9-1-2016

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

Gema Zamarro
University of Arkansas, Fayetteville, gzamarro@uark.edu

Kaitlin Anderson
Michigan State University

Jennifer L. Steele
University of Arkansas, Fayetteville

Trey Miller
RAND Corporation

UNIVERSITY OF ARKANSAS | College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

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

Gema Zamarro, Kaitlin Anderson,

Jennifer Steele, and Trey Miller

September 2016

EDRE Working Paper 2016-15

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

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

Kaitlin Anderson (kaitlina@uark.edu), The University of Arkansas

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

Trey Miller, PhD (Trey_Miller@rand.org), RAND Corporation

ABSTRACT:

In randomized controlled trials, it is common for attrition rates to differ by lottery status, jeopardizing the identification of causal effects. Inverse probability weighting methods (Hirano et al, 2003; Busso et al., 2014) and estimation of informative bounds for the treatment effects (e.g. Lee, 2009; Angrist et al., 2006) have been used frequently to deal with differential attrition bias. This paper studies the performance of various methods by comparing the results using two datasets: a district-sourced dataset subject to considerable differential attrition, and an expanded state-sourced dataset with much less attrition, differential and overall. We compared the performance of different methods to correct for differential attrition in the district dataset, as well as we conducted simulation analyses to assess the sensitivity of bounding methods to their underlying assumptions. In our application, methods to correct differential attrition induced bias, whereas the unadjusted district level results were closer and more substantively similar to the estimated effects in the benchmark state dataset. Our simulation exercises showed that even small deviations from the underlying assumptions in bounding methods proposed by Angrist et al. (2006) increased bias in the estimates. In practice, researchers often do not have enough information to verify the extent to which these underlying assumptions are met, so we recommend using these methods with caution.

KEYWORDS: differential attrition, bounding methods, simulation

JEL Codes: C18, C15, C90

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through State and Local Policy Programs and Systems grant # R305E120003 to the RAND Corporation, the American Councils for International Education, and the Portland Public Schools. This effort has benefitted enormously from ongoing research collaboration with Deborah Armendariz, Director of Dual Language in the Portland Public Schools, and from ongoing feedback from Allen Ruby at the Institute of Education Sciences. It would not have been possible without outstanding data assistance by Joseph Suggs, Karin Brown, and Jennifer Miller in the Portland Public Schools and Jonathan Wiens in the Oregon Department of Education. The opinions expressed are those of the authors and do not represent views of IES, the U.S. Department of Education, or the research partner organizations.

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 arise at causal effects of educational programs (see, e.g. Rouse (1998); Angrist, Bettinger, Bloom, Kremer, & King (2002); Hoxby & Rockoff (2004); Cullen, Jacob, & Levitt (2006); Abdulkadiroglu et al. (2009); Hoxby & Murarka (2009); Dobbie & Fryer (2009), Deming, Hastings, Kane, & Staiger (2014); Engberg, Epple, Imbrogno, Sieg, & Zimmer (2014) among others).

In education, it is common for school districts around the country to use lotteries to determine access to oversubscribed educational programs. Then, those winning the lottery have the possibility of enrolling in the specific program while those non-placed would not have the option to participate but would have multiple other outside options. By comparing average outcomes of lottery winners with average outcomes of those non-placed, the hope is to arise at causal effects not affected by bias due to selection into the program.

However, it is not uncommon that students who are not placed by the lottery seek alternative options outside the district, e.g. by choosing a charter school, choosing a private school, or moving to a different school district instead. For those who leave the school district, it is uncommon to have data of those students and this creates a missing data problem. In particular, if attrition rates differ considerably depending on the lottery status, this creates a differential attrition bias problem, jeopardizing the identification of causal effects trough the randomization induced by the lottery. This differential attrition problem is quite common. In a review of development economics studies published between 2009 and the first quarter of 2015, Molina & Macours (2015) find that of the 68 RCTs, about 19 percent had differential attrition,

and in many cases, authors simply restrict the analysis to a subsample in which attrition was balanced. (Molina & Macours, 2015).

Removing all the bias induced by selective attrition would be possible if either all covariates determining the outcome are known (Steyer, Gabler, von Davier, & Nachtigall, 2000); or if the selection process is completely known (Cook, 2008; Goldberger, 1972; Shadish, Cook, & Campbell, 2002). Often, however, researchers are not fully able to directly observe these covariates or accurately model the selection process, and selection bias due to attrition continues to be a potential issue. A common approach among researchers in education, as well as other fields, aimed at minimizing the bias due to differential attrition or differential nonresponse, has been the use of inverse probability weighting[1]. In this case, observations in the treatment and control group are reweighted to remain comparable in important observed characteristics. The success of this method, however, relies on the availability of enough information to control for the key differences between treatment and controls induced from attrition bias. Alternatively, researchers have used various bounding methods to arrive at estimates of the possible range of estimated effects under different attrition scenarios. Molina and Macours (2015) find that bounding methods were used in almost 15 percent of the 68 studies included in their review.

In our review of the literature on the use of bounding methods, focusing primarily on studies in the field of education, we find that the most popular bounding method used is that proposed by Lee (2009).[2] Another approach sometimes used, similar to Lee (2009), is Manski's worst-case scenario bounds (Manski, 1990; Manski, 1995; Horowitz & Manski, 1998; Horowitz

[1] See for example: Imbens & Wooldridge, 2009; Reynold, Temple, Ou, Arteaga, & White, 2011; Bailey, Hopkins, & Rogers, 2016; Muralidharan & Sundararaman, 2013; Frölich & Huber, 2014; Molina & Macours, 2015.
[2] See for example: DiNardo, McCrary, & Sanbonmatsu, 2006; Kremer, Miguel, & Thornton, 2009; Glewwe, Illias, & Kremer, 2010; Hastings, Neilson, & Zimmerman, 2012; Karlan, Fairlie, & Zinman, 2012; Aron-Dine, Einav, & Finkelstein, 2013; Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2013; Muralidharan & Sundararaman, 2013; Boo, Palloni, & Urzua, 2014; Engberg et al., 2014; Molina & Macours, 2015; Aker & Ksoll, 2015.

& Manski, 2000, Imbens & Manski, 2004)[3]. Both these bounding methods obtain bounds for extreme case scenarios under relatively weak assumptions on the type of respondents that are attriting. As a result of their weak assumptions on the attrition process, these methods tend to provide wide bounds that in occasions result uninformative.

Still other researchers have proposed bounding the estimates based on certain assumptions about the type of respondents attriting from the sample. These include the parametric and non-parametric bounding approaches proposed by Angrist et al. (2006), used for example in Barrow, Richburg-Hayes, Rouse, and Brock (2014), as well as other extensions and modifications of these bounding methods (e.g., Huber & Mellace, 2013; Engberg et al., 2014; Zhang and Rubin, 2003; Grilli and Mealli, 2008; Zhang, Rubin, and Mealli, 2008; and Lechner and Melly, 2010).

With the exception of the bounding method proposed by Lee (2009) and Manski's worst-case scenario bounds, all these other alternative bounding approaches make restrictive assumptions about the type of respondents that are attriting from the sample. For example, Angrist et al. (2006)'s bounding approach assumes that those attriting come from only one side of the outcome distribution. Similarly, Huber and Mellace (2013) derive bounds under the assumption of stochastic dominance. The authors describe stochastic dominance as the assumption 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" (p. 17). This assumption, which is not imposed by Lee (2009), has also been imposed in other bounding approaches like for example, Grilli and Mealli (2008) and Lechner and Melly (2010).

[3] See for example: DiNardo, McCrary, & Sanbonmatsu, 2006; Holm & Jaeger, 2009; Lechner & Melly, 2010; Karlan, Fairlie, & Zinman, 2012; Aron-Dine et al., 2013; Bailey, Hopkins, & Rogers, 2013; Ksoll, Aker, Miller, Perez-Mendoza, & Smalley, 2014.

In another paper, Engberg et al. (2014) estimate informative bounds around the treatment effects in a magnet program using an approach based on the "worst-case" scenarios from Manski (1990) and Horowitz and Manski (2000), 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 approaches in Angrist et al. (2006). By imposing assumptions on the attrition process, these bounding approaches tend to provide tighter and more informative bounds than those provided by more relaxed approaches like Lee (2009). However, we often lack a clear understanding of the type of respondents that attrited from the sample, and to our knowledge there is no prior research on the consequences for the estimated effects of imposing these restrictive assumptions on the attrition process if they turn out no to be true.

In this paper, we study the performance of inverse probability weighting methods (Hirano, Imbens, & Ridder, 2003; Busso, DiNardo, & McCrary, 2014) and two common approaches for the estimation of informative bounds for the treatment effects (Angrist et al., 2006 and 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. A unique feature of our study is that we are able to complement a district–sourced student-level dataset, which suffers from differential attrition, with state of Oregon administrative data, provided by the Oregon Department of Education (ODE), which presents much lower rates of overall and differential attrition. This provides us with the unique opportunity of comparing the performance of different approaches to correct for differential attrition in the estimation of the effect of attending dual language immersion on student

achievement using the district level data set, and having the more complete state level data set as a benchmark for the desired estimates.

The use of a benchmark to test the relative performance of correction methods is not new (LaLonde, 1986; Heckman, Ichimura, Smith, & Todd, 1998; Dehejia & Wahba, 1999; Smith & Todd, 2005; Cook, Steiner, & Pohl, 2009; Steiner, Cook, Shadish, & Clark, 2010; Garlick & Hyman, 2016), however, there have been several studies that have lacked an experimental or quasi-experimental benchmark against which correction models could be evaluated (Mroz, 1987; Newey, Powell, & Walker, 1990; Melenberg & van Soest, 1996; Clark, Rothstein, & Schanzenbach, 2009). While we do not claim to be the first to use a benchmark to test the performance of various correction methods, we do add simulation analyses that alter the degree to which the assumptions for various methods are met, and then quantitatively assess at what point the methods become unable to correctly estimate the parameters.

The differential attrition bias-correction methods we study in this paper differ in the assumptions they use to estimate causal effects. Inverse probability weighting methods assume that we have enough observable information to model the decision to attrite from the sample. The idea is to weight observations so that the average characteristics of treated and control students remaining in the district sample look alike in key observable characteristics. On the other hand, bound estimation approaches (Angrist et al., 2006; Lee, 2009) relax the assumption that we have information on key variables driving attrition decisions and offer the estimate of potential bounds for the treatment effect under alternative assumptions about who attrites (e.g., students leaving the district are those with potentially higher outcomes if they were to stay in the district). Specifically, the main research questions in our paper are:

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?*

We begin by comparing the results of various correction methods to the benchmark, state-level dataset, but since the performance of these methods also relies on various assumptions being met, we go a step further to also compare the accuracy of the bounding methods proposed by Angrist et al. (2006) on data with artificially simulated attrition. Our results using the true scenario (non-simulated data) show that, despite differential attrition rates, estimates of the effect of attending DLI on student achievement using PPS district data were very similar to those using the more complete ODE dataset. In fact, the different methods we tried to correct for attrition did not seem to lead to less biased estimates of the effect of dual language immersion on student achievement. Therefore, it appears that some of the assumptions that differential attrition correction methods are based on might not be satisfied in our specific application. Therefore, we turn to our simulation analysis and find that the parametric and non-parametric bounding methods proposed by Angrist et al. (2006) work quite well at recovering the parameters in the benchmark case, even if the assumptions are not perfectly met. These Angrist methods require attrition that is highly correlated with potential test scores, and we find that even when attrition is based in part on random error or unobservable characteristics, as long as it is primarily based on test scores, these methods often still provide more accurate estimates than an uncorrected model.

These results highlight the fact that a variety of different types of correction methods work well if there are not big deviations from their underlying assumptions. However, our results show the importance of being aware of the assumptions that different methods are imposing on

the attrition process and of having the ability to observe important covariates or have a good understanding on what might be driving the selection or differential attrition problem. Our results are aligned with those of other studies that have also evaluated these methods. Many find that, for successful bias reduction, the selection of covariates is much more important than selection of a particular method (Cook et al., 2009; Steiner et al., 2010; Garlick & Hyman, 2016).

The rest of the paper proceeds as follows. Section 2 describes the data and sample for the analysis of performance of methods to correct for differential attrition. Section 3 describes the empirical methods studied in this paper. Finally, section 4 presents the results of the analysis using the non-simulated data and section 5 describes our simulation analysis. Finally, section 6 outlines our main conclusions.

2. DATA AND SAMPLE

This study utilizes data of seven cohorts of students who applied to attend a language immersion program in pre-k or kindergarten for the fall terms of 2004 through 2010 in Portland Public Schools (PPS) in Portland, Oregon. Slots to language immersion were assigned through a lottery system. PPS serves about 47,000 students and is among the largest two public school districts in the Pacific Northwest[4]. Outcome data were measured through the 2013-14 academic year, so the oldest cohort can be tracked through ninth grade, and include reading test scores from the state of Oregon administered tests in grades 3 to 9.[5]

In the spring prior to their child's pre-k or kindergarten year, families were able to apply to up to three school programs (including immersion or other programs types). Many of these

[4] For more details on the lottery process or the language immersion programs in PPS see Steele et al., forthcoming.
[5] Math test scores in grades 3 through 9 were also available, but we focus on the reading test scores in the current study, as there was only one grade level of significant treatment effects on math test scores. For the simulation analyses, we focus only on grade 3 outcomes, although these models could theoretically be extended to future grades as well.

programs established multiple preference categories (e.g. for native speakers of the partner language, students who live in the school's catchment neighborhood, or students living in other neighborhoods). Within each lottery round, slots in a given school and preference category are first filled by students with siblings currently attending the school, then applicants who reside within the school district, then applicants from outside the school district. A randomization lottery only occurs for one of these three categories (but most occurred within the in-district, no sibling category). Our (non-simulation) analysis focuses only on applicants whose first choice was an immersion program, and who were participants in a binding lottery. Lotteries were considered binding only if there are winners and not placed students within a given lottery category and subcategory in a given year[6]. The lottery applicants sample includes 3,457 students, 1,946 (56.3%) of which participated in binding lotteries. Of the 1,946 that participated in binding lotteries, 864 (44.4%) won immersion slots and 1,082 (55.6%) did not.

Attrition

This current paper is motivated by an issue of differential attrition from PPS between the treatment and control groups. Of the 864 students in bounding lottery categories who were originally assigned a spot in a DLI program, only 684 treatment students were enrolled in PPS in kindergarten (attrition rate of 20.8%). Of the 1,082 lottery applicants originally not assigned a spot in a DLI program, only 728 students were enrolled in PPS in kindergarten (attrition of 32.7%), yielding differential attrition of nearly 12 percentage points. See Table 1 for enrollment rates by grade. Using a student-level longitudinal dataset provided by PPS, the outcomes of these students are lost to attrition. However, using a supplemental dataset provided by ODE, we observe 752 treatment students and 873 control students, reducing attrition to 13% in the

[6] We classified the lottery-winning status of pre-k applicants based on their first application. Steele et al. (forthcoming) showed that results were not sensitive to this decision.

treatment group and 19% in the control group, for a differential of only 6 percentage points. For purposes of this analysis, we treat this state-sourced data as the benchmark dataset whose estimates we seek to replicate with weighting and bounding methods in the district-sourced data.

To better understand what types of students either leave or remain in PPS, we estimated probit models to predict whether a student is enrolled in PPS and has an observed reading test score in a given grade. Table 2 presents the results for the combined sample of treated and control students. The propensity to be initially enrolled in PPS in kindergarten is also provided in the first column, although the condition of having a test score observation was not included at this grade level. Indicators for winning the lottery were the most significant predictor of enrollment in PPS, consistent with the differential attrition rates described above. Females were more likely to be observed than males in grades six and eight. If a student was missing a race indicator due to missing data issues, this negatively predicted whether at student was observed in PPS in kindergarten, third grade, and fifth grade. Students eligible for free- and reduced-price lunch were more likely to be enrolled in PPS in kindergarten, but less likely to be enrolled in PPS in eighth grade. Students with special needs at the time of application were more likely to be observed in PPS in kindergarten as well as in grades four and five. Students whose first language was not English were less likely to be observed in PPS in kindergarten and grade three, and marginally less likely to be observed in PPS in grades 4 and 5. In only one grade (fifth) was a lagged test score predictive of observation in PPS. Students with higher fourth grade test scores were more likely to be observed in PPS in fifth grade. In summary, while the only consistent predictor of enrolling and having a reading test score in PPS was winning the lottery, there is not a clear indication that either the most advantaged or least advantaged students tend to enroll in PPS in the later grades. The probit analysis for kindergarten does indicate, however, that those

who leave the district might be relatively well off (non-FRPL, non-special needs). On the other hand, we also observe that those who leave PPS tend to be non-English speakers and for fifth grade, the leavers also tend to be lower performers in the previous year. It is then unclear based on Table 2 whether it is the top performing or bottom performing students (or a mix of both) that tend to leave Portland Public Schools. Next, to further analyze what could be driving differential attrition, we run similar probit models for the treatment and control groups, separately.

Table 3 shows the results of probit models predicting enrollment in PPS in kindergarten, and having a reading test score observed in PPS in grades three through nine, for those in the treatment group. Students in the treatment group, i.e. won a spot in a DLI program, were more likely to enroll in PPS in kindergarten if they had special needs in kindergarten. Students whose first language was not English were less likely to enroll in PPS in third grade, and marginally less likely to enroll in kindergarten as well.[7] If a student was missing a race indicator due to missing data issues, this negatively predicted whether at student was observed in PPS in kindergarten (as well as third grade). In addition, free- and reduced-price lunch eligibility also is associated with a decrease in the propensity to enroll in PPS in grade 4, and a marginally significant association with a decrease in the propensity to enroll in PPS in grade 3. Overall this tends to indicate that for the students who won the lottery, the ones who do decide to enroll in the district are relatively more advantaged.

Table 4 shows the results of similar probit models for students that were not placed in a DLI program, i.e. the control group. Students in the control group were more likely to enroll in PPS in kindergarten if they were eligible for free-and reduced-price lunch in kindergarten. Having a first language other than English negatively predicted enrollment in PPS in grade three

[7] Analysis of retention by language immersion program type indicated that Spanish program winners, in particular, were less likely to be retained in PPS, while Mandarin program winners were more likely to be retained.

and negatively, but only marginally, predicted enrollment in PPS in kindergarten. If a student was missing a race indicator due to missing data issues, this negatively predicted whether that student was observed in PPS in kindergarten (as well as grades three and five, and marginally in grade four). Control group students who had special needs in kindergarten were more likely to enroll in PPS (an increase in likelihood of about 14% to 19% in grades three through six). In fact, out of the 28 special needs students in the control group, all enrolled in PPS in kindergarten.

In summary, separating these probit analyses into treatment and control groups, we see a clearer, yet not definitive, picture of the types of students that are leaving PPS at higher rates. For the control group (who are the students more likely to leave the district, and therefore the ones of primary interest), the leavers tend to be a relatively advantaged group (non-FRPL and non-special needs in kindergarten). For the treatment group, however, it seems that the less-advantaged students are leaving with higher probability.

3. EMPIRICAL METHODS

Following Steele et al. (forthcoming) our main specification for estimating the effect of attending a DLI program on student academic performance is the following:

$$y_{it} = \beta_o + \beta_1(DLI_i^{kg} G_{it}) + \beta_2 G_{it} + \beta_3 l_i + \beta_4 X_i + \varepsilon_{it} \qquad (1)$$

Where y_{it} represents test scores in reading for student *i* at time *t*. G_{it} is a vector of grade-level fixed effects and l_i denotes lottery strata fixed effects. The key variables of interest are $DLI_i^{kg} G_{it}$ and denote being placed in an immersion program in kindergarten interacted with grade level. X_i denotes time-invariant student demographic characteristics observed in kindergarten, including the child's race or ethnicity, gender, free or reduced price lunch status, whether the

child's first language is English, and whether the child is classified in kindergarten as needing special education services.

To be able to apply all model corrections for differential attrition to the same model specification, we obtain estimates of the model in (1) using pooled ordinary least squares and obtained clustered-robust standard errors at the student level.[8] Estimates using this model will represent the average intent-to-treat parameter of attending DLI on student achievement, which we consider our parameter of interest when exploring the performance of different methods to correct for selection attrition.

Our benchmark estimates are obtained by estimating model (1) on the most complete state-sourced dataset (ODE). In the other extreme, estimates of equation (1) restricting the sample to those who enrolled in PPS would provide us with the estimated ITT effects affected by differential attrition. The main empirical challenge when differential attrition is present in a lottery analysis like the one considered here comes from the potential of selection bias in the estimated treatment effect. This is so because, among lottery applicants, information is only available from school district records if they decide to enroll in the district after the lottery results are known. If, as it is usually the case, lottery status affects the decision to finally enroll in the district, then even the randomization induced from the lottery cannot guarantee that lottery winners and lottery losers are comparable in the observed dataset of those who enrolled in the district.

There are several methods that have been proposed in the literature to address issues of differential attrition that we study in this paper. Under the strong assumptions of selection on observables, or conditional independence, and common support between treated and controls,

[8] Note that Steele et al. (forthcoming) estimated a student random effects model instead. Using this more efficient estimation approach lead to slightly more significant effects of attending DLI in several grades.

one could reweight the observations of treated and controls so they remain comparable in a set of observed characteristics. We define $\widehat{\prod(X_i)}$ as the estimated probability of being in the treatment group among binding lottery participants that decide to enroll in PPS as a function of observed characteristics. We then define weights as $\frac{1}{\widehat{\prod(X_i)}}$ for lottery winners enrolled in PPS and $\frac{1}{1-\widehat{\prod(X_i)}}$ for not placed students. We then use weighted least squares to obtain estimates of the average ITT effect of attending the DLI program. We compute separate weights for each grade and so treated and controls remain comparable within grades in terms of their demographic characteristics: child's race or ethnicity, gender, free or reduced price lunch status, whether the child's first language is English, and whether the child is classified in kindergarten as needing special education services.

Although relatively easy to compute, the inverse probability weighting approach relies on the strong assumption that we have enough information about the characteristics of treated and controls so that, by controlling for these observable characteristics, we can guarantee that treated and controls are also comparable on other unobserved characteristics. This is with no doubt a strong assumption, given the limited demographic information that is usually available from education records. Two approaches have been proposed in the literature to relax the assumption of selection on observables: Parametric selection model corrections (Heckman (1979)) and bounding analysis. Despite relaxing the assumption of selection on observables, parametric methods like Heckman's (1979) selection model often require an exclusion restriction for identification. That is, one would need to find a variable that affects the decision of enrolling in the district but that does not affect student achievement directly. Given the limited family background information available in administrative records, it is difficult to find such an

exclusion restriction in our case. Therefore, we only study the performance of bounding approaches.

The first bounding analysis that we test in this paper is the method proposed by Lee (2009). The idea behind Lee's (2009) approach 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 either that the best students in terms of test scores are the ones deciding not to enroll in the district or that the worst students in terms of test scores are the ones deciding not to enroll.

As it is also the case in Heckman's (1979) selection model, this approach is based on the following assumptions:

1) The regressor of interest (treatment variable) is independent of the errors in the outcome and selection equation. This is guaranteed through the randomization induced by the lottery in our case.

2) The selection equation can be written as a standard latent variable binary choice model. This implies that we have to assume that treatment assignment only affects enrollment in the district in one direction (i.e. we rule out heterogeneous effects of winning the lottery on enrolling in the district). Winning has to either make everybody more probable to enroll in the district or less probable. Following our estimates presented in Tables 2 through 4, we assume that it increases the likelihood for all students to enroll in the district.

Note, however, that in contrast with Heckman (1979), Lee's (2009) bounding analysis does not require exclusion restrictions for identification. In theory, Heckman's (1979) selection

model does not require these either but in practice the method does not usually work well if they are not imposed.

Specifically, Lee's (2009) bounding method works as follows. Assuming a standard latent variable sample selection model and assuming that winning the lottery induces students to enroll in the district, we know that the observed distribution of test scores for those who win the lottery is a mixture of two distributions: 1) the distribution for those who would have decided to enroll in the district irrespective of the lottery outcome and 2) the distribution of those induced into enrolling in the district because of winning the lottery. Comparing the proportion of lottery winners that enroll in the district with the proportion of not placed students that do so, we can estimate the proportion of lottery winners that were induced to enroll in the district because of winning the lottery in the following way:

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

Where each of the probabilities in (2) is estimated from the data. As in most cases, when only district data are available, it is not possible to know the characteristics of those induced to enroll in the district because of winning the lottery. This method proposes to construct extreme-case scenarios by assuming that they are either the very best students in terms of test scores or the very worst. Thus, trimming the data for lottery winners by the estimated proportion of excess students (p), estimated following equation (2), in the top and the bottom of the distribution of test scores, will provide us with bounds for the average ITT effect of those who would enroll in PPS irrespective of the treatment ("always enrollees.")

Lee's (2009) bounding approach has the advantage of relying on very few assumptions for identification. In practice, however, it can lead to bounds that are too wide and that turn out to be uninformative. In this respect, the original Lee (2009) approach did not consider covariates,

but covariates could be included in the analysis and could help estimate tighter bounds (Tauchmann, 2013; Ksoll et al., 2014). In the case of Lee (2009) bounds including covariates, 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. A weighted average of the computed bounds, weighting by the proportion of the sample in each cell, would provide us with an estimate of the average ITT effect among "always enrollees." In this line, Behaghel, Crépon, Gurgand, and Le Barbanchon (2015) suggested to incorporate information about how difficult to reach a respondent is in a bounding approach similar to Lee's (2009). Information about difficulty to reach a respondent was based on the number of attempts that were made to reach a respondent. These type of paradata information is not observable in our administrative dataset, however.

The final two bounding methods that we study in this paper are the parametric and non-parametric bounding approaches proposed by Angrist et al., (2006). The nonparametric method (Angrist et al., 2006) generally leads to tighter bounds than Lee's (2009) approach, but at the cost of making the additional assumption that selection bias only affects one part of the test score distribution, either those not enrolling in the district are the highest performing students or the lowest performing students. As in Lee (2009), this method also requires the assumption that winning the lottery affects enrollment in the district only in one direction, i.e. we assume it makes all students more likely to enroll in the district. Finally, we also need to assume that treatment affects test scores positively. With these assumptions, we then define $q_o(\theta)$ as the value in the test score distribution corresponding to the θ quantile for those who lost the lottery. Similarly, $q_1(\theta)$ is the value in the test score distribution corresponding to the θ quantile for those who won the lottery. Note that under the assumption that attending the DLI program has

positive effects, $q_1(\theta) > q_o(\theta)$. Thus, Angrist et al. (2006) showed that non-parametric bounds can be obtained 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 $q_o(\theta)$.

Bounds are obtained using linear regression models, so including covariates to tighten the bounds is easy.

Under the additional assumption that test scores are normally distributed, Angrist et al. (2006) also propose a parametric approach to correct for attrition in the estimated effects. In this case, if we assume that those who decide not to enroll in the district are those with higher potential test scores, the idea of this approach is to censor the observed distribution of outcomes at a given quantile (q1). Under the normal distribution of test scores, one could then recover the effect of treatment using a Tobit regression. This method requires the assumption that those not observed in PPS would not have scored below the chosen censoring point (q1). For robustness one could choose different censoring points, and they should lead to similar estimates of the effect of the program.

4. RESULTS

4.1 Inverse Probability Weighting

The inverse probability weights for this analysis were created using predicted probabilities of the probit model estimates presented in Table 5. Out of the subsample of students that remained in PPS, the treatment group is generally more likely to be Asian (grades three

through six) and marginally less likely to be female (grades five and six). Students that remained in PPS in sixth grade were also marginally less likely to be of another race, and students that remained in eighth grade were marginally less likely to be Hispanic.

Table 6 compares the readings results for three models: the benchmark ODE model, an unweighted PPS model, and an inverse probability weighted PPS model. The benchmark ITT effect in the benchmark ODE sample was positive and significant in grade 5 (0.150 standard deviations) and grade 8 (0.232 standard deviations). Interestingly, we see significant and positive effects in these two grades in the unweighted (uncorrected) PPS sample, indicating that even without weighting, treatment effects estimated using the PPS sample are at least somewhat similar to those estimated using the benchmark ODE sample. In five out of seven grades, the unweighted PPS results were negatively biased, and in two out of seven grades, the unweighted PPS results were positively biased.

Turning next to the inverse probability weighted reading results, we see the grade five treatment effect estimate is very close to the unweighted PPS, and now only marginally significant, perhaps indicating that we would have been about as well off, even without weighting. The grade eight treatment effect estimate is between the benchmark ODE estimate and the unweighted PPS estimate, indicating that at least in one case, inverse probability weighting produces estimates that are closer to the benchmark results. The last column of Table 6 lists the change in absolute value of bias. The magnitude of the bias only went down in two cases out of seven (grade 3 and grade 8), indicating that IPW is getting us *further from* the ODE benchmark results, in five out of seven cases. In ninth grade in particular, in which there are fewer observations, the inverse probability weighted result includes a considerable amount of bias and becomes particularly noisy.

4.2 Lee (2009) Bounds

Next, we discuss the results of the Lee (2009) bounding method. First, we report the estimated proportion of the sample of lottery winners to be trimmed following equation (2), representing the percent of lottery winners who were induced to enroll in PPS by winning the lottery. Depending on the grade level, the proportion to be trimmed ranged from about 15.7% to about 21.9% (see Table 7). These percentages indicate the proportion of observations trimmed from the upper and lower tails of the test score distribution to create the Lee (2009) bounds.

Our results following Lee's (2009) bounding approach,[9] however, lead to large and uninformative bounds. In all cases, including those in which covariates were used to tighten the bounds, the bounds included zero. The key covariates that we attempted to use to tighten the bounds were the covariates that predicted enrollment into PPS (Ksoll et al., 2014), yet their predictive value is quite weak, as indicated in Tables 2-4, so the bounds remain wide and uninformative. Wide, uninformative bounds are a common practical issue with this method (Tauchmann, 2014). To illustrate, Table 8 provides the bounds for the reading impacts. Even when the bounds are tightened using FRPL-eligibility and first language not English-status, the bounds tend to range from about -0.25 to about 0.3, and all include zero.[10] Part of the issue here, is that we lack variables that highly predict enrollment in a district. If there was a more apparent relationship between observable characteristics of students and enrollment status, the bounds could theoretically be tightened (Ksoll et al., 2014), but unlike the case of survey non-response illustrated in Behaghel et al.'s (2015) work for example, we do not have a highly predictive

[9] Lee bounds were estimated using the command "leebounds" in Stata (Tauchmann, 2014).
[10] We attempted to tighten bounds using every available combination of variables that were significantly predicting enrollment status as indicated in Tables 2-4, but in no cases did the bounds exclude zero.

variable such as the number of times that a call to a survey respondent was made before obtaining a response.

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

We next study the performance of Angrist et al.'s (2006) proposed non-parametric bounding strategy. As explained above, this approach aims to obtain tighter bounds that those of Lee (2009) by making the additional assumption that attrition only affects one side of the distribution of test scores. As we presented in Tables 2-4, we do not have strong evidence suggesting that, for the case of DLI immersion programs in PPS, those deciding to enroll in the district are either those with potentially higher test scores or those with lower test scores. Judging by special education and free-reduced lunch status the results suggest that those leaving the district, particularly from the control group, are more advantaged and potentially have higher test scores. Under this assumption, the unweighted PPS ITT estimates would be biased downward. While most of the unadjusted PPS estimates in Table 9 were biased downward, they were still biased upward in two out of seven cases (and one of the two grades in which there was a significant estimate defect of DLI in the unweighted case). Therefore, we recognize that the assumption that those who leave the district are from the top end of the potential test score distribution is not fully supported by our data. Still, we present the primary findings under this assumption. In fact, the results from Tables 2-4 suggest that the type of students leaving the treatment and control group might be different, more advantaged in the control group and less advantaged in the treatment group. Therefore, we caution that this assumption might not be satisfied in our case.[11]

[11] 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 Tables 2 and 3 do indicate that non-English speakers and those with lower lagged test scores, 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 2 out

Table 9 presents the reading results using the Angrist et al.'s (2006) non-parametric bounds. The results in Table 9 indicate that the non-parametric bounding method, including covariates, generally negatively biased results. In 22 out of 28 comparisons grade-bound combinations, the result was negatively biased, relative to the benchmark ODE results. In only 6 out of 28 of these cases was there positive bias. Grade 8 results are in line with the benchmark ODE results, however, the bounding method would generally lead to a false negative in the grade 5 results. For the grade 5 results, the negative bias has essentially pushed the estimated effect to zero. It should also be noted, that in some cases, the upper bounds were less than the corresponding lower bounds (same percentile and grade comparison). This again could indicate a misspecification issue.

4.4 Angrist, Bettinger, & Kremer (2006) Parametric Bounds

Angrist et al. (2006) suggest the use of a modified Tobit procedure to adjust for differential attrition bias. Following this method, a censored dataset was constructed by censoring the observed PPS scores at or above a particular value or quantile. Any PPS students scoring above this point, as well as the students in the ODE data who were not observed in PPS, were assigned the censoring point. Under the assumption of normality of the uncensored latent score distribution, as well as the assumption that we are only missing students from one particular tail of the distribution (upper), one could recover the censored portion using a Tobit regression model.

As a first check of the normality assumption, we performed skewness and kurtosis tests. We reject the null hypothesis that the test score distribution is normally distributed. See Figure 1

of 3 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.

for the reading test score distributions of the entire ODE benchmark sample. Skewness and kurtosis tests of each test score distribution by grade rejected normality for reading in grades three, four, five, and nine.

As the normality assumption is not met, and we have concerns that not only students in one tail of the distribution are leaving the district, the results of the parametric bounds are not reported in detail here. However, if we use this method, against the failed normality assumption, we consistently obtain large, negative impacts of the treatment, which is inconsistent with the benchmark ODE estimates, a further indication that the assumptions have failed.

5. SIMULATION ANALYSIS OF PERFORMNACE OF ANGRIST ET AL. (2006) BOUNDING APPROACHES

As discussed in previous section, we worry that the assumptions behind Angrist et al. (2206) bounding methods are not satisfied in our case. Thus, we wonder to what extend our results of bias when using these methods are driven but the fact that the underlying assumptions are not satisfied. To better understand the practicality and performance of these correction methods under various situations on the degree their underlying assumptions are met, we also performed analyses using real test score data from PPS, but simulating artificial assignment of treatment status and simulating attrition under various assumptions. First, we created a subsample of PPS students who were non-applicants to dual language immersion programs. We also limited the sample to those who were at least present in third grade as a baseline.

Using this subsample of 17,249 students, we assigned treatment status to approximately half (8,625) and control status to the other 8,624 students. Under simulated random assignment of treatment, the expected average treatment effect (ATE) is zero. This random assignment of artificial treatment status in conducted 100 times to create 100 different samples, each with a

different treatment and control group. Next, in each of these 100 datasets, we created artificial attrition of the control group under various scenarios ranging from completely random attrition to attrition based solely on third grade test scores. In grade three, we created attrition equal to 5% and 10% of the control group. It is important to note that this sample (pre-artificial simulated attrition) already will present some real attrition in grades later than three, due to the fat that it is based on real administrative data. Therefore, for the purposes of this simulation analysis, we focused on estimating effects in grade three, where all attrition is controlled by our simulation exercise and no real attrition of students has taken place in the data yet.

Under the case of completely random attrition, test scores are uncorrelated with predictors of attrition, and differential attrition should not bias the results. Therefore, under random attrition, the estimated effect even without correction should still be equivalent to the actual ATE which is zero in this exercise. Theoretically, the Angrist et al. (2006) correction methods (both parametric and non-parametric) should work best under the case of attrition based solely on third grade test scores, although the parametric method still requires an underlying test score distribution that is normal. 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, but if there is too much random attrition, then differential attrition becomes a non-issue. This simulation analysis seeks to find the situations under which Angrist et al. (2006) bounding methods do or do not correct any existing differential attrition bias.

We present results for both the parametric and non-parametric methods proposed by Angrist et al. (2006) under 14 scenarios. For attrition amounts of both 5 and 10 percent, we present results under control group attrition using seven different mechanisms: attrition based solely on test scores, attrition that is completely random, and five cases of attrition that is 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 create these attrition variables in different ratios of test score based attrition and random attrition, we used z-scores (test scores normalized to a mean of zero and standard deviation of one), and created a random variable with the same standard normal distribution. Then, we apply a propensity to attrite, by adding these two variables in various proportions. We then replicated this exercise in 100 samples and study average performance in estimating the actual ATE, which in this simulation exercise should be zero.

5.1 Angrist et al. (2006) Parametric Results under Artificially Simulated Attrition

The results of the parametric approach (Angrist et al., 2006) for the simulated sample are presented in Table 10 and Figure 2. To briefly summarize before analyzing these results in detail, this method does appear to work quite well when the assumptions underlying this approach are met, as Angrist et al. (2006) predicted. The problem, however, is that its ability to correct for differential attrition falls apart if the attrition is also based on random error, or on factors other than test scores.

In Table 10, we present statistics for each combination of attrition type (test scores, random, or a mix), and model (a naïve OLS and Tobit models censored at various percentiles). In these simulations, we assign treatment status randomly in a way such that the expected ATE should be zero. Therefore, methods that work should return estimates not significantly different from zero. Table 10 includes the average estimate calculated over each of 100 loops and the share of accurate estimates (the proportion of times out of those 100 loops where the estimate calculated was not significantly different from the expected ATE of zero). Under the case where the attrition is driven entirely by test scores (the 5 percent of control group students with the highest test scores artificially attrite), the naïve OLS model would be biased 100 percent of the

time, but the Angrist et al. (2006) parametric results using Tobit censored at the 95th percentile or lower would be accurate in at least 95% of cases.

Figure 2 also shows the "shares of accurate estimates" from the Angrist et al. (2006) parametric correct method graphically. These shares represent the number of samples out of 100 for which we calculate accurate estimates. There is a monotonic relationship between the amount of attrition that is based on test score and the share of accurate estimates. When attrition is driven entirely by test scores, the parametric results using Tobit at the 99th percentile corrected the bias in 68 out of 100 cases, but using Tobit at the 95th percentile or lower corrected the bias in at least 95 out of 100 cases. Under other types of attrition, e.g. when attrition is as much as 50 percent random, the correction method still works in the vast majority of cases, using Tobit models censored at the 90th percentile or lower. Where the randomness really begins to become an issue, it appears, is in the case of 60 percent randomness, where the Tobit censored at the 90th percentile only corrects the bias in about 64 out of 100 cases. As expected, when attrition is completely random, there is no systematic relationship between test score outcomes and treatment group status, so the estimated ATE even in the naïve OLS model is equal to the actual ATE and there is no bias to correct for. Under this situation, however, attempting to correct for this attrition actually introduces bias, as indicated in Table 10 by the fact that 0 percent of the estimates for the Tobit models were accurate in the 100 percent random case.

We also perform similar simulations but increase the level of attrition to 10 percent of control group students. These results are presented in Table 11 and Figure 3. Figure 3 indicates that, similar to the case of 5 percent attrition, the Angrist et al. (2006) parametric method works quite well as long as attrition is at least primarily test score-based. When the level of attrition is higher, the correction method does appear more sensitive to the introduction of randomness.

Similar to the case of 5 percent attrition, under the case where attrition is driven entirely by test scores, the naïve OLS model would be biased 100 percent of the time, but the Angrist et al. (2006) parametric results using Tobit censored at the 90th percentile or lower corrected the bias in all at least 96 out of 100 cases.

In other cases, where attrition is as much as 40 percent random, the correction method still has accurate estimates in at least 96 out of 100 cases using Tobit models censored at 80 percent or lower. Again, as with the case of 5 percent attrition, the randomness appears to become an issue when it reaches 60 percent. As expected, when attrition is completely random, there is no systematic relationship between test score outcomes and treatment group status, so the ATE even in the naïve OLS model is ATE and there is no bias to correct for. Under this situation, however, attempting to correct for this attrition would actually introduce bias, as with the case of 5 percent attrition that is completely random.

5.2 Angrist et al. (2006) Non-Parametric Bounding Results under Artificially Simulated Attrition

The results of the non-parametric proposed by Angrist et al. (2006) are presented in Table 12 and Figure 4, with simulated attrition of 5 percent. Overall, as it was also the case for Angrist et al. (2006) parametric bounding approach, this method does appear to work quite well even when the assumptions are not fully met, as long as most of the attrition is based on test scores, rather than random error.

In Table 12, we present three statistics for each combination of attrition type (test scores, random, or a mix), and model (a naïve OLS and estimates of bounds at various quantiles). In these simulations, we present the mean of both the lower and upper bounds, as well as the proportion of times that this range includes the expected actual ATE value of zero.

Figure 4 shows the proportion of cases in which the range included zero. Under the case where the attrition is driven entirely by test scores (the 5 percent of control group students with the highest test scores artificially attrite), the range of estimates created by the lower and upper bounds at various quantiles included the expected ATE of zero in the vast majority of cases (at least 97 out of 100). In addition, as long as the attrition is based at least primarily (60 percent or more) on test scores, the method appears to work rather well, resulting in few Type I errors (false positives) at certain quantiles. For example, as long as attrition is based at least 60% on test scores, at least 96 percent of the bounds included zero with censoring at the 90th percentile or below. Interestingly, when the attrition is based completely on random error, the range of estimates generally does include zero, and when most of the attrition is based on random error (60 percent or more, the method still works in about 68 to 99 percent of the cases), at least with censoring at the 95th percentile or below.

Next, we present the results of the non-parametric method (Angrist et al., 2006) when attrition is increased to 10 % of the control group in Table 13 and Figure 5. In Table 13, we present the mean of both the lower and upper bounds, as well as the proportion of times that this range includes the expected actual value of ATE which should be zero. Figure 5 graphically shows the proportion of cases in which the range includes zero. Under the case where the attrition is driven entirely by test scores, the range of estimates created by the lower and upper bounds at various quantiles included the expected ATE of zero in the vast majority of cases.

In addition, as with the case of 5 percent attrition, as long as the attrition is based primarily (60 percent or more) on test scores, the method appears to work rather well, resulting in few Type I errors (false positives) at certain quantiles. For example, when attrition is at least 60 percent based on test scores, and with censoring at the 85th percentile or below, the range of

estimates included zero in at least 96 percent of the cases. When the attrition is based completely on random error, the range of estimates generally does include zero, and when most of the attrition is based on random error (60 percent or more, the method still works in about 63 to 93 percent of the cases), at least with censoring at the 80th percentile or below.

Overall, the results of the simulation analysis indicated that, when the assumptions of the Angrist et al. (2006) correction methods are met (when attrition is at least primarily based on test scores), these methods generally are successful at correcting differential attrition. Therefore, we believe that the lack of accurate correction methods in the non-simulated data might be due to the fact that the underlying assumptions for these methods were not clearly met.

6. CONCLUSIONS

This study exploited a unique opportunity to test and compare the performance of various correction methods for differential attrition, a practical issue that is common in lottery-based studies. We find that, using the real (non-simulated) data, where the source of attrition is unknown, the results from the attrition-affected PPS dataset were actually similar to the results from the augmented state-provided ODE dataset, which we treat as the benchmark data source. Despite the apparent similarity between the unadjusted PPS and benchmark ODE results, we tried various correction methods, many of which seemed to add more bias, or at least more noise, to the estimates. The Angrist et al. (2006) parametric bounds and the Lee (2009) bounds did not appear to work in this particular case, indicating the limited practicality of these methods in a situation such as the one considered in this paper.

In the non-simulated data, inverse probability weighting often increased the magnitude of the bias, rather than decreasing it. In five out of seven grade-level treatment effect estimates, the

inverse probability weighted results were further from the ODE benchmark results than the unweighted PPS results were.

The second method for which we report results using the non-simulated data is the Angrist et al. (2006) non-parametric bounding method. These bounds tended to provide negatively biased results. For the grade five reading results, in which there was a positive effect in the benchmark ODE model, this method lead to false negatives, and the unadjusted PPS results were actually closer and more substantively similar to the benchmark. Adding to this confusion, we note that in several cases, the upper bound was actually lower than the corresponding lower bound.

In our non-simulation analysis, the bias and noise from using correction methods, without certainty that the necessary assumptions are met, leads to false positives or false negatives in many cases. In addition, some would have changed the overall interpretation of the results. In most cases, the unadjusted PPS results were closer and more substantively similar to the true benchmark ODE results. Keeping in mind that we normally would not have had the opportunity to compare these methods to the benchmark ODE result, the main result of the non-simulation analysis is that attempting to adjust results using these various methods probably would have not been the right choice in our case, and that researchers should approach the use of these methods with caution.

Serious concerns about whether the assumptions upon which these methods rely on were actually met motivated the second part of our study, a simulation analysis to test the performance of the Angrist et al. (2006) parametric and non-parametric methods under various types of attrition. Overall, we find these Angrist et al. (2006) bounding methods do work quite well, even

when the assumption about what types of students attrite is not fully met, as long as the majority of the attrition is based on student test scores.

The problem, however, is that in most studies or datasets, the researchers are not able to observe (or simulate) what is driving the attrition, and that little reliable information is available to determine whether the necessary conditions are met. Therefore, the most important conclusion of this study is that researchers must only use these types of correction methods with caution, and need to have strong evidence that the assumptions are met, because otherwise, using these methods incorrectly might actually introduce more bias. The results are important for researchers who use these types of methods, or for those who are consumers of information from studies that use these types of methods. In the absence of clear knowledge about what is driving attrition, there must be a strong theoretical reason for these assumptions to be met if we are to rely upon these methods.

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.

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 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.

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

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.

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. Retreived 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 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 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.

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 from: http://www.irp.wisc.edu/publications/dps/pdfs/dp12372.pdf

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 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.

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.

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. & Murarka, S. (2009). Charter schools in New York City: Who enrolls and how they affect their students' achievement. No. w14852. *National Bureau of Economic Research.* Retrieved from: http://www.nber.org/papers/w14852.pdf

Hoxby and Rockoff (2004). The impact of charter schools on student achievement. Unpublished manuscript. Retrieved 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 from: https://www.alexandria.unisg.ch/70307/1/sample_selection_bounds_incl_appendix.pdf

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

Imbens, G.W. and 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.

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.

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 from: https://www.brown.edu/academics/economics/sites/brown.edu.academics.economics/files/uploads/2010-8_paper.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 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. (2013). The aggregate effect of school choice: Evidence from a two-stage experiment in India. NBER Working Paper 19441. Retrieved 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.

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

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.

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.

Stata Manual. *Teffects ipw – Inverse-Probability Weighting.* Retrieved from: http://www.stata.com/manuals13/teteffectsipw.pdf

Steele, J.L, Slater, R.O., Zamarro, G., Miller, T., Li, J., & Burkhauser, S. (Forthcoming.) Effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal.* (October 1, 2015). Retrieved as EDRE Working Paper No. 2015-09 from http://dx.doi.org/10.2139/ssrn.2693337

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.

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 from: http://www.dgps.de/fachgruppen/methoden/mpr-online/issue11/art4/steyerCRII.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.

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: Enrollment in PPS by Grade, Binding Lottery Applicants Only

	Treatment		Control		Total
Original Participants	**864**		**1,082**		**1,946**
K	684	79%	728	67%	1,412
1	661	77%	675	62%	1,336
2	633	73%	633	59%	1,266
3	610	71%	590	55%	1,200
4	462	53%	460	43%	922
5	343	40%	357	33%	700
6	205	24%	238	22%	443
7	125	14%	171	16%	296
8	73	8%	103	10%	176
9	26	3%	62	6%	88

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

	K	3rd	4th	5th	6th	7th	8th	9th
Won Lottery	0.0685 ***	0.134 ***	0.127 ***	0.173 ***	0.141 ***	0.201 ***	0.229 ***	-0.188 ***
	(0.0160)	(0.0220)	(0.0230)	(0.0256)	(0.0357)	(0.0440)	(0.0582)	(0.0708)
Female	-0.0125	0.00896	-0.0111	0.00340	0.0756 **	0.0751	0.145 **	0.0612
	(0.0161)	(0.0225)	(0.0235)	(0.0266)	(0.0367)	(0.0466)	(0.0624)	(0.0770)
Asian	-0.0443	0.0213	0.0228	-0.00543	0.0122	-0.117	-0.0705	0.313 **
	(0.0311)	(0.0372)	(0.0379)	(0.0439)	(0.0565)	(0.0840)	(0.108)	(0.159)
Black	-0.0149	-0.101 *	-0.0195	0.0412	0.0351	0.0156	-0.00748	0.0976
	(0.0424)	(0.0569)	(0.0545)	(0.0513)	(0.0738)	(0.0898)	(0.116)	(0.193)
Hispanic	0.00798	-0.00529	0.0183	0.0467	0.0129	0.0314	0.119	0.206
	(0.0268)	(0.0378)	(0.0384)	(0.0405)	(0.0581)	(0.0698)	(0.0911)	(0.133)
Other Race	-0.0427	-0.0490	-0.0791	-0.125	-0.145	-0.196	-0.365	
	(0.0370)	(0.0499)	(0.0599)	(0.0847)	(0.141)	(0.160)	(0.243)	
Missing Race	-0.420 ***	-0.457 ***	-0.162	-0.352 ***	-0.197	-0.296	-0.235	
	(0.0862)	(0.0803)	(0.119)	(0.132)	(0.185)	(0.309)	(0.327)	
FRPL	0.0497 **	0.0395	-0.0487	-0.0587	-0.0308	-0.105	-0.182 **	0.0567
	(0.0203)	(0.0305)	(0.0361)	(0.0420)	(0.0542)	(0.0645)	(0.0916)	(0.110)
Special Needs (t=0)	0.110 ***	0.0145	0.112 ***	0.103 **	0.0909	-0.0465	-0.0520	0.0503
	(0.0177)	(0.0549)	(0.0404)	(0.0433)	(0.0743)	(0.113)	(0.134)	(0.160)
First Language Not English	-0.0892 **	-0.153 ***	-0.0849 *	-0.0976 *	-0.0890	-0.0667	-0.00330	-0.0610
	(0.0364)	(0.0448)	(0.0493)	(0.0554)	(0.0680)	(0.0796)	(0.113)	(0.108)
Lagged Test Score			0.0113	0.037 **	0.00382	-0.00189	0.0184	-0.0858 *
			(0.0134)	(0.0162)	(0.0220)	(0.0275)	(0.0372)	(0.0443)
Observations	1,625	1,581	1,095	847	591	416	251	135

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

Note: There were no test score outcomes in kindergarten, so the outcome in Kindergarten is simply enrolled in PPS.

Table 3: Propensity to Enroll in PPS for Treatment Group Only; Dependent Variable: Has Reading Test Score and Enrolls in Portland (Marginal Effects)

	K	3rd	4th	5th	6th	7th	8th	9th
Female	0.00504	0.0139	0.0362	0.0324	0.0674	0.128 **	0.126	0.0855
	(0.0203)	(0.0293)	(0.0284)	(0.0279)	(0.0487)	(0.0602)	(0.0847)	(2.487)
Asian	-0.00605	0.00581	0.0400	-0.0316	-0.00633	-0.185	-0.123	0.221
	(0.0326)	(0.0453)	(0.0382)	(0.0469)	(0.0725)	(0.122)	(0.135)	(4.919)
Black	-0.0379	-0.248 ***	-0.0745	-0.0661	0.0366	0.136 *	0.0535	
	(0.0612)	(0.0919)	(0.0798)	(0.0842)	(0.0997)	(0.0695)	(0.130)	
Hispanic	0.0243	-0.0553	0.0559	0.0536	-0.0515	-0.0167	0.0144	0.992
	(0.0324)	(0.0561)	(0.0410)	(0.0365)	(0.0881)	(0.0971)	(0.142)	(2.979)
Other Race	-0.00466	-0.0197	0.00854	-0.0739				
	(0.0444)	(0.0704)	(0.0684)	(0.133)				
Missing Race	-0.432 ***	-0.580 ***						
	(0.162)	(0.136)						
FRPL	0.0132	0.0643 *	-0.105 **	-0.0866 *	-0.0279	-0.129	-0.277	-0.474
	(0.0279)	(0.0381)	(0.0506)	(0.0525)	(0.0831)	(0.106)	(0.198)	(62.59)
Special Needs (t=0)	0.0688 ***	-0.0745	0.0577	0.0383	-0.0322	-0.166	-0.392	
	(0.0244)	(0.0729)	(0.0459)	(0.0401)	(0.112)	(0.160)	(0.240)	
First Language Not English	-0.0839 *	-0.139 **	-0.0994	-0.0747	-0.129	0.000468	0.0582	0.165
	(0.0488)	(0.0603)	(0.0655)	(0.0629)	(0.102)	(0.0996)	(0.114)	(3.915)
Lagged Test Score			-0.0120	0.0223	-0.0433	-0.0337	-0.0649	-0.1000
			(0.0160)	(0.0184)	(0.0306)	(0.0341)	(0.0537)	(2.958)
Observations	752	721	498	369	244	151	89	31

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

Note: There were no test score outcomes in kindergarten, so the outcome in Kindergarten is simply enrolled in PPS.

Table 4: Propensity to Enroll in PPS for Control Group Only; Dependent Variable: Has Reading Test Score and Enrolls in Portland (Marginal Effects)

Dependent Variable: Has Reading Test Score and Enrolls in Portland (Marginal Effects)

	K	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**
Female	-0.0328	0.0032	-0.0616 *	-0.0274	0.0747	0.0355	0.158 *	0.0889
	(0.0258)	(0.0325)	(0.0355)	(0.0414)	(0.0519)	(0.0622)	(0.0849)	(0.104)
Asian	-0.109 *	0.0214	-0.0106	0.0276	0.0004	-0.0485	-0.0345	0.270
	(0.0581)	(0.0593)	(0.0674)	(0.0716)	(0.0867)	(0.113)	(0.156)	(0.201)
Black	0.0122	0.0017	0.0600	0.175 ***	0.0319	-0.0675	-0.0543	0.301
	(0.0628)	(0.0732)	(0.0724)	(0.0607)	(0.105)	(0.126)	(0.155)	(0.261)
Hispanic	-0.0054	0.0309	-0.0073	0.0430	0.0615	0.0834	0.170	0.184
	(0.0432)	(0.0518)	(0.0588)	(0.0650)	(0.0775)	(0.0908)	(0.122)	(0.156)
Other Race	-0.0805	-0.0645	-0.15 *	-0.158	-0.150	-0.217	-0.341	
	(0.0582)	(0.0675)	(0.0854)	(0.108)	(0.152)	(0.168)	(0.226)	
Missing Race	-0.438 ***	-0.401 ***	-0.262 *	-0.457 ***	-0.262	-0.291	-0.208	
	(0.0995)	(0.0942)	(0.140)	(0.129)	(0.198)	(0.292)	(0.321)	
FRPL	0.0957 ***	0.0234	0.0000	-0.0471	-0.0434	-0.113	-0.170	0.103
	(0.0310)	(0.0451)	(0.0523)	(0.0641)	(0.0743)	(0.0821)	(0.110)	(0.136)
Special Needs (t=0)		0.149 **	0.167 ***	0.185 ***	0.192 **	0.0197	0.0938	0.272
		(0.0753)	(0.0630)	(0.0683)	(0.0977)	(0.160)	(0.167)	(0.245)
First Language Not English	-0.0954 *	-0.151 **	-0.0734	-0.120	-0.0379	-0.103	-0.0365	-0.0932
	(0.0570)	(0.0639)	(0.0741)	(0.0849)	(0.0913)	(0.106)	(0.153)	(0.147)
Lagged Test Score			0.0348 *	0.0427 *	0.0335	0.0147	0.0588	-0.0978 *
			(0.0208)	(0.0247)	(0.0305)	(0.0376)	(0.0482)	(0.0562)
Observations	845	860	594	475	345	264	162	95

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

Note: There were no test score outcomes in kindergarten, so the outcome in Kindergarten is simply enrolled in PPS.

Table 5: Propensity to be in treatment group (won lottery), conditional on enrollment in PPS and having a reading score

Dep Var: Won Lottery	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**
Female	-0.0421	-0.0373	-0.0746 *	-0.0834 *	-0.0809	-0.119	-0.141
	(0.0297)	(0.0336)	(0.0385)	(0.0489)	(0.0594)	(0.0804)	(0.155)
Asian	0.127 ***	0.154 ***	0.156 ***	0.162 **	0.0842	0.221 *	0.203
	(0.0479)	(0.0512)	(0.0574)	(0.0729)	(0.107)	(0.124)	(0.290)
Black	-0.0738	-0.0242	-0.0609	0.0751	0.147	0.102	
	(0.0691)	(0.0757)	(0.0819)	(0.0945)	(0.105)	(0.143)	
Hispanic	-0.0037	0.0391	0.0613	0.0876	-0.0027	-0.239 *	0.0789
	(0.0503)	(0.0563)	(0.0654)	(0.0778)	(0.0924)	(0.123)	(0.191)
Other Race	-0.0099	-0.0068	-0.124	-0.289 *	-0.237		
	(0.0631)	(0.0798)	(0.112)	(0.155)	(0.185)		
Missing Race	-0.181	-0.171	-0.0222	-0.151			
	(0.151)	(0.154)	(0.193)	(0.247)			
FRPL	0.00153	0.00620	0.0187	0.0531	0.0580	0.0319	-0.114
	(0.0418)	(0.0484)	(0.0567)	(0.0687)	(0.0799)	(0.120)	(0.145)
Special Needs	0.0126	0.0121	-0.0125	-0.0284	0.0493	-0.0649	
	(0.0744)	(0.0838)	(0.0948)	(0.114)	(0.152)	(0.185)	
First Language Not English	0.0494	-0.0001	0.00533	-0.0437	0.0672	0.143	0.0260
	(0.0547)	(0.0609)	(0.0708)	(0.0854)	(0.107)	(0.150)	(0.168)
Observations	1,164	908	692	440	292	169	30

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

Table 62: Comparison of Reading Results: Inverse Probability Weighting

	Benchmark ODE Sample	Unweighted PPS	Bias in Unweighted PPS	Inverse Probability Weighted PPS	Bias in IPW Portland	Change in Absolute Value of Bias
Grade 3 ITT	0.0585	0.0774	0.019	0.0756	0.017	-0.002
	(0.0508)	(0.0552)		(0.0549)		
Grade 4 ITT	0.0779	0.0648	-0.013	0.0603	-0.018	0.005
	(0.0564)	(0.0618)		(0.0615)		
Grade 5 ITT	0.150 **	0.123 *	-0.027	0.122 *	-0.028	0.001
	(0.0600)	(0.0660)		(0.0660)		
Grade 6 ITT	0.120	0.119	-0.001	0.106	-0.014	0.013
	(0.0747)	(0.0818)		(0.0821)		
Grade 7 ITT	0.117	0.0909	-0.026	0.0615	-0.056	0.029
	(0.0809)	(0.0942)		(0.0950)		
Grade 8 ITT	0.232 **	0.313 ***	0.081	0.279 **	0.047	-0.034
	(0.101)	(0.118)		(0.124)		
Grade 9 ITT	0.0917	-0.123	-0.215	-0.260	-0.352	0.137
	(0.292)	(0.310)		(0.286)		
Time Dummies	Y	Y		Y		
Demographic Controls	Y	Y		Y		
Binding Lottery Strata Fixed Effects	Y	Y		Y		
Constant	0.120	0.103		0.126		
	(0.172)	(0.190)		(0.200)		
Observations	4,594	3,705		3,695		
Students	1,447	1,208		1,208		
Adjusted R-Squared	0.3112	0.3098		0.3114		

Robust standard errors in parentheses

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

Table 3: Proportion to be trimmed in Lee Bounds Analysis, by Grade

	3rd	4th	5th	6th	7th	8th
Reading	16.5%	17.2%	20.1%	15.7%	19.2%	21.9%

Note: 9th grade calculations are not accurate due to small sample size

Table 8: Lee (2009) Bounds on Reading Treatment Effects (Grades 3-8)

	Lower Bound	Upper Bound
Grade 3	-0.26	0.30
Grade 4	-0.35	0.24
Grade 5	-0.30	0.34
Grade 6	-0.27	0.22
Grade 7	-0.35	0.17
Grade 8	-0.14	0.50

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

Table 9: Comparison of Reading Results: Non-Parametric Bounds (Angrist et al., 2006)

				Lower Bounds				Upper Bounds			
	Benchmark ODE Sample	**Unadjusted PPS**	**Bias in Unadjusted PPS**	**95% Lower Bound**	**Bias in 95% LB**	**90% Lower Bound**	**Bias in 90% LB**	**95% Upper Bound**	**Bias in 95% UB**	**90% Upper Bound**	**Bias in 90% UB**
Grade 3 ITT	0.0585	0.0774	0.019	0.0317	-0.027	0.00666	-0.052	0.0499	-0.009	0.0373	-0.021
	(0.0508)	(0.0552)		(0.0490)		(0.0475)		(0.0496)		(0.0477)	
Grade 4 ITT	0.0779	0.0648	-0.013	-0.0173	-0.095	-0.0405	-0.118	0.0248	-0.053	-0.0203	-0.098
	(0.0564)	(0.0618)		(0.0554)		(0.0536)		(0.0564)		(0.0537)	
Grade 5 ITT	0.150 **	0.123 *	-0.027	0.0880	-0.062	0.0616	-0.088	0.100	-0.050	0.0798	-0.070
	(0.0600)	(0.0660)		(0.0603)		(0.0589)		(0.0610)		(0.0590)	
Grade 6 ITT	0.120	0.119	-0.001	0.0850	-0.035	0.0618	-0.058	0.0780	-0.042	0.0430	-0.077
	(0.0747)	(0.0818)		(0.0738)		(0.0719)		(0.0743)		(0.0714)	
Grade 7 ITT	0.117	0.0909	-0.026	0.144	0.027	0.133	0.016	0.123	0.006	0.108	-0.009
	(0.0809)	(0.0942)		(0.0907)		(0.0885)		(0.0906)		(0.0878)	
Grade 8 ITT	0.232 **	0.313 ***	0.081	0.310 ***	0.078	0.228 *	-0.004	0.344 ***	0.112	0.318 ***	0.086
	(0.101)	(0.118)		(0.117)		(0.119)		(0.117)		(0.117)	
Grade 9 ITT	0.0917	-0.123	-0.215	-0.147	-0.239	0.0198	-0.072	-0.232	-0.324	-0.0846	-0.176
	(0.292)	(0.310)		(0.302)		(0.298)		(0.343)		(0.334)	
Observations	4,594	3,705		3,510		3,266		3,470		3,283	
Students	1,447	1,208		1,187		1,124		1,161		1,128	
Adjusted R-Squared	0.311	0.310		0.315		0.304		0.316		0.305	

Robust standard errors in parentheses

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

Note: 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).

Table 10: Parametric Results Under Artificially Simulated Attrition of 5% (Grade 3 Reading)

	Naïve OLS	Tobit 99%	Tobit 95%	Tobit 90%	Tobit 85%	Tobit 80%	Tobit 75%	Tobit 70%
100% Test Score								
Estimate Mean	0.10	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.68	0.95	0.97	0.96	0.96	0.96	0.96
75% Score, 25% Random								
Estimate Mean	0.10	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.60	0.96	0.97	0.96	0.96	0.96	0.96
60% Score, 40% Random								
Estimate Mean	0.09	-0.03	-0.01	0.00	0.00	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.37	0.92	0.95	0.96	0.96	0.95	0.95
50% Score, 50% Random								
Estimate Mean	0.08	-0.04	-0.02	-0.01	-0.01	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.12	0.82	0.90	0.96	0.97	0.96	0.96
40% Score, 60% Random								
Estimate Mean	0.06	-0.06	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01
Share of Accurate Estimates	0.01	0.00	0.32	0.64	0.80	0.88	0.91	0.92
25% Score, 75% Random								
Estimate Mean	0.03	-0.09	-0.06	-0.05	-0.04	-0.03	-0.03	-0.03
Share of Accurate Estimates	0.24	0.00	0.00	0.05	0.16	0.26	0.47	0.58
100% Random								
Estimate Mean	0.00	-0.14	-0.10	-0.08	-0.08	-0.07	-0.06	-0.06
Share of Accurate Estimates	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: Standard deviation of the estimates was approximately 0.01 in all cases.

Table 11: Parametric Bounding Results Under Artificially Simulated Attrition of 10%

	Naïve OLS	Tobit 99%	Tobit 95%	Tobit 90%	Tobit 85%	Tobit 80%	Tobit 75%	Tobit 70%
100% Test Score								
Estimate Mean	0.17	-0.08	-0.01	0.00	0.00	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.00	0.88	0.97	0.97	0.97	0.96	0.98
75% Score, 25% Random								
Estimate Mean	0.16	-0.09	-0.02	0.00	0.00	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.00	0.81	0.96	0.96	0.97	0.96	0.98
60% Score, 40% Random								
Estimate Mean	0.16	-0.10	-0.03	-0.02	-0.01	0.00	0.00	0.00
Share of Accurate Estimates	0.00	0.00	0.26	0.84	0.93	0.96	0.96	0.96
50% Score, 50% Random								
Estimate Mean	0.14	-0.12	-0.05	-0.03	-0.02	-0.01	-0.01	-0.01
Share of Accurate Estimates	0.00	0.00	0.03	0.36	0.71	0.85	0.91	0.94
40% Score, 60% Random								
Estimate Mean	0.11	-0.14	-0.08	-0.05	-0.04	-0.03	-0.02	-0.02
Share of Accurate Estimates	0.00	0.00	0.00	0.02	0.16	0.40	0.59	0.75
25% Score, 75% Random								
Estimate Mean	0.07	-0.20	-0.13	-0.10	-0.08	-0.07	-0.06	-0.05
Share of Accurate Estimates	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
100% Random								
Estimate Mean	0.00	-0.28	-0.20	-0.17	-0.16	-0.14	-0.13	-0.12
Share of Accurate Estimates	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: Standard deviation of the estimates was approximately 0.01 in all cases.

Table 12: Non-Parametric Bounding Results at Various Percentiles, Under Artificially Simulated Attrition of 5%

	Naïve OLS	99th Percentile	95th Percentile	90th Percentile	85th Percentile	80th Percentile	75th Percentile	70th Percentile
100% Test Score								
Lower Bound Mean	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.10	0.09	0.07	0.06	0.06	0.04	0.04	0.06
Proportion of Bounds Including Zero	0.00	0.97	1.00	1.00	0.99	0.98	0.98	1.00
75% Score, 25% Random								
Lower Bound Mean	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.10	0.09	0.07	0.06	0.06	0.04	0.04	0.06
Proportion of Bounds Including Zero	0.00	0.93	0.99	1.00	0.99	0.98	0.98	1.00
60% Score, 40% Random								
Lower Bound Mean	0.09	0.02	0.01	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.09	0.08	0.07	0.06	0.05	0.04	0.04	0.06
Proportion of Bounds Including Zero	0.00	0.58	0.94	0.96	0.97	0.98	0.98	1.00
50% Score, 50% Random								
Lower Bound Mean	0.08	0.03	0.01	0.01	0.01	0.00	0.00	0.00
Upper Bound Mean	0.08	0.07	0.07	0.06	0.05	0.04	0.04	0.06
Proportion of Bounds Including Zero	0.00	0.33	0.83	0.91	0.94	0.97	0.98	0.99
40% Score, 60% Random								
Lower Bound Mean	0.06	0.03	0.02	0.01	0.01	0.01	0.00	0.00
Upper Bound Mean	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.04
Proportion of Bounds Including Zero	0.01	0.23	0.68	0.86	0.94	0.96	0.94	0.99
25% Score, 75% Random								
Lower Bound Mean	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01
Upper Bound Mean	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.03
Proportion of Bounds Including Zero	0.25	0.47	0.69	0.83	0.87	0.90	0.93	0.90
100% Random								
Lower Bound Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Proportion of Bounds Including Zero	0.97	0.97	0.95	0.95	0.82	0.93	0.94	0.72

Note: Standard deviation of the lower bound estimates was approximately 0.01 in all cases. Standard deviation of the upper bound estimates was approximately 0.01 to 0.03 in all cases.

Table 13: Non-Parametric Bounding Results at Various Percentiles, Under Artificially Simulated Attrition of 10%

	Naïve OLS	99th Percentile	95th Percentile	90th Percentile	85th Percentile	80th Percentile	75th Percentile	70th Percentile
100% Test Score								
Lower Bound Mean	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.17	0.16	0.13	0.12	0.10	0.10	0.08	0.11
Proportion of Bounds Including Zero	0.00	1.00	1.00	0.99	0.98	0.98	1.00	1.00
75% Score, 25% Random								
Lower Bound Mean	0.17	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.17	0.15	0.14	0.11	0.10	0.10	0.08	0.11
Proportion of Bounds Including Zero	0.00	0.63	0.96	0.99	0.98	0.98	1.00	1.00
60% Score, 40% Random								
Lower Bound Mean	0.15	0.05	0.02	0.01	0.00	0.00	0.00	0.00
Upper Bound Mean	0.15	0.14	0.13	0.12	0.11	0.10	0.08	0.11
Proportion of Bounds Including Zero	0.00	0.04	0.73	0.93	0.96	0.98	0.99	0.99
50% Score, 50% Random								
Lower Bound Mean	0.14	0.06	0.03	0.02	0.01	0.01	0.01	0.01
Upper Bound Mean	0.14	0.13	0.12	0.10	0.11	0.09	0.08	0.10
Proportion of Bounds Including Zero	0.00	0.01	0.36	0.67	0.87	0.93	0.93	0.94
40% Score, 60% Random								
Lower Bound Mean	0.12	0.06	0.04	0.02	0.02	0.01	0.01	0.01
Upper Bound Mean	0.12	0.11	0.10	0.10	0.09	0.08	0.08	0.08
Proportion of Bounds Including Zero	0.00	0.00	0.14	0.42	0.65	0.83	0.86	0.87
25% Score, 75% Random								
Lower Bound Mean	0.07	0.05	0.04	0.03	0.02	0.02	0.02	0.01
Upper Bound Mean	0.07	0.07	0.06	0.06	0.07	0.05	0.05	0.07
Proportion of Bounds Including Zero	0.01	0.03	0.20	0.36	0.50	0.63	0.74	0.77
100% Random								
Lower Bound Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Upper Bound Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Proportion of Bounds Including Zero	0.97	0.95	0.92	0.95	0.85	0.93	0.92	0.74

Note: Standard deviation of the lower bound estimates was approximately 0.01 in all cases. Standard deviation of the upper bound estimates was approximately 0.01 to 0.03 in all cases.

Figure 1: ODE Reading Test Score Distribution

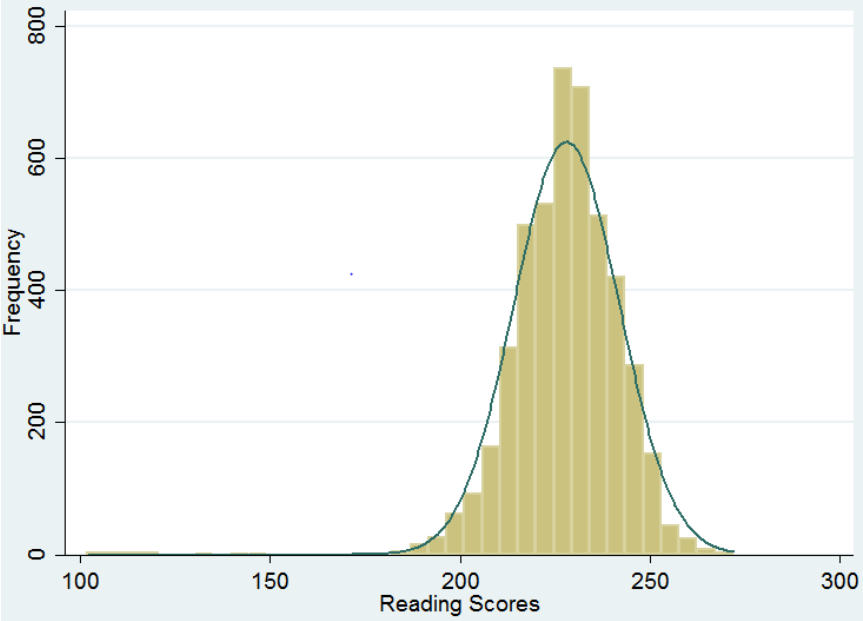

Figure 2: Parametric Results Under Artificially Simulated Attrition of 5% (Grade 3 Reading)

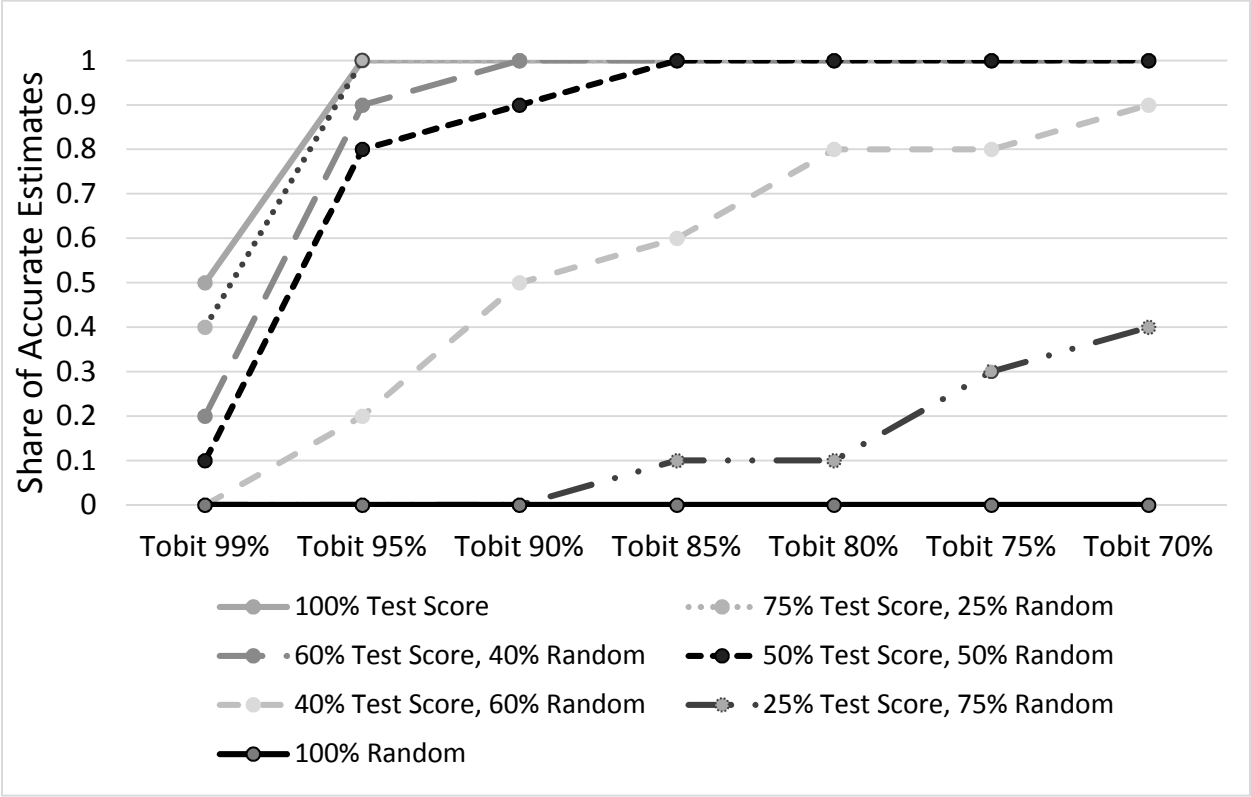

Figure 3: Parametric Bounding Results Under Artificially Simulated Attrition of 10%

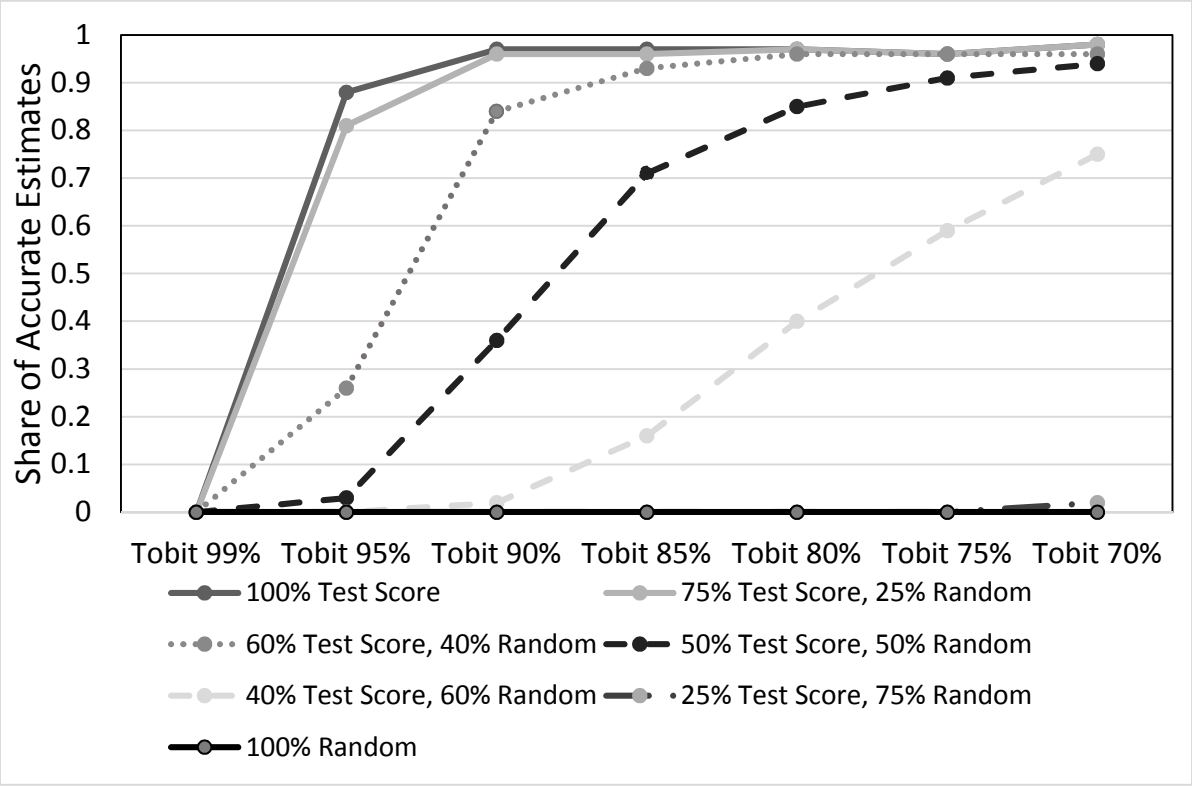

Figure 4: Non-Parametric Bounding Results at Various Percentiles, Under Artificially Simulated Attrition of 5%

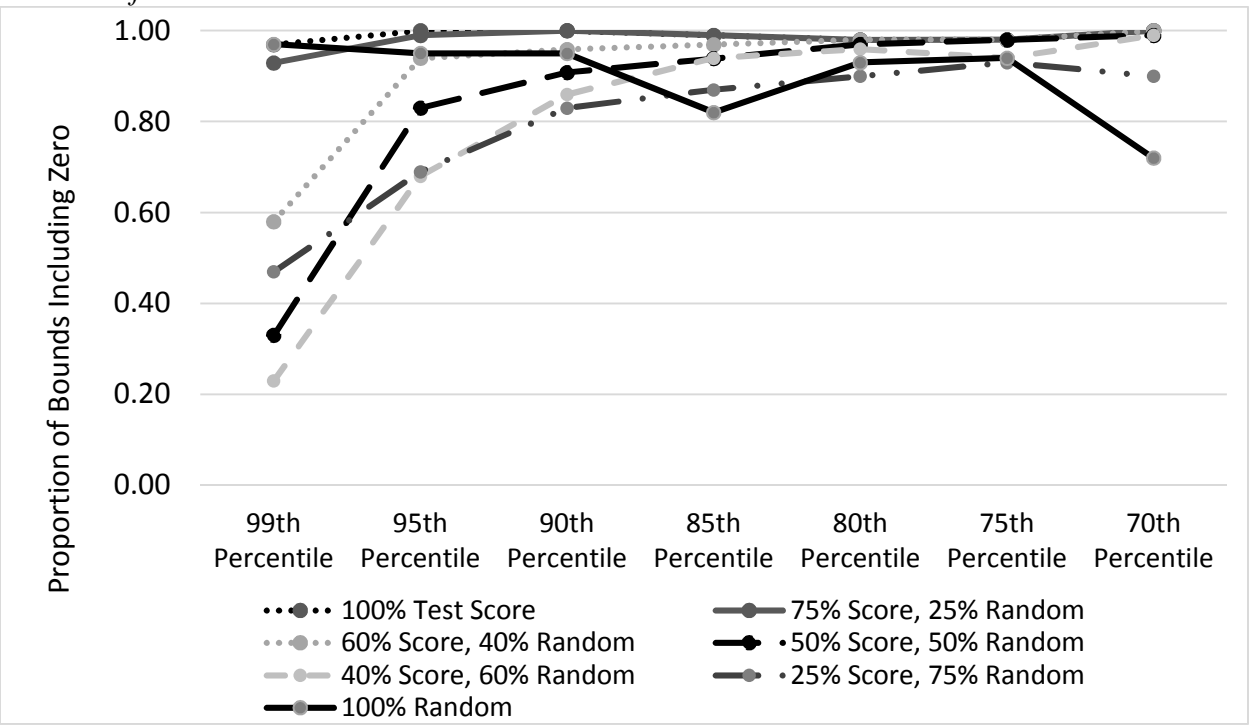

Figure 5: Non-Parametric Bounding Results at Various Percentiles, Under Artificially Simulated Attrition of 10%