

[10] While in theory, one could test this if they had outcomes for the full distribution (as we do in our unique situation), in practice, researchers generally do not have outcome measures for the full (pre-attrition) sample.
[11] Similarly, Behaghel, Crépon, Gurgand, and Le Barbanchon (2015) suggest using information about how difficult it is to reach respondents such as the number of attempts made to reach each. This type of paradata is not observable in our data.

Tobit regression. This method also requires the assumption that those not observed in PPS would not have scored below the chosen censoring point (q1), although various censoring points can be tested to assess the robustness of the results. The normality assumption was not met in our case,[12] so we do not use this approach to attempt to correct for differential attrition in the PPS sample. It is worth noting, however, that in practice, one would not be able to test this assumption without the availability of the augmented dataset.

The non-parametric method (Angrist et al., 2006) relaxes the normality assumption. It generally leads to tighter bounds than Lee's (2009) and requires three key assumptions: Selection bias only affects one part of the test score distribution; those not enrolling in PPS are either the highest performing students or the lowest performing students. In practice, researchers generally do not have outcome measures for the full (pre-attrition) sample, however, in our unique situation, we can test this using the outcomes for the full distribution. We test this assumptions two ways: 1) using baseline characteristics, we predict what type of students do not enroll in the district, following Table 2, a strategy that would be applicable to researchers who do not have the opportunity to recover missing observations with an augmented dataset and 2), we use the augmented dataset to compare the overall test score distribution to that of non-responders – a strategy that is only available to us given the augmented dataset.

For this first test, as shown in Table 2, we do not have strong evidence that those deciding to enroll in the district are likely to be either the highest or lowest performing students. Judging by special needs status and FRPL status, the results suggest that those leaving the district, particularly control units, are relatively advantaged and thus, we may assume they would have higher potential test scores. However, students whose first language is not English leave at

[12] Specifically, we performed skewness and kurtosis tests, separately by grade level and subject. We reject the null hypothesis that the test score distribution is normally distributed in grades three, four, five, and nine.

higher rates, and in a highly diverse area, it's not clear whether this is likely to be associated with advantage or disadvantage. Thus, these data do not provide a clear indication that this assumption has been met.

In addition, we show histograms of the test scores of control group students who had an observed grade 3 reading test score in in the augmented data, and for the subset of these students who did not enroll in PPS, respectively, in Figures 1 and 2. The results clearly indicate that the PPS non-enrollees come from across the full distribution of potential test scores, indicating that the required assumption – that they come from one part of the distribution - has not been met.

[FIGURE 1 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

1) As in Lee (2009), we must assume that winning the lottery affects enrollment in the district only in one direction. Again, this is not directly testable, but we assume it makes all students more likely to enroll.

2) We also must assume that treatment affects test scores in one direction. In our context, this would require assuming that winning the lottery does not harm individuals (or that individuals can opt out rather than suffer harm). For example, families choosing to apply for DLI may do so for one or more different reasons including: (1) to get their kids into a better school that may improve their learning generally, (2) to help them become proficient in another language, or (3) to get them into a positive and diverse environment. Rational parents are maximizing a complex objective function that includes these three outcomes, and they may be willing to trade off (1) to improve (2) and (3). Some parents may choose to put their children in DLI even if they think it may lead to slightly reduced reading test scores, so that they can expose their children to a diverse environment and help them become bilingual. Subgroup results from

the main analysis (Authors, 2017) indicates positive or null effects – and no negative impacts – by grade eight for each of the groups assessed.[13] However, we acknowledge that (based on an understanding of why families might choose DLI) this assumption may not be met.

We define $q_o(\theta)$ and $q_1(\theta)$ as the value in the test score distribution corresponding to the θ quantile for those who lost the lottery and for those who won the lottery, respectively. Under the assumption that winning the lottery has positive effects, $q_1(\theta) > q_o(\theta)$, Angrist et al. (2006) showed that non-parametric bounds can be obtained, using linear regression, as follows:

Upper-bound: The average ITT effect estimated when the distribution of test scores for treated students is smaller than $q_1(\theta)$ and the distribution of test scores for students who lost the lottery is smaller than $q_o(\theta)$.

Lower-bound: Estimated average effect when the distribution of test scores of both treated and controls is conditioned to be lower than $q_o(\theta)$.

Assessing Correspondence

To assess whether differential attrition-correction methods produce results similar to the benchmark results, we combine statistical tests of equivalence and difference (Steiner & Wong, 2018). The null hypothesis of the difference test (Tryon, 2001) is that the difference between the estimates equals zero: $H_0: \beta_A - \beta_B = 0$. Given that the samples are not independent, the estimated standard error of the difference must account for this dependency. Bifulco (2012) demonstrates that upper and lower bounds of the standard error can be calculated assuming that the sampling distributions of estimates have a correlation of zero and one respectively.[14]

[13] Subgroup analyses were conducted by program type, and native language (by whether home language is English and by whether home language is the program's partner language).

[14] Specifically, assuming that the correlation of the sampling distributions for two different estimates equals zero, the standard error of the bias estimate can be computed as the following, where $\hat{\delta}_{est1}$ and $\hat{\delta}_{est2}$ are the two estimates: $\mathrm{SE}_{\mathrm{bias}} = \sqrt{\mathrm{var}(\hat{\delta}_{\mathrm{est1}}) + \mathrm{var}(\hat{\delta}_{\mathrm{est2}})}$. Similarly, assuming the correlation of the sampling distributions for two

The second test is a test of statistical equivalence within a given threshold, δ_E. The null hypothesis is that the absolute difference between the estimates is larger than this threshold: $H_0: |\beta_A - \beta_B| \geq \delta_E$. This can also be stated as two one-sided tests: $H_{01}: \beta_A - \beta_B \geq \delta_E$ and $H_{02}: \beta_A - \beta_B \leq -\delta_E$. To reject the null hypothesis of non-equivalence, the null hypotheses of both one-tailed tests must be rejected. Following Steiner and Wong (2018), we assess statistical difference and statistical equivalence at a tolerance threshold of 0.1 s.d., and to align with the What Works Clearinghouse's (2013) bias threshold for creating its attrition thresholds, we report correspondence at an equivalence threshold of 0.05 s.d. as well.

Steiner and Wong (2018) explain that the results of these two tests, combined, lead to one of four conclusions about the degree of correspondence between two estimates. To conclude statistical equivalence, both the difference and equivalence tests must suggest equivalence. To conclude statistical difference, both tests must suggest a difference. If both tests fail to reject the null hypotheses, perhaps due to low power, we conclude the tests are indeterminate. When both tests reject the null hypotheses (the equivalence tests indicates equivalence and the difference tests indicates a difference), perhaps due to high power, we conclude there is a trivial difference.

6. RESULTS

Inverse Probability Weighting

Inverse probability weights were created using predicted probabilities of the probit model estimates in Table 2. Table 4 compares the reading results for three models: the benchmark ODE model, an unweighted ("naïve") PPS model, and an IPW PPS model. The benchmark ITT effect

different estimates equals one, the standard error of the bias estimate is the following: $\text{SE}_{\text{bias}} = \sqrt{\text{var}(\hat{\delta}_{\text{est1}}) + \text{var}(\hat{\delta}_{\text{est2}}) - 2\sqrt{\text{var}(\hat{\delta}_{\text{est1}})\text{var}(\hat{\delta}_{\text{est2}})}}$. Alternatively, standard errors could be bootstrapped to account for dependencies in the samples, although in our case, using bootstrapped samples would arbitrarily impose a new set of attrition mechanisms on each sample would create more noise and uncertainty.

in the benchmark ODE sample was positive and significant in grade 5 (0.150 s.d.) and grade 8 (0.232 s.d.).[15] We also estimate significant and positive effects in these grades in the "naïve" PPS sample, while the IPW PPS models also estimate positive effects in grade 6 as well. The bias columns indicate the difference between the PPS estimates (naïve or IPW) and the ODE benchmark estimate. To indicate statistically difference and/or equivalence, we present the results of the Steiner and Wong (2018) correspondence test (equivalence, difference, trivial difference, or indeterminate) under three different assumptions about the correlation between the sampling distribution of the two estimates (corr. = 0, 0.5, 0.75, or 1). These correlations are also used to calculate the standard error of the bias, in parentheses in the same columns. For equivalence testing, we report results for equivalence thresholds of 0.05 and 0.1 s.d. of bias. The results of the difference test are the same across these threshold increases, all else equal, it is more likely that we will conclude equivalence (absolute value of the bias less than that threshold).

When the correlation is assumed to be either 0 or 0.5 (unlikely assumptions given the strong overlap between the PPS and ODE samples) all findings are indeterminate. At an assumed correlation of 0.75, and an equivalence threshold of 0.5, all findings are indeterminate as well. At an assumed correlation of 0.75, and with a higher threshold for equivalence, 0.1 s.d., both the naïve PPS and the IPW PPS models find "equivalence" in 3 out of 6 cases. At a correlation of 1, the estimated standard error of the bias is reduced enough that we reject the nulls of equivalence and difference in many cases, leading to many conclusions of "trivial differences," however, in this case, saying a difference within 0.1 s.d. is perhaps misleading and merely a function of a larger threshold. Perhaps most importantly, when looking at the grade levels for which we

[15] For further discussion of these results, see Authors (2017).

estimate statistically significant effects, the results were generally trivially different in grade 5, but the IPW PPS model was biased significantly upwards in grade 8, and would lead to a false positive in grade 6. Overall, these findings indicate that the IPW is not performing better than simply using the "naïve," unweighted estimates (not producing more equivalence), and if anything, is introducing more bias (more differences).

[TABLE 4 ABOUT HERE]

Lee (2009) Bounds

Next, we discuss the results of the Lee (2009) bounding method. In Table 5, we report the estimated proportion of lottery winners to be trimmed following equation (2), representing the percent of lottery winners induced to enroll in PPS by winning the lottery and the proportion of observations trimmed from the upper and lower tails of the test score distribution to create the Lee (2009) bounds. Depending on grade level, the proportions ranged from 15.7% to 21.9%.

These results[16] are uninformative, because even with tightening, all the estimated bounds included zero. The covariates used to tighten the bounds were those that predicted enrollment into PPS (Ksoll et al., 2014), yet their predictive value is quite weak, as indicated in Table 2. Table 5 provides the bounds for the reading impacts, tightened using FRPL-eligibility and first language not English-status, all of which include zero.[17] A key issue here, is that we lack variables that highly predict enrollment in a district (Ksoll et al., 2014).

[TABLE 5 ABOUT HERE]

Angrist, Bettinger, & Kremer (2006) Non-Parametric Bounds

We next study the performance of Angrist et al.'s (2006) non-parametric bounding

[16] Lee bounds were estimated using the command "leebounds" in Stata (Tauchmann, 2014).
[17] We attempted to tighten bounds using every available combination of variables that were significantly predicting enrollment status as indicated in Table 2, but in no cases did the bounds exclude zero.

strategy. As mentioned previously, we do not have strong evidence that one of the key assumptions is met – that district leavers come from one side of the distribution – either the highest performing or the lowest performing. The third column in Table 6 reports the amount of bias in the naive PPS model. Under the assumption that those leaving the district, particularly from the control group, are relatively advantaged and potentially have higher test scores, the naïve PPS ITT estimates should be biased downward. While four out of six of the "naïve" PPS estimates in Table 6 were biased downward, two were biased upward, suggesting – as we expected earlier - that the assumption that those who leave the district are from the top end of the potential test score distribution is not fully supported.[18]

Table 6 also presents results using Angrist et al.'s (2006) non-parametric bounds, despite lacking a clear indication that the necessary assumptions have been met. This method produced bounds that contained the point estimate from the ODE benchmark sample in only two cases (grade 7 and grade 8 using $\theta = 0.9$). The 95% confidence intervals (CIs) of the bounds, however, overlap with the 95% CIs of the benchmark ODE point estimate in all cases, so we cannot conclude that the bounds do not include the benchmark estimate. In some cases, the upper bounds were less than the corresponding lower bounds (same percentile and grade comparison), continuing to raise concerns that the assumptions required for this method are not met.

[TABLE 6 ABOUT HERE]

[18] Note that the reverse assumption, that the students leaving the district are those in the bottom tail of the test score distribution, is also not fully supported by our data. The estimates in Table 2 do indicate that students with certain disadvantages, in some grades, leave PPS with higher probability. In this case, the estimated ITT unweighted PPS effects would be downward biased and this seems to be the case for one out of two of the significant estimated effects. To test whether this could be the case, we conducted the same non-parametric bounding approach under the assumption that those who leave the district are actually the *lowest* performers. We find that the results do not improve, and in some cases there are large (0.3 standard deviation) biases.

7. SIMULATION ANALYSIS OF PERFORMANCE OF ANGRIST ET AL. (2006) BOUNDING APPROACHES

As discussed previously, there is evidence that the required assumptions for the Angrist et al. (2006) bounding methods are not satisfied, so we wonder to what extent this harms performance. These types of methods are often used without a full assessment of whether the assumptions are met, so we use the simulation analysis to model the performance of these methods when such assumptions are and are not fully met.

To better understand the practicality and performance of these correction methods in various contexts, we simulated artificial assignment of treatment status and differential attrition under various assumptions. We created a simulation sample of 17,249 PPS students who were DLI non-applicants and who were present in PPS in third grade to ensure we had outcome data in this year and focused on estimating effects in grade three, where all attrition can be controlled by our simulation exercise. We assigned treatment status randomly (8,625 treatment, 8,624 control) such that the expected average treatment effect (ATE) is zero. Random assignment was conducted 100 times to create 100 samples, each with a different treatment and control group. Next, in each of these 100 datasets, we created artificial attrition of 5% and 10% of control units under various scenarios ranging from completely random attrition to attrition based solely on third grade test scores. The assumed attrition rate for treatment units is 0%, ensuring differential attrition rates of 5% and 10%.

In the case of completely random attrition, test scores are uncorrelated with predictors of attrition, and no selection bias should exist. Theoretically, the Angrist et al. (2006) non-parametric correction method works best under the case of attrition based solely on potential test scores. Under attrition that is a mix of test scores and random error, there is theoretically some

point at which too much random attrition causes the method to fail. Yet, if attrition is essentially random, it is a non-issue for bias. This simulation analysis seeks to find the situations under which Angrist et al. (2006) bounding methods are able to bound the true effect of zero.

We present results for the Angrist et al. (2006) non-parametric methods under 14 scenarios. For attrition amounts of 5 and 10 percent, we present results following seven types of control group attrition: attrition based solely on test scores, attrition that is completely random, and five cases of attrition driven by a mix of test scores and random error in the following ratios: 25/75, 40/60, 50/50, 60/40, and 75/25. To simulate attrition based on different ratios of test score-based and random attrition, we used z-scores (test scores normalized to a mean of zero and standard deviation of one), and a random variable with the standard normal distribution. We calculated the propensity to not be observed as the sum of these two variables in different proportions and use these to select the proportion that are unobserved. We replicated this exercise in 100 samples and assess how often we are able to estimate the expected ATE of zero.

Results of the non-parametric Angrist et al. (2006) approach under 5 percent simulated attrition are in Figure 3. The points represent the proportion of the bounds that included the true simulated ATE of zero for each combination of attrition type (test scores, random, or a mix), and model (naïve OLS or estimates of bounds at various quantiles). Recall, these quantiles are used to identify the upper and lower bounds, as described in section 3.1. Overall, this method works well, as long as attrition is primarily based on test scores.

[FIGURE 3 ABOUT HERE]

When attrition is driven entirely by test scores (the 5 percent of control group students with the highest test scores are not observed), the non-parametric bounds at various quantiles included the expected ATE of zero in the vast majority of cases (at least 97 out of 100). As long

as attrition is based at least 60 percent or more on test scores, at least 96 percent of the bounds created at the 90th percentile or below included the expected ATE of zero. "Naïve" OLS continues to work well when attrition is completely random (there is no bias to correct), but not if attrition is non-random. When attrition is completely random, the range of estimates using the non-parametric approach generally does include the expected ATE of zero, indicating, as expected, that this method does not perform better than OLS under completely random attrition.

Next, Figure 4 graphs the results using the non-parametric method (Angrist et al., 2006) to correct for simulated attrition of 10% of the control group in Figure 4. The results are similar to those in Figure 3, except when attrition is higher, the method appears to be more sensitive to the degree to which attrition is based on test scores.

[FIGURE 4 ABOUT HERE]

Overall, the results of the simulation analysis indicated that, when the assumptions of the Angrist et al. (2006) correction methods are met and attrition is primarily based on test scores, these methods are generally successful at correcting differential attrition. However, when attrition is random, we would have been better off using OLS without bounding. This supports our conclusion that the poor performance of these correction methods in the non-simulated data may be due to the underlying assumptions not being fully met.

8. CONCLUSION

This study provided a unique opportunity to test and compare the performance of various correction methods for differential attrition, a common practical issue in RCTs, against a benchmark estimate. The use of an augmented dataset enabled this study, but is often not practically attainable for researchers conducting lottery-based evaluations or RCTs more generally. One clear implication from this work is that similarly designed studies should try to

access an augmented dataset whenever possible. Using the observed (non-simulated) data, where the sources of attrition are largely unknown, the "naïve" PPS results were similar to the results from the augmented state-provided ODE dataset, where differential attrition was limited and which we use as the benchmark data source. Despite the apparent similarity between the "naïve" PPS and benchmark ODE results – which of course, would be unknowable without the benchmark dataset – we tested the performance of various correction methods against this benchmark, and concluded that the Angrist et al. (2006) non-parametric bounds and the Lee (2009) bounds had little success in a situation such as ours, particularly when it is difficult to directly test the extent to which underlying assumptions of correction methods are supported.

In the non-simulated data, IPW performed worse than unweighted PPS models. For five out of six estimates, the IPW results were further from the ODE benchmark results than the unweighted PPS results. Indeed, despite differential attrition rates that indicate strong potential for selection bias, it turned out that the "naïve" results may not have been extremely biased.

Perhaps this is not surprising, as IPW relies on propensity score models that accurately model the selection process. Luckily, in our case, we have an augmented dataset with which to compare the results, but in practice, many researchers may rely on IPW without enough observable characteristics to be confident that selection is appropriately modelled. Using the non-simulated data, the Angrist et al. (2006) non-parametric bounds contained the ODE benchmark point estimate in only two out of twelve cases, and in several cases, the upper bounds were lower than the corresponding lower bounds, suggesting issues of misspecification.

The bias and noise from using correction methods without evidence to support the underlying assumptions could lead to an incorrect conclusion about program effects. Thus, the main result of the non-simulation analysis is that using these various correction methods would

have not been the right choice in our case, and that – while researchers should carefully consider whether the assumptions are supported – this is often not possible in practice.

For two key reasons, the non-simulated results might not generalize to most program evaluations. First, despite differential attrition in the PPS sample, we had relatively balanced treatment and comparison groups in the "naïve" PPS sample (see Table 3), so there was not much bias to correct for. A second unique aspect is the lack of baseline test scores for applicants, who were applying prior to pre-K or kindergarten. Lack of important baseline covariates might make it harder to correctly model selection or to assess whether the assumptions for various methods are met. Thus, the results might be most relevant for evaluations that use administrative data to conduct ad-hoc analyses and those without a robust set of baseline covariates.

Given concerns of generalizability of these non-simulation results, the results of our simulation analysis have broader implications for researchers dealing with differential attrition. Due to uncertainty about whether the underlying assumptions were supported, we conducted a simulation analysis to test the performance of the Angrist et al. (2006) non-parametric methods under various types of attrition. Overall, we find these methods work quite well if the attrition is primarily based on student test scores. Unfortunately, it is often difficult to tell in practice to what extent this may be true.

We recommend researchers consider and test whenever possible the assumptions attrition correction methods area based on. The problem, however, is that researchers are generally not able to fully observe and model what drives attrition, particularly in cases where baseline outcome measures are unavailable. Thus, our results provide an important word of caution for researchers and for consumers of research using these types of methods.

References*

**One citation hidden to protect the integrity of blind review process.*

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. Retrieved 2/28/2018 from https://folio.iupui.edu/bitstream/handle/10244/726/InformingTheDebate_Final.pdf?sequence=2

Aker, J. C. & Ksoll, C. (2015). Call me educated: Evidence from a mobile monitoring experiment in Niger. Center for Global Development Working Paper 406. Retrieved 2/28/2018 from http://www.cgdev.org/publication/call-me-educated-evidence-mobile-monitoring-experiment-niger-working-paper-406.

Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security administrative records. *The American Economic Review, 80*(3), 313-336.

Angrist, J., Bettinger, E., Bloom E., Kremer, M, and King, E. (2002). The effects of school vouchers on students: Evidence from Colombia. *American Economic Review*, 92(5), 1535-1558.

Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review*, 96(3), 847-862.

Aron-Dine, A., Einav, L., & Finkelstein, A. (2013). The RAND health insurance experiment, three decades later. *Journal of Economic Perspectives, 27*(1), 197-222.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine, 34*(28), 3661-3679.

Bailey, M. A., Hopkins, D.J., & Rogers, T. (2016). Unresponsive, unpersuaded: The unintended consequences of voter persuasion efforts. *Political Behavior, 38*(3), 713-746.

Barnow, B. S., & Greenberg, D. (2015). Do estimated impacts on earnings depend on the source of the data used to measure them? Evidence from previous social experiments. *Evaluation Review*, *39*(2), 179-228.

Barrow, L., Richburg-Hayes, L., Rouse, C. E., & Brock, T. (2014). Paying for performance: The education impacts of a community college scholarship program for low-income adults. *Journal of Labor Economics*, *32*(2), 563-599.

Behaghel, L. Crépon, B., Gurgand, M., & Le Barbanchon, T. (2015). Please call again: Correcting nonresponse bias in treatment effect models. *The Review of Economics and Statistics, 97*(5), 1070-1080.

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, 729-751.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). Scaling up what works: Experimental evidence on external validity in Kenyan Education. Center for Global Development Working Paper 321. Retrieved 2/28/2018 from: http://www.cgdev.org/publication/scaling-what-works-experimental-evidence-external-validity-kenyan-education-working

Boo, F. L., Palloni, G., & Urzua, S. (2014). Cost-benefit analysis of a micronutrient supplementation and early childhood stimulation program in Nicaragua. *Annals of the New York Academy of Sciences, 1308,* 139-148.

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.

Clark, M., Rothstein, J., & Schanzenbach, D. W. (2009). Selection bias in college admissions test scores. *Economics of Education Review, 28,* 295-307.

Cook, T. D. (2008). "Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics, 142,* 636-654.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research, 44*(6), 828-847.

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica,* 74(5), 1191-1230.

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.

Deming, D. J., Hastings, J. S., Kane, T. J. & Staiger, D. O. (2014). School choice, school quality, and postsecondary attainment. *The American Economic Review, 104*(3), 991-1023.

DiNardo, J., McCrary, J., & Sanbonmatsu, L. (2006). Constructive proposals for dealing with attrition: An empirical example. Retrieved 2/28/2018 from: http://eml.berkeley.edu/~jmccrary/DMS_v9.pdf

Dobbie, W. & Fryer, R. G. (2009). Are high quality schools enough to close the achievement gap? Evidence from a social experiment in Harlem. No. w15473. *National Bureau of Economic Research.* Retrieved 2/28/2018 from: http://www.nber.org/papers/w15473.pdf

Engberg, J., Epple, D., Imbrogno, J., Sieg, H., & Zimmer, R. (2014). Evaluating education programs that have lotteried admission and selective attrition. *Journal of Labor*

Economics, 32(1), 27–63.

Frölich, M. & Huber, M. (2014). Treatment evaluation with multiple outcome periods under endogeneity and attrition. *Journal of the American Statistical Association, 109*(508), 1697-1711.

Garlick, R. & Hyman, J. (2016). *Data vs. methods: quasi-experimental evaluation of alternative sample selection corrections for missing college entrance exam score data.* ERID Working Paper Number 221. Economic Research Initiatives at Duke, Duke University. Retrieved 2/28/2018 from https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/13249/SSRN-id2793486.pdf?sequence=1

Glewwe, P., Illias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics, 2,* 205-227.

Goldberger, A. S. (1972). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations.* (Discussion Paper No. 123). Madison, WI: Institute for Research on Poverty, University of Wisconsin – Madison. Retrieved 2/28/2018 from: http://www.irp.wisc.edu/publications/dps/pdfs/dp12372.pdf

Greenberg, D., & Barnow, B. S. (2014). Flaws in evaluations of social programs: Illustrations from randomized controlled trials. *Evaluation review, 38*(5), 359-387.

Grilli, L. & Mealli, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *Journal of Educational and Behavioral Statistics, 33.*

Hastings, J. S., Neilson, C. A., & Zimmerman, S. D. (2012). *The effect of school choice on intrinsic motivation and academic outcomes.* NBER Working Paper 18324. Retrieved 2/28/2018 from: http://www.nber.org/papers/w18324

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153-162.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*(5), 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.

Hogan, J. W., & Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research, 13*(1), 17-48.

Holm, A. & Jaeger, M. M. (2009). Selection bias in educational transition models: Theory and empirical evidence. University of Copenhagen Department of Economics Working Paper No. 2009-05. Retrieved 2/28/2018 from https://core.ac.uk/download/pdf/6476769.pdf

Horowitz, J. L. & Manski, C. F. (1998). Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations. *Journal of Econometrics,* 84, 37-58.

Horowitz, J. L. & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95*(449), 77–84.

Hoxby, C. M., & Rockoff, J. (2004). The impact of charter schools on student achievement. Unpublished manuscript. Retrieved 2/28/2014 from: https://www0.gsb.columbia.edu/faculty/jrockoff/hoxbyrockoffcharters.pdf

Huber, M. & Mellace, G. (2013). Sharp bounds on causal effects under sample selection. *University of St. Gallen, Dept. of Economics.* Retrieved 2/28/2018 from: https://www.alexandria.unisg.ch/70307/1/sample_selection_bounds_incl_appendix.pdf

Imbens, G. W., & Manski, C. F. (2004): Confidence intervals for partially identified parameters. *Econometrica, 72*, 1845-1857.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5-86.

Karlan, D., Fairlie, R. W., & Zinman, J. (2012). Behind the GATE experiment: Evidence on effects of and rationales for subsidized entrepreneurship training. Yale University Department of Economics Working Paper No. 95. Retrieved 2/28/2018 from https://www.dartmouth.edu/~jzinman/Papers/GATE_2012_11.pdf

Kremer, M., Miguel, E. & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics, 91*(3), 437-456.

Ksoll, C., Aker, J. Miller, D., Perez-Mendoza, K. C., & Smalley, S. L. (2014). Learning without teachers? A randomized experiment of a mobile phone-based adult education program in Los Angeles. (Working Paper 368). Center for Global Development. Retrieved 2/28/2018 from https://www.cgdev.org/publication/learning-without-teachers-randomized-experiment-mobile-phone-based-adult-education

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review, 76*(4), 604-620.

Lechner, M. & Melly, B. (2010). Partial identification of wage effects of training programs. Brown University Department of Economics Working Paper. Retrieved 2/28/2018 from: https://www.alexandria.unisg.ch/38673/4/Bounds.pdf

Lee, D.S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies,* 76, 1071-1102.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings, 80,* 319–23.

Manski, C. F. (1995). *Identification problems in the social sciences.* Cambridge, MA: Harvard University Press.

Melenberg, B. & van Soest, A. (1996). Parametric and semi-parametric modelling of vacation expenditures. *Journal of Applied Econometrics, 11*(1), 59-76.

Molina, T. & Macours, K. (2015). Attrition in randomized control trials: Regular versus intense tracking protocols. Retrieved 2/28/2018 from http://lacer.lacea.org/bitstream/handle/123456789/52356/lacea2015_attrition_randomized_control_trials.pdf?sequence=1

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica, 55*(4), 75-799.

Muralidharan, K. & Sundararaman, V. (2014). The aggregate effect of school choice: Evidence from a two-stage experiment in India. NBER Working Paper 19441. Retrieved 2/28/2018 from http://www.nber.org/papers/w19441

Newey, W. K., Powell, J. L., & Walker, J. R. (1990). Semiparametric estimation of selection models: Some empirical results. *The American Economic Review, 80*(2), 324-328.

Reynolds, A. J., Temple, J. A., Ou, S, Arteaga, I. A., & White, B. A. B. (2011). School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups. *Science, 333,* 360-364.

Robins, P. K., & West, R. W. (1986). Sample attrition and labor supply response in experimental panel data: A study of alternative correction procedures. *Journal of Business & Economic Statistics*, *4*(3), 329-338.

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.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin Company.

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*(3), 250-267.

Steiner, P. M. & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation Review, 42*(2), 214-247.

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal Regression Models II: Unconfoundedness and Causal Unbiasedness. *Methods of Psychological Research Online*

2000, 5(3). Retrieved 2/28/2018 from http://136.199.86.12/fachgruppen/methoden/mpr-online/issue11/art4/steyerCRII_corrected.pdf

Tauchmann, H. (2014). Lee's treatment effect bounds for non-random sample selection – an implementation in Stata. *The Stata Journal*, 14(4), 884-894.

Tryon, W. W. (2001). Evaluation statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis tests. *Psychological Methods*, 6, 371-386.

What Works Clearinghouse. (2013). Assessing Attrition Bias V 2.1. Washington, DC: What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. Retrieved 1/16/2020 from https://ies.ed.gov/ncee/wwc/Document/243

What Works Clearinghouse. (2014). *Procedures and Standards Handbook: Version 3.0.* Institute of Education Sciences. Washington, DC: U.S. Department of Education. Retrieved 2/28/2018 from: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, *141*(2), 1281-1301.

Zhang, J. L. & Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics, 28*(4), 353-368.

Zhang, J. L., Rubin, D. B., & Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. In Fomby, T., Hill, C.R., Millimet, D.L, Smith, J., & Vytlacil, E.J. (Eds.), *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics* (117-145). Emerald Group Publishing Limited.

TABLE 1. Randomized Sample by Lottery-Assigned Status and Share Observed in Benchmark and PPS Reading Analysis Samples.

Valid Lottery Applicants Randomized by Cohort			Size of Randomized Sample by Highest Observable Grade (if Zero Attrition)			Fraction Observed in Benchmark ODE Analysis Sample		Fraction Observed in PPS Analysis Sample	
Fall Term	Won	Lost	Grade	Won	Lost	Won	Lost	Won	Lost
2004	63	133							
2005	74	111	8	137	244	0.67**	0.68**	0.53	0.42
2006	103	148	7	240	392	0.63**	0.65**	0.52	0.44
2007	125	153	6	365	545	0.63*	0.60*	0.56	0.44
2008	166	171	5	531	716	0.70*	0.66*	0.65	0.50
2009	148	168	4	679	884	0.74*	0.68*	0.68	0.52
2010	185	198	3	864	1,082	0.75*	0.69*	0.71	0.55
Total	864	1,082							

**Meets What Works Clearinghouse Version 3.0 Conservative Attrition Thresholds (WWC, 2014)

*Meets What Works Clearinghouse Version 3.0 Liberal Attrition Thresholds (WWC, 2014)

TABLE 2. Propensity to Enroll in PPS (Full Sample); Dependent Variable: Has Reading Test Score and Enrolls in Portland (Marginal Effects)

	Full Sample		Treatment Group Only		Control Group Only		Diff. in Coefficients (T-C)	
	K	3rd	K	3rd	K	3rd	K	3rd
Won Lottery	0.0685 ***	0.134 ***						
	(0.0160)	(0.0220)						
Female	-0.0125	0.0090	0.0050	0.0139	-0.0328	0.0032	0.0378	0.0107
	(0.0161)	(0.0225)	(0.0203)	(0.0293)	(0.0258)	(0.0325)	(0.0328)	(0.0438)
Asian	-0.0443	0.0213	-0.0061	0.00581	-0.109 *	0.0214	0.1030	-0.0156
	(0.0311)	(0.0372)	(0.0326)	(0.0453)	(0.0581)	(0.0593)	(0.0666)	(0.0746)
Black	-0.0149	-0.1010 *	-0.0379	-0.248 ***	0.0122	0.0017	-0.0501	-0.2497 **
	(0.0424)	(0.0569)	(0.0612)	(0.0919)	(0.0628)	(0.0732)	(0.0877)	(0.1175)
Hispanic	0.00798	-0.00529	0.0243	-0.0553	-0.0054	0.0309	0.0297	-0.0862
	(0.0268)	(0.0378)	(0.0324)	(0.0561)	(0.0432)	(0.0518)	(0.0540)	(0.0764)
Other Race	-0.0427	-0.0490	-0.0047	-0.0197	-0.0805	-0.0645	0.0758	0.0448
	(0.0370)	(0.0499)	(0.0444)	(0.0704)	(0.0582)	(0.0675)	(0.0732)	(0.0975)
Missing Race	-0.420 ***	-0.457 ***	-0.432 ***	-0.580 ***	-0.438 ***	-0.401 ***	0.0060	-0.1790
	(0.0862)	(0.0803)	(0.162)	(0.136)	(0.0995)	(0.0942)	(0.1901)	(0.1654)
FRPL	0.0497 **	0.0395	0.0132	0.0643 *	0.0957 ***	0.0234	-0.0825 **	0.0409
	(0.0203)	(0.0305)	(0.0279)	(0.0381)	(0.0310)	(0.0451)	(0.0417)	(0.0590)
Special Needs (t=0)	0.1100 ***	0.0145	0.0688 ***	-0.0745		0.149 **	N/A	-0.2235 **
	(0.0177)	(0.0549)	(0.0244)	(0.0729)		(0.0753)	N/A	(0.1048)
First Language Not Englisl	-0.0892 **	-0.153 ***	-0.0839 *	-0.139 **	-0.0954 *	-0.151 **	0.0115	0.0120
	(0.0364)	(0.0448)	(0.0488)	(0.0603)	(0.0570)	(0.0639)	(0.0750)	(0.0879)
Observations	1,625	1,581	752	721	845	860		

Note. There were no test score outcomes in kindergarten, so the kindergarten outcome is simply enrollment in PPS. Models predicting observations in later years include lagged test scores as additional explanatory variables; results available by request. Some control group units were dropped from the kindergarten outcome model due to special needs status perfectly predicting enrollment in the district in kindergarten.

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

TABLE 3. Covariate balance

Panel A: Kindergarten	Benchmark ODE Sample				
	Binding Lottery Applicants Only				
	Won Slot	Not Placed	Difference (Unadj)	Strata adj. Difference	Strata adj. p-value
N	752	873			
Proportion	0.463	0.537			
Female	0.508	0.546	-0.038	-0.041	0.15
Asian	0.178	0.115	0.064	0.008	0.61
Black	0.052	0.060	-0.008	0.004	0.77
Hispanic	0.177	0.164	0.013	0.008	0.65
White	0.517	0.559	-0.042	0.028	0.25
Other Race	0.063	0.073	-0.011 **	-0.036 **	0.01
FRPL	0.273	0.250	0.023	-0.009	0.63
Special Needs in K	0.052	0.032	0.020	0.012	0.29
ELL in K	0.153	0.105	0.048	-0.002	0.91
First Language Not English	0.206	0.157	0.049	-0.013	0.42

Panel B: Grade 3	Benchmark ODE Sample					Initial PPS Sample - Unweighted					Inverse Probability Weighted Sample				
	Binding Lottery Applicants with ODE-Observed 3rd Grade Reading Score					Binding Lottery Applicants with PPS-Observed 3rd Grade Reading Score					Binding Lottery Applicants with PPS-Observed 3rd Grade Reading Score				
	Won Slot	Not Placed	Difference (Unadj)	Strata adj. Difference	Strata adj. p-value	Won Slot	Not Placed	Difference (Unadj)	Strata adj. Difference	Strata adj. p-value	Won Slot	Not Placed	Difference (Unadj)	Strata adj. Difference	Strata adj. p-value
N	642	745				583	581				583	581			
Proportion	0.463	0.537				0.501	0.499				0.501	0.499			
Female	0.511	0.566	-0.056 *	-0.059 *	0.06	0.513	0.552	-0.040	-0.040	0.23	0.504	0.552	-0.048	-0.036	0.29
Asian	0.192	0.121	0.071	0.011	0.53	0.185	0.110	0.075	0.006	0.76	0.182	0.115	0.067	0.004	0.84
Black	0.051	0.062	-0.010	-0.001	0.95	0.045	0.064	-0.019	-0.013	0.36	0.055	0.060	-0.006	-0.003	0.86
Hispanic	0.174	0.172	0.003	-0.004	0.84	0.170	0.165	0.005	0.004	0.84	0.186	0.163	0.023	0.015	0.45
White	0.519	0.553	-0.034	0.037	0.17	0.539	0.585	-0.046	0.035	0.22	0.509	0.563	-0.054	0.028	0.31
Other Race	0.058	0.071	-0.014 **	-0.034 **	0.02	0.057	0.064	-0.007 **	-0.030 **	0.06	0.055	0.068	-0.013 **	-0.037 **	0.04
FRPL	0.293	0.262	0.031	-0.008	0.71	0.271	0.250	0.021	-0.016	0.48	0.285	0.253	0.032	-0.012	0.64
Special Needs in K	0.044	0.035	0.009	-0.0001	0.99	0.046	0.040	0.007	-0.001	0.92	0.053	0.033	0.021	0.010	0.44
ELL in K	0.160	0.115	0.045	-0.011	0.48	0.142	0.100	0.042	-0.008	0.63	0.157	0.111	0.046	-0.009	0.61
First Language Not English	0.215	0.161	0.054	-0.012	0.49	0.194	0.139	0.055	-0.007	0.70	0.217	0.161	0.056	-0.013	0.52

Note. The "strata-adj" differences and p-values reflect the differences/balance within lottery strata.

TABLE 4. Comparison of Reading Results: Inverse Probability Weighting.

				Steiner & Wong (2018) Correspondence at the 95% Confidence Level; (Standard Errors of Estimated Bias in Parentheses)							
	Benchmark ODE Sample	Naïve PPS	Bias in Naïve PPS	Assuming Equivalence Threshold of 0.05 s.d.				Assuming Equivalence Threshold of 0.1 s.d.			
				Corr = 0	Corr = 0.5	0.75	Corr = 1	Corr = 0	Corr = 0.5	0.75	Corr = 1
Grade 3 ITT	0.059	0.077	0.019	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Equivalence	Trivial Diff.
	(0.051)	(0.055)		(0.075)	(0.053)	(0.038)	(0.004)	(0.075)	(0.053)	(0.038)	(0.004)
Grade 4 ITT	0.078	0.065	-0.013	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Equivalence	Trivial Diff.
	(0.056)	(0.062)		(0.084)	(0.059)	(0.042)	(0.005)	(0.084)	(0.059)	(0.042)	(0.005)
Grade 5 ITT	0.150 **	0.123 *	-0.027	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Indet.	Trivial Diff.
	(0.060)	(0.066)		(0.089)	(0.063)	(0.045)	(0.006)	(0.089)	(0.063)	(0.045)	(0.006)
Grade 6 ITT	0.120	0.119	-0.001	Indet.	Indet.	Indet.	Equivalence	Indet.	Indet.	Equivalence	Equivalence
	(0.075)	(0.082)		(0.111)	(0.078)	(0.056)	(0.007)	(0.111)	(0.078)	(0.056)	(0.007)
Grade 7 ITT	0.117	0.091	-0.026	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Indet.	Trivial Diff.
	(0.081)	(0.094)		(0.124)	(0.088)	(0.063)	(0.013)	(0.124)	(0.088)	(0.063)	(0.013)
Grade 8 ITT	0.232 **	0.313 ***	0.081	Indet.	Indet.	Indet.	Difference	Indet.	Indet.	Indet.	Difference
	(0.101)	(0.118)		(0.155)	(0.110)	(0.079)	(0.017)	(0.155)	(0.110)	(0.079)	(0.017)
Obs.	4,594	3,705									
Students	1,447	1,208									
Adj. R-squared	0.3112	0.3098									

				Steiner & Wong (2018) Correspondence at the 95% Confidence Level; (Standard Errors of Estimated Bias in Parentheses)							
	Benchmark ODE Sample	IPW PPS	Bias in IPW PPS	Assuming Equivalence Threshold of 0.05 s.d.				Assuming Equivalence Threshold of 0.1 s.d.			
				Corr = 0	Corr = 0.5	0.75	Corr = 1	Corr = 0	Corr = 0.5	0.75	Corr = 1
Grade 3 ITT	0.059	0.089	0.030	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Equivalence	Trivial Diff.
	(0.051)	(0.055)		(0.075)	(0.053)	(0.037)	(0.004)	(0.075)	(0.053)	(0.037)	(0.004)
Grade 4 ITT	0.078	0.097	0.019	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Equivalence	Trivial Diff.
	(0.056)	(0.062)		(0.084)	(0.059)	(0.042)	(0.005)	(0.084)	(0.059)	(0.042)	(0.005)
Grade 5 ITT	0.150 **	0.126 *	-0.024	Indet.	Indet.	Indet.	Trivial Diff.	Indet.	Indet.	Equivalence	Trivial Diff.
	(0.060)	(0.065)		(0.089)	(0.063)	(0.045)	(0.005)	(0.089)	(0.063)	(0.045)	(0.005)
Grade 6 ITT	0.120	0.177 **	0.057	Indet.	Indet.	Indet.	Difference	Indet.	Indet.	Indet.	Trivial Diff.
	(0.075)	(0.083)		(0.112)	(0.079)	(0.056)	(0.008)	(0.112)	(0.079)	(0.056)	(0.008)
Grade 7 ITT	0.117	0.147	0.030	Indet.	Indet.	Indet.	Difference	Indet.	Indet.	Indet.	Trivial Diff.
	(0.081)	(0.094)		(0.124)	(0.088)	(0.063)	(0.013)	(0.124)	(0.088)	(0.063)	(0.013)
Grade 8 ITT	0.232 **	0.353 ***	0.121	Indet.	Indet.	Indet.	Difference	Indet.	Indet.	Indet.	Difference
	(0.101)	(0.116)		(0.154)	(0.109)	(0.078)	(0.015)	(0.154)	(0.109)	(0.078)	(0.015)
Grade 9 ITT	0.0917	-0.286	-0.378								
	(0.292)	(0.238)									
Obs.	4,594	3,660									
Students	1,447	1,204									
Adj. R-squared	0.3112	0.3198									

Note. Robust standard errors in parentheses. All models include year fixed effects, lottery strata fixed effects, and demographic controls.. IPW = Inverse Probability Weighted. The standard error estimates included in the columns with the results of the difference and equivalence tests (Steiner & Wong, 2018), are the standard errors of the estimated bias, assuming correlations between the effect estimates of 0, 0.5, and 1. Grade 9 results hidden from table due to small sample size (<50 observations). ODE = Oregon Department of Education. PPS = Portland Public Schools. Naïve PPS models are calculated based on equation (1), among only the PPS sample, without inverse probability weighting.

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

TABLE 5. Lee Bounds Analysis: Proportion to be Trimmed, Upper and Lower Bounds on Reading Treatment Effects.

	Proportion to be Trimmed	Lower Bound	Upper Bound
Grade 3	16.5%	-0.26	0.30
Grade 4	17.2%	-0.35	0.24
Grade 5	20.1%	-0.30	0.34
Grade 6	15.7%	-0.27	0.22
Grade 7	19.2%	-0.35	0.17
Grade 8	21.9%	-0.14	0.50

Note. Covariates used for tightening include FRPL-eligibility and first language not English.

TABLE 6. Comparison of Reading Results: Non-Parametric Bounds (Angrist et al., 2006).

				Bounds Using θ = .95				Bounds Using θ = .90			
	Benchmark ODE Sample	Naive PPS	Bias in Naive PPS	95% Lower Bound	95% Upper Bound	Benchmark Estimate within	Confidence Intervals Overlap?	90% Lower Bound	90% Upper Bound	Benchmark Estimate within	Confidence Intervals Overlap?
Grade 3 ITT	0.0585	0.0774	0.019	0.0317	0.0499	No	Yes	0.0067	0.0373	No	Yes
	(0.051)	(0.055)		(0.049)	(0.050)			(0.048)	(0.048)		
Grade 4 ITT	0.0779	0.0648	-0.013	-0.0173	0.0248	No	Yes	-0.0405	-0.0203	No	Yes
	(0.056)	(0.062)		(0.055)	(0.056)			(0.054)	(0.054)		
Grade 5 ITT	0.150 **	0.123 *	-0.027	0.0880	0.100	No	Yes	0.0616	0.0798	No	Yes
	(0.060)	(0.066)		(0.060)	(0.061)			(0.059)	(0.059)		
Grade 6 ITT	0.120	0.119	-0.001	0.0850	0.0780	No	Yes	0.0618	0.0430	No	Yes
	(0.075)	(0.082)		(0.074)	(0.074)			(0.072)	(0.071)		
Grade 7 ITT	0.117	0.0909	-0.026	0.144	0.123	No	Yes	0.133	0.108	Yes	Yes
	(0.081)	(0.094)		(0.091)	(0.091)			(0.089)	(0.088)		
Grade 8 ITT	0.232 **	0.313 ***	0.081	0.310 ***	0.344 ***	No	Yes	0.228 *	0.318 ***	Yes	Yes
	(0.101)	(0.118)		(0.117)	(0.117)			(0.119)	(0.117)		
Observations	4,594	3,705		3,510	3,470			3,266	3,283		
Students	1,447	1,208		1,187	1,161			1,124	1,128		
Adj. R-Squared	0.311	0.310		0.315	0.316			0.304	0.305		

Note. Robust standard errors in parentheses. Other covariates include year indicators, binding lottery strata fixed effects, and demographic controls (gender, race, special needs in kindergarten, first language not English in kindergarten, and FRPL in kindergarten). Grade 9 results hidden from table due to small sample size (<50 observations). Whether or not confidence intervals overlap is based on the 95% CI of the benchmark ODE sample, and the 95% confidence intervals of the upper and lower bounds. Naïve PPS models are calculated based on equation (1), among only the PPS sample, without bounding.
*** p<0.01, ** p<0.05, * p<0.1.

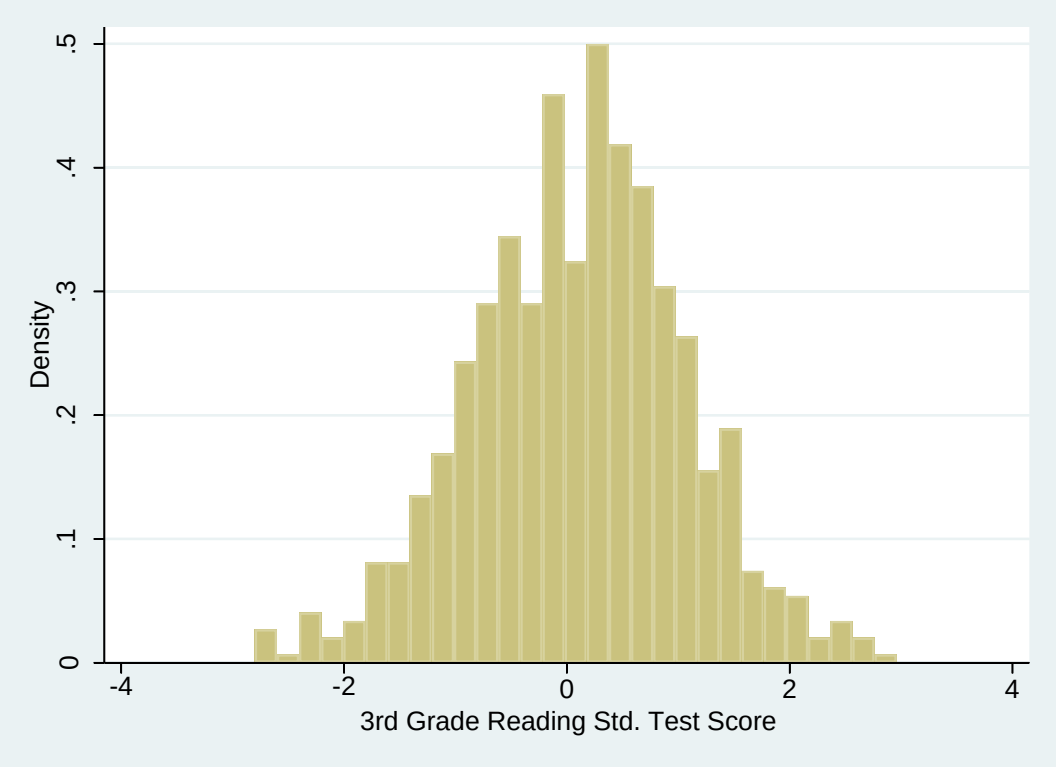

FIGURE 1. *Distribution of third grade reading scores, all control group students observed in augmented state-level dataset, including PPS-enrollees and non-enrollees*

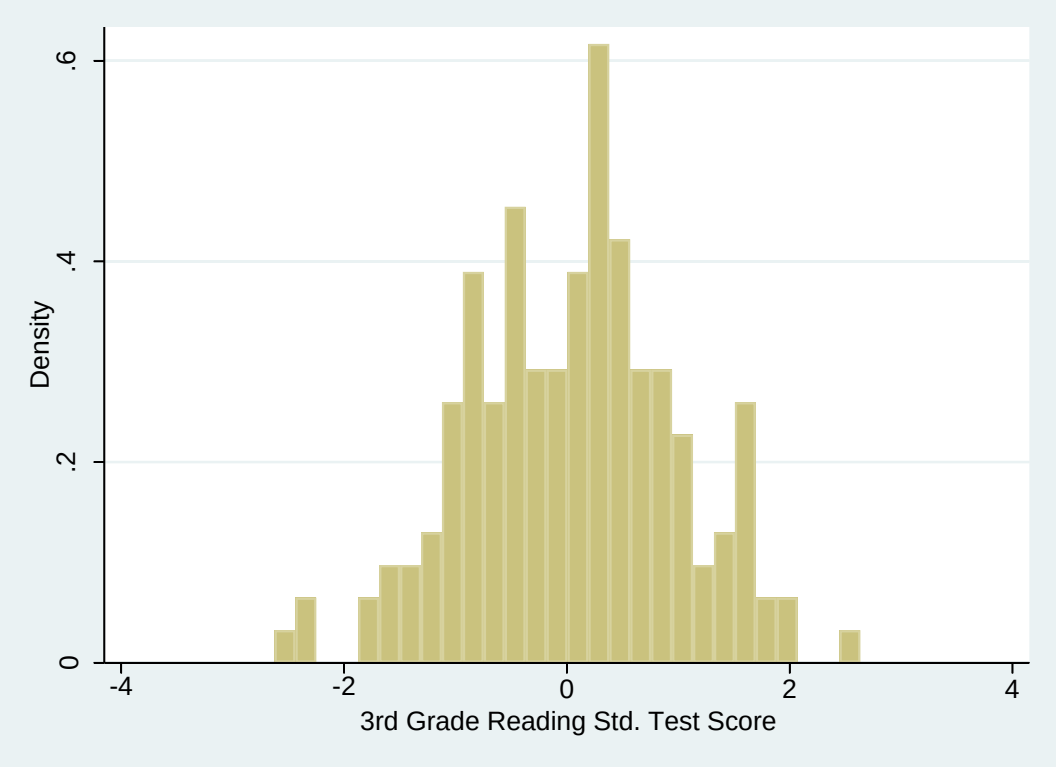

FIGURE 2. *Distribution of third grade reading scores, control group students who did not enroll in PPS district*

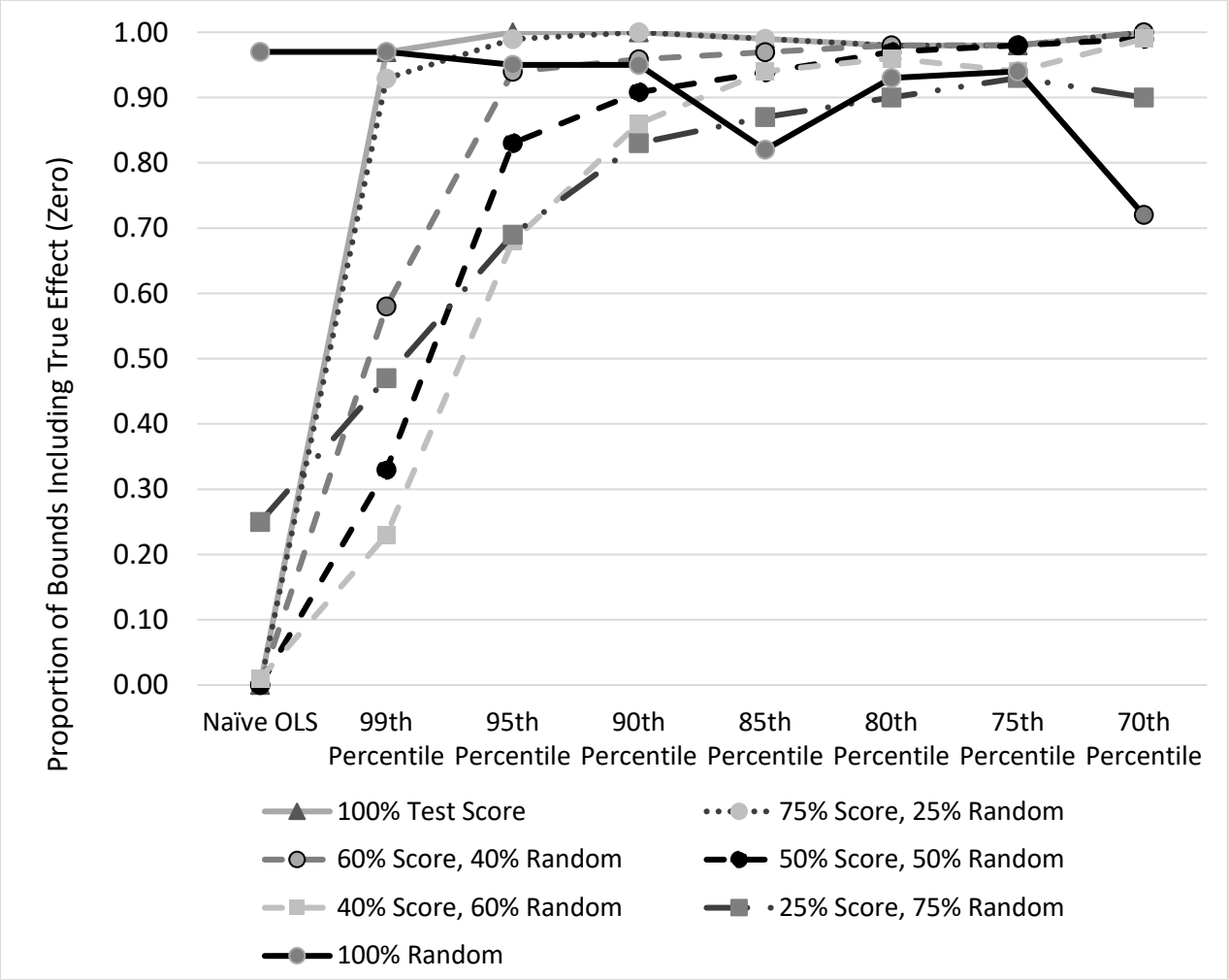

FIGURE 3. *Non-parametric bounding results at various percentiles, under artificially simulated attrition of 5% (grade 3 reading). Tabular summaries of these results are available by request.*

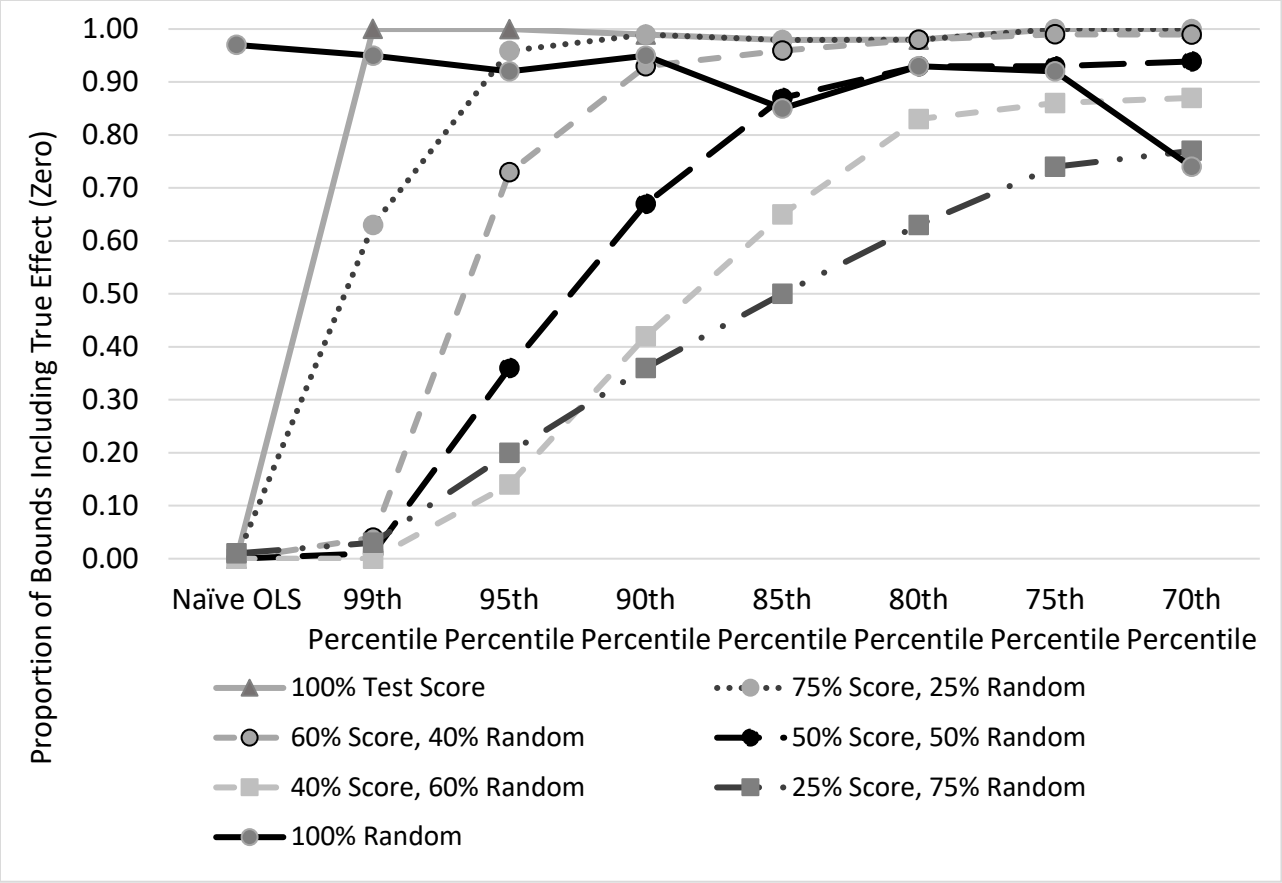

FIGURE 4. *Non-parametric bounding results at various percentiles, under artificially simulated attrition of 10% (grade 3 reading). Tabular summaries of these results are available by request.*