LEADERSHIP AND POLICY STUDIES

# A COMPARATIVE ANALYSIS OF NON-EXPERIMENTAL METHODS FOR ESTIMATING THE IMPACT OF MAGNET SCHOOL ENROLLMENT ON STUDENT ACHIEVEMENT

DAVID A. STUIT

Dissertation under the direction of Professor R. Dale Ballou

The objective of this dissertation is to understand the circumstances under which non-experimental methods yield unbiased estimates of the effect of magnet school attendance on student achievement. This dissertation has two main analyses. In the first analysis, non-experimental estimates (via multiple regression with observed covariates, analysis of covariance with student fixed effects, and propensity score matching) of the effect of attending one academically selective magnet school on 5th and 6th grade math and reading achievement are compared to experimental estimates found using lottery status as an instrumental variable (IV) for magnet school attendance. This analysis finds that multiple regression and propensity score matching yield estimates with sizeable positive bias that would likely lead a policymaker conclude the magnet school is more effective than it really is; in some cases, this bias represents over half a school year's worth of learning. Student fixed-effects modeling performs the best of the non-experimental methods and in reading yields estimates of the magnet effect on 5th and 6th grade achievement that are not meaningfully different from the experimental IV estimates. The second analysis tests how well the experimental and non-experimental methods perform under various forms and rates of sample attrition. To investigate this issue I create a variety of samples with different forms and rates of artificial attrition among lottery winners, lottery losers, and non-participants and then run the experimental and non-experimental estimators on these simulated samples. The second analysis finds that the experimental IV estimates are less biased than the non-experimental estimates in almost all scenarios. The one exception is the student fixed-effects estimator, which performed as well or better than the experimental IV estimator as attrition rates exceeded 40%. Collectively, the findings raise caution against using multiple regression and propensity score matching to evaluate the causal impact of school choice programs, even under situations where attrition in the experimental sample is severe. Student fixed-effects modeling shows promise, particularly in reading, but only under high rates of sample attrition can one expect it to perform better than an analysis using randomly assigned comparison groups.

Approved_______________________________________________Date____________________

# A COMPARATIVE ANALYSIS OF NON-EXPERIMENTAL METHODS FOR ESTIMATING THE IMPACT OF MAGNET SCHOOL ENROLLMENT ON STUDENT ACHIEVEMENT

A COMPARATIVE

By

David A. Stuit

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

December, 2009

Nashville, Tennessee

Approved:

Professor R. Dale Ballou

Professor Mark A. Berends

Professor Thomas M. Smith

Professor Matthew G. Springer

To Katie, for unending patience and encouragement

and

To William and Wyatt, reminders of what really matters.

# ACKNOWLEDGEMENTS

I would like to acknowledge the U.S. Department of Education's Institute of Education Sciences for generous financial support of my doctoral training through a grant to Vanderbilt's Experimental Education Research Training (ExPERT) program (R305B040110). I am particularly grateful to my advisor, Dale Ballou, for setting a high standard and patiently supporting me as I worked towards it. I also want to thank my other dissertation committee members –Mark Berends, Tom Smith, and Matthew Springer – for their thoughtful suggestions. I am indebted to Ellen Goldring and Dale Ballou for allowing me to use data from their magnet school research to conduct my analyses. I would also like to acknowledge David Cordray for his insights during the early stages of this project as well as the many learning opportunities he made possible through his role as director of the ExPERT program. Finally, I wish to thank my family for their love and encouragement through this process. This would not have been possible without the understanding and support of my wife Katie.

TABLE OF CONTENTS

Page

# LIST OF TABLES

## LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

### *Overview of the Research*

This dissertation investigates non-experimental methods for estimating the causal effect of magnet school attendance on student achievement.  The overriding objective is to understand the circumstances under which non-experimental methods yield unbiased estimates of the magnet school effect and therefore can serve as adequate substitutes for experiments.  An empirical investigation of these issues is possible because I have estimates from a random assignment evaluation of magnet schools that can be used to judge the performance of the non-experimental methods.  This research builds off of a three year investigation of magnet schools by Dale Ballou and Ellen Goldring, which utilized administrative records on magnet school admissions lotteries to create randomized control groups (lottery winners and lottery losers).

Following the approach taken by Wilde and Hollisert (2007) in their investigation of non-experimental methods for evaluating class size reduction, I frame this research in the context of the decisions researchers must make when evaluating programs. The first decision is on the methods to use in the evaluation.  For those interested in causal effects, an experimental design is the gold-standard, nevertheless resource limitations or other practical concerns may lead to the selection of a non-experimental method. The ability of these non-experimental methods to yield unbiased estimates of the causal effect of the program is contingent upon a number of assumptions.

Even if the researcher is able to conduct an experiment, they may still need to decide whether or not to use non-experimental methods to address threats to the experiment's random assignment. In school choice evaluations that utilize admissions lotteries for random-assignment, the most prevalent threat is selective attrition. Selective attrition is a problem because lottery losers (i.e., the control group) often leave the district or enroll in a private school, which typically means their future test scores are unavailable to the researcher. If the control group attrition is not completely random, the experimental estimates will be biased and the researcher will need to decide on the appropriate course of action. Faced with this situation, one option is to simply ignore the threat of attrition bias and proceed with the analysis of the experimental sample. Another option is to abandon the experimental sample altogether and construct a different comparison group using a non-experimental method.

Presently there is little empirical evidence to guide these important decisions. This dissertation aims to fill this gap by comparatively evaluating the performance of non-experimental and experimental estimators of the effect of an academically selective middle school on students' math and reading achievement. It is guided by two questions that are framed in the context of research decisions:

(1) If a random-assignment study was not possible, would using a non-experimental method lead to a biased estimate of the magnet school effect?

(2) How do the experimental and non-experimental methods perform under different forms and rates of selective attrition and are there circumstances where non-experimental estimators are less biased than the experimental estimator?

To answer the first question, estimates of the magnet school effect from three common non-experimental methods (multiple regression with observed covariates, analysis of covariance with student fixed effects, and propensity score matching) are compared to the estimates from the random-assignment evaluation. The bias in the non-experimental estimates is then estimated as the difference between the non-experimental estimate and the experimental estimate.

The second question is answered through a simulated data exercise.  In this exercise I start with a complete sample that has no attrition.  I then create a variety of subsamples that have different forms and rates of artificial attrition among lottery winners, lottery losers, and non-participants.  The experimental and non-experimental estimators are then run on these simulated samples.  By comparing the experimental and non-experimental estimates from the simulated samples to the experimental estimate from the complete sample (i.e. the unbiased experimental estimate), I am able to determine the scenarios where non-experimental methods perform better than the experimental method.

*Research Motivation*

This dissertation contributes to the literature on direct empirical comparisons of experimental and non-experimental estimators.  To date, these comparisons are almost exclusively found in the field of labor economics, where researchers have compared experimental and non-experimental findings from welfare-to-work, job training, and employment services interventions.   I was only able to locate two empirical comparisons in the field of education (Rouse, 1997; Wilde & Hollister, 2002).

The paucity of evaluations of non-experimental estimators in the field of education is disparaging in light of the rise in demand among education researchers for better empirical understanding of the utility of non-experimental estimators.  This rise in demand is partly due to two recent pieces of federal legislation: the *No Child Left Behind Act of 2001* (NCLB) and the *Education Sciences Reform Act of 2002* (ESRA). NCLB raised demand by mandating that schools and districts adopt interventions that have demonstrated positive effects on student outcomes via experimental or rigorous non-experimental research. NCLB requires education interventions to be supported by research that:

> …is evaluated using experimental or quasiexperimental designs in which individuals, entities programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls. (Title IX, General Provisions, Part A Section 9101)

NCLB's mandate raised demand from education leaders for more rigorous evidence on education interventions. In turn, this raised demand from the research community for better understanding of the best practices for designing and conducting experimental research and the conditions under which non-experimental methods may produce unbiased estimates of causal effects and thus serve as adequate substitutes for experiments (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007).

ESRA also contributed to the demand for comparative evaluations of experimental and non-experimental quantitative methods by creating new standards for federal research that prioritized experimental and quasi-experimental designs. The law states that for an education evaluation to be considered "scientifically valid", it must be one that "employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible" (Title I Education Sciences Reform Act, Section 102).

ESRA converted the U.S. Department of Education's Office of Educational Research and Improvement into the Institute of Education Sciences (IES). This new federal agency was charged with implementing the new research standards and turning education into "an evidence-based field by providing decision makers with the best available research to inform their practice" (IES, 2009).

Together, ESRA and NCLB led to greater interest from the education research community in the empirical strategies for overcoming the common problems researchers encounter when trying to execute an experiment in the dynamic context of public education. They also sparked interest in rigorous non-experimental methods that have potential to answer causal questions when experiments are infeasible or when they break down due to non-random processes, such as participant non-compliance or selective attrition.

While experiments present the best opportunity for estimating an unbiased treatment effect, there are many reasons why researchers do not use them. For one, random assignment is often impossible because of the nature of the program under investigation. This is the case for interventions that target rare populations (such as special education students) for whom it is difficult to recruit a sample large enough to satisfy the statistical power requirements for an experimental design. In many situations experiments researchers cannot recruit participants who are willing to be randomly assigned to treatment conditions. Experiments can also be prohibitively expensive because they typically require the deployment of large research teams for multiple years. In some situations experiments cannot be used because they pose ethical concerns, particularly if they involve subjecting a participant to an intervention whose effects are unknown, or denying an intervention that is expected to benefit all participants, For these reasons and others, researchers often elect to utilize a non-experimental method that is cheaper, less time consuming, and/or poses less ethical concerns.  Three of the most common non-experimental methods are: using multiple regression to control for observed covariates (MR); propensity score matching (PSM); and analysis of covariance with student fixed effects (FE).

The ability of each of these methods to estimate an unbiased estimate of a treatment effect is contingent upon a number of assumptions.  A main contribution of this paper is that it discusses the assumptions behind these three non-experimental estimators and then empirically tests how well these assumptions hold in the magnet school evaluation.  This is done by comparing the non-experimental estimates of the magnet school effect on student achievement to the estimates from the random-assignment evaluation where lottery outcomes are used as an instrumental variable (IV) for magnet school attendance.  The results illuminate some of the strengths and weaknesses of the non-experimental methods as they relate to accurately estimating a school choice program effect.

This dissertation contributes directly to our understanding of the methods used to evaluate magnet schools as well as other school choice programs.   Magnet schools have been staples of urban public school systems since the late 1960s.  However, they gained new prominence in recent years because of increased public support and legislative action for expanding public school choice programs.  This is evidenced by the fact that the number of magnet schools in operation has more than doubled from 1989 to 2006, increasing from 1,165 to 2,736 (NCES, 2006).  A trend that will likely continue given the federal support for magnet schools; the president’s 2008 budget included approximately $100 million for the Magnet Schools Assistance Program (MSAP), which provides funds to assist school districts in opening new magnet schools. During the 2006-2007 school year, MSAP helped 52 school districts open 218 new magnet schools.

The proliferation of magnet schools continues despite inconclusive evidence that they are more effective than their traditional public school counterparts. The poor empirical basis for magnet schools is partly due to the methodological challenge of isolating the effect of the magnet school on student achievement from the other effects on student achievement that are independent of the school. Of particular concern is that students whose families seek out magnet schools are different from those who remain in their traditional public schools. For example, they may have more motivation to improve their child's education or more resources to transport their child to and from the school every day. It is plausible that these differences will cause the magnet students to perform better than their non-magnet counterparts regardless of the actual school they attend. This dissertation speaks directly to the merits of different methods that can be used to overcome this concern.

### *Paper Organization*

The rest of this dissertation is organized as follows: Chapter II presents the logic of causal inference and the problem of selection bias, which are central to the conceptual foundation for evaluating the experimental and non-experimental estimators. Chapter III reviews the research on empirical evaluations of non-experimental estimators, most of which is found in the field of labor economics. Chapter IV presents the methods and results of the randomized lottery evaluation of the magnet schools that are the basis for evaluating the non-experimental estimators. Chapter V discusses the assumptions and specifications of the non-experimental estimators, presents their respective estimates of the magnet school effect, and comparatively evaluates their accuracy in relation to the experimental estimates. Finally, chapter VI comparatively evaluates the experimental and non-experimental estimators under different assumptions on sample attrition via a simulated data exercise.

# CHAPTER II

## CAUSAL INFERENCE AND THE PROBLEM OF SELECTION BIAS

This section reviews the logic of causal inference using the framework pioneered by Rubin (1974, 1977, 1978, 1980) and commonly referred to as *Rubin's Causal Model* (RCM; Holland, 1986). This framework defines causal effects in terms of potential outcomes and counterfactual conditions rather than in terms of parameters of a regression model (Imbens & Angrist, 1994). In this chapter, and the rest of the dissertation, I use the standard notation developed by Rubin (1974, 1977) to discuss the assumptions of the experimental and non-experimental estimators.

To begin, consider a treatment $D$, where $D$ = 1 if a participant receives treatment and $D = 0$ if a participant does not. The objective of evaluation research is to determine the causal effect of $D$ on a designated outcome $Y$.

Three conditions must hold to allow a causal inference of the effect of $D$ on $Y$. The first is *temporal order*, where it must be established that $D$ occurred prior to the observed effect on $Y$. The second condition is *association*, where it must be established that a change in $D$ associates with a positive or negative change in $Y$. The third condition is that it must be possible to render all rival explanations for the effect implausible.

Modern statistical theory relies on the notion of counterfactuals to satisfy the third condition (Roy, 1951; Quandt, 1972; Holland, 1986; Rubin, 1974; Heckman, Ichimura, Smith, and Todd 1998). A counterfactual is defined as the condition that would have been observed had the treatment not occurred. Each person $i$ is conceived to have two possible outcomes, $Y_{i0}$ and $Y_{i1.}$ $Y_{i1}$ indicates the outcome when person $i$ receives treatment ($D$=1) and $Y_{i0}$ indicates the outcome when treatment is not received (*D=0.*) Finding the difference between $Y_{i1}$ and $Y_{i0}$ at the same point in time for the same individual allows the researcher to rule out alternative causes and identify the causal effect of the treatment as: $\delta = Y_{i1} - Y_{i0}$.

### *The Fundamental Problem of Causal Inference*

It is evident that δ cannot be identified because one cannot observe both $Y_{i1}$ and $Y_{i0}$ at the same point in time for the same individual. If $D = 1$, we observe $Y_{i1}$, but not $Y_{i0}$. Conversely, if $D = 0$, we observe $Y_{i0}$, but not $Y_{i1}$. Holland (1986) refers to this as the *fundamental problem of causal inference*. Others refer to it as the evaluation problem (Heckman et al., 1998) or the problem of unobservability (Deheija & Wahba, 2002). In essence, it is a problem of missing data because the researcher is always missing observations on either $Y_{i1}$ or $Y_{i0}$.

### *The Average Treatment Effect*

The fundamental problem of causal inference makes it impossible to *observe δ.* Holland (1986) posits that one statistical solution to this problem is to *estimate* the average treatment effect (*ATE*) of *Y* on a population. The *ATE* is found as the average difference between the outcomes for individuals that receive treatment and the outcomes for individuals that do not receive treatment: $\delta_{ATE} = E(Y_1) - E(Y_0)$.

The *ATE* is estimated using different observational units observed under different treatment conditions. For the ATE to be an unbiased estimate of *δ,* the expected outcomes of $Y_0$ and $Y_1$ must be independent of treatment assignment. Holland (1986) and others refer to this as the *independence assumption,* formally defined as: $D \perp\!\!\!\perp E(Y_0, Y_1)$.

### *Heterogeneous Treatment Effect*

The *ATE* is estimated over a population, which presents two concerns. First, it may be one is less interested in how a program impacts the entire population and more interested in how a program impacts a particular population subgroup. If the causal effect varies within the population then the estimate of the *ATE* will be a poor estimate of the treatment effect for any particular individual or subgroup within the population.

The second concern is a practical one; estimating the *ATE* requires one to have access to either the entire population of interest or a random sample of the entire population of interest. If this requirement cannot be met, the researcher will only be able to make causal inferences to the sample that is accessible. These inferences may not be of particular value to the research objectives. This is a particular concern in evaluations of programs that are optional, where individuals must voluntarily select into treatment. Individuals who seek out treatment may differ from the population of which they came in ways that influence their response to treatment and therefore it is impossible to yield a consistent estimate of the *ATE* using a self-selected sample.

### *The Average Effect of Treatment on the Treated*

In many situations, researchers are interested in learning how a program impacts those who actually receive treatment rather than learning the average effect of a treatment for a population. For illustration, consider a hypothetical case of an evaluation of a remedial math tutoring program. In this case, researchers will be primarily interested in how the program impacts low performing students (for whom the program is designed) and less interested in the effects of the program on those students performing at or above grade-level (who do not require remedial tutoring). The effect these researchers are interested in is termed the average effect of treatment on the treated (*ATT*) and can be expressed formally as:

$$\delta_{ATT} = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1) \tag{2.1}$$

From a practical perspective, estimating the *ATT* has some notable advantages to estimating the *ATE.* For one, it eliminates concern over heterogeneous treatment effects because the causal effect only generalizes to those who seek out treatment. However, if the treatment effect is constant for the population, the *ATT* will equal the *ATE*. The *ATT* may also be more practical and cost-effective parameter to estimate because it does not require random sampling or access to an entire population.

The fundamental problem of causal inference is still present in the estimation of the *ATT*. This is revealed by the fact that we can estimate $E(Y_1 \mid D = 1)$, but not $E(Y_0 \mid D = 1)$. That is, we cannot observe what would have happened to individuals in the treatment group had they not received treatment. To estimate the *ATT*, the researcher must find a substitute counterfactual for $E(Y_0 \mid D = 1)$. One option is to use the average outcome of non-participants, E($Y_0$|D=0), but this will be biased if the unobserved potential outcomes of treatment recipients differ from the observed outcomes of non-recipients. To illustrate, consider the case of a researcher who wants to know how a job training program impacts the wages of those who voluntarily enroll in the program. It would be wrong to assume the wages of those who did not enroll in the program are equivalent to the wages of those who did enroll in the program had they not received treatment. Those who enroll in the program are likely to have more motivation to improve their employment situation, which will likely positively impact their wages whether they participate in the training program or not.

*The Role of Randomization*

The random assignment of participants to treatment and control groups allows us to assume $E(Y_0 \mid D = 1)$ = E($Y_0$|D=0), by establishing the independence of treatment assignment to potential outcomes. Assuming the randomization is valid and the sample is sufficiently large, the average expected value of $Y_0$ will be equivalent for treatment participants and control group members. This allows us to estimate the ATT as: $E(Y|D=1) – E(Y|D=0)$.

*Selection Bias*

Selection bias arises when the independence assumption fails and there are unobserved differences between treatment participants and non-participants that associate with the expected values of $Y_0$: $E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0) \neq 0$

In absence of randomization, or when randomization breaks down due to non-compliance with treatment assignment, sample attrition, or other non-random events, researchers must use other methods to capture a causal inference. Holland (1986) states that the emphasis of all these methods is "…on the ways that pre-exposure variables can be used to replace the independence assumption with less stringent conditional independence assumptions" (p. 949).

The conditional independence assumption requires that conditional on variables X, treatment status is independent of the expected outcome: $D \perp\!\!\!\perp E(Y_0, Y_1|X)$. This paper is fundamentally about determining the extent to which three standard non-experimental methods (MR, FE, PSM) are able to establish conditional independence and produce estimates of the magnet school effect that are free of selection bias.

# CHAPTER III

## REVIEW OF THE LITERATURE ON COMPARATIVE EVALUATIONS OF NON-EXPERIMENTAL ESTIMATORS

This chapter summarizes the relevant literature on the empirical comparisons of non-experimental methods. These comparisons take two forms: between-study comparisons and within-study comparisons (Glazerman, Levy, & Myers, 2003). Within-study comparisons replicate the findings from an experiment using one or more non-experimental method. Between-study comparisons use meta-analytic techniques to compare findings of experimental research from non-experimental research on a given program.

### *Between-Study Comparisons*

Between-study comparisons gather the results of all experimental and non-experimental studies on a given topic and compare the mean effect sizes from the experimental designs to the mean effect sizes of the non-experimental designs using meta-analysis. This method helps discern if there are systematic differences between the research findings on a given program (or intervention) based on the research methods that were used.

The results of these comparisons are expectantly mixed. While most between-study comparison found that the experimental effects were larger, on average, than the non-experimental effects, they all revealed substantial variation in the differences between experimental and non-experimental effects. This is an expected result because the meta-analyses evaluated studies of different programs and the performance of non-experimental methods will vary depending on the characteristics of the program and its participants.

Heinsman (1993) conducted a meta-analysis on 99 studies within four seemingly unrelated areas: scholastic aptitude test coaching, ability grouping of children within classrooms, adolescent drug use prevention, and presurgical psychological interventions to improve surgery outcomes.  The author found that the mean effect size of the randomized experiments (0.42) was significantly larger than that from non-randomized experiments (0.03).

Shadish and Ragsdale (1996) conducted a meta-analysis of marital and family therapy studies aimed at understanding the difference in effect sizes between randomized and non-randomized comparison groups.  They found that the mean effect size of the 64 randomized experiments in their sample was larger than the mean effect size of the 36 non-experimental designs.  The randomized designs yielded a mean effect size of 0.60, whereas the non-randomized designs only showed an effect of 0.08 – a difference of 0.52.  When the authors accounted various covariates, including pretest effect size differences, the difference in effect sizes dropped to 0.27.

In their comprehensive review of 302 meta-analyses on studies of psychological treatments, Lipsey and Wilson (1993) found 74 meta-analyses that compared treatment effects of randomized designs to non-randomized designs.  They found the mean effect size for nonrandomized designs (0.41) to be slightly smaller than that of randomized designs (0.46).  The authors concluded that there is not a systematically strong bias in either direction that stems from the experimental or non-experimental designs.  However, when the authors graphed the distribution of the differences in effect sizes between experimental and non-experimental designs for the 74 meta-analyses, it showed wide variation, with differences distributed normally around zero. The majority of the differences were between -.40 and +.40. In some of the meta-analyses on certain psychological interventions, the bias in non-experimental estimators was substantially negative or substantially positive. The differences between experimental and non-experimental effect sizes ranged from negative 1 to positive 1.6.

The findings of Lipsey and Wilson (1993) underscore an important point: the bias in a non-experimental estimator will depend on the intervention itself and the likelihood that a selection effect would be present if random assignment were impossible. Many of the psychological interventions studied in the design may have had little bias simply because the intervention did not incentivize people to self-select into treatment for reasons that were unobserved.

The between-study design has limitations for evaluating the bias in non-experimental estimators. For one, it relies on the presence of a large enough body of evidence on a given issue to be able to conduct a meta-analysis. For most education programs policies, there are not enough studies to achieve the power necessary to test for significant differences between experimental and non-experimental estimators.

A second limitation is that between-study designs compare experimental and non-experimental studies conducted at different sites and different time periods. Moreover, the interventions that are grouped together within a meta-analysis are often very different. Consequently, the results of the between-study design still leave some uncertainty as to which methods work best because the differences in findings between experimental and non-experimental designs may be due to other factors, such as differences in the study samples or features of the intervention

.

*Within-Study Comparisons*

The within-study comparative approach enables the researcher to determine if they can eliminate selection bias through various statistical techniques. Within-study comparisons aim to estimate the selection bias in non-experimental estimators.  The estimation of selection bias is done in one of two ways: (1) estimate the selection bias as the difference between the impact estimate of the experimental method and the impact estimates of the non-experimental methods; (2) compare the average outcomes of the experimental control groups with the average outcomes of the non-experimental comparison groups.  These two approaches will yield almost identical findings if the same treatment group is used in the experimental and non-experimental methods (Glazerman et al, 2003).

The most popular program for conducting within-study evaluations has been the National Supported Work Demonstration (NSW) that was conducted during the mid-1970s in 10 sites across the U.S by the Manpower Demonstration Research Corporation (MDRC).  The NSW was a temporary employment program designed to provide work experience and counseling to disadvantaged workers.  NSW randomly assigned applicants to either participate in the NSW program or serve as a control group and receive no support. By tracking the behavior of the treatment and control group participants in the labor market over time, the study was able to determine the experimental impact of the NSW training program.  Comparing the earnings of the treatment participants to the control group, they found that males and females in the treatment group earned 9% and 8.5% more respectively than they would have without the program. However, these estimates may be biased due to sample attrition.

A number of economists have used the experimental NSW data to test non-experimental methods by replacing the experimental control group with comparison groups drawn from two national surveys: the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). The common assumption of these studies is that if the non-experimental statistical models are specified correctly, they should produce the same impact estimate of NSW on annual earnings as the experiment.

LaLonde (1986) was the first to use the NSW data to test whether the experimental estimates could be replicated with non-experimental estimators. He used different samples from both the CPS and the PSID as comparison groups and specified a variety of models that included controls for age, schooling, race and individual fixed-effects. The author's main conclusion was that non-experimental estimators are poor substitutes for experimental estimators. While some combinations of comparison groups and models came close to the experimental estimate of the impact of the NSW training program on annual earnings in 1978, others were off by more than 10%. The results for the females tend to be positive and larger than the experimental estimate, while the results for the males tend to be negative and smaller than the experimental estimate. His findings demonstrated that the specification of the comparison group as well as the econometric model can greatly influence the accuracy of the non-experimental method.

Heckman and Hotz (1989) reanalyzed LaLonde's NSW data and argued that specification tests can be used to separate the good non-experimental estimators from the bad. Their specification tests involved running the non-experimental models with the same sample, but with data from participants before they enrolled in NSW. These tests assume that if the selection-correction procedures of the non-experimental estimators are accurate, there should be no difference in earnings of future treatment participants and comparison group members prior to treatment.[1] Using this approach, the authors find that linear control models and fixed-effects models are biased, while a random growth estimators perform well and yield the same conclusion as the experiment.

Dehejia and Wahba (1999) also use NSW data to estimate the impact of NSW on earnings using propensity score matching. The authors use a subset of LaLonde's sample of adult men for whom data on two years of pre-program earnings are available. They use PSM to balance program and comparison groups on pre-program earnings and a number of other covariates. They find that PSM, relative to the estimators LaLonde evaluates, come closer to the experimental estimate.

Smith and Todd (2005) reanalyzed the data by Dehejia and Wahba (1999) to assess the sensitivity of their findings and reconcile how they were able to produce non-experimental estimators that performed better than LaLonde (1986). The authors found that the primary reason that Dehejia and Whaba's PSM estimators performed better than LaLonde's non-experimental estimators is that they used a different sample for their comparison group. Dehejia and Whaba included variables in their estimation of the propensity score that caused high earners to be dropped from their final samples. Smith and Todd show that the exclusion of the high earners is the primary reason why Dehejia and Whaba's estimators perform better than LaLonde. They show that the traditional regression and difference-in-differences estimators that LaLonde employed also perform well when Dehejia and Whaba's sample is used. Moreover, they find that the matching estimators exhibit substantial bias because of differences in how earnings are recorded in the NSW and how it is recorded in the CPS and PSID surveys. They also find that

[1] I present the results of the Heckman and Hotz (1989) specification tests for the magnet school study in the appendix, where I use 2nd, 3rd, and 4th grade data on the students used in the non-experimental methods as 5th and 6th graders.

matching estimators produce bias when treatment participants are matched to nonparticipants from different local labor markets.

Friedlander and Robins (1995) assessed non-experimental methods using experimental data from an evaluation of a mandatory welfare-to-work program. The welfare-to-work evaluation took place at multiple welfare offices within four states. To construct non-experimental comparison groups they used the experimental control from one site as the non-experimental comparison group for a different site. They used OLS regression and propensity score matching to estimate the non-experimental program impact. The authors found that the program impacts varied depending on the comparison groups that were used. Their non-experimental estimates were closer to the experimental estimates when the authors used comparison groups from the same state rather than comparison groups from different states. Contrary to Heckman and Hotz (1989) the authors did not find that specification tests could rule out bad estimators, other than “wildly inaccurate outlier estimates” (p. 935).

Heckman et al. (1998) present the most comprehensive evaluation of non-experimental estimators. The authors use data from four National Job Training Partnership Act (JTPA) randomized field trial study sites. To create non-experimental comparison groups, they used survey data on non-JTPA participants who met the JTPA criteria and were from the same neighborhoods as the JTPA sites. This eliminates the problems created by selecting samples from different geographic areas. To estimate the bias in non-experimental methods, the authors compared the outcomes of the experimental control groups with those of the non-experimental comparison groups.

The authors employed an extensive series of tests to a variety of propensity score methods and econometric models. They used variants of propensity score matching, including kernel matching and local regression matching in conjunction with matched difference-in-difference estimators. They were able to decompose the selection bias into three fundamentally different components: (1) bias due to experimental control groups with no observationally similar counterparts in the comparison group; (2) bias due to differential representation of observationally similar people in the two groups, and (3) bias due to unobserved differences between observationally similar people.

Of the econometric methods they tested, their combination of propensity score matching with a difference-in-difference estimation performs the best at eliminating bias. The main finding from this evaluation is that it is observable characteristics rather than unobservable characteristics that are the main source of bias. Most of the selection bias was due to comparing the wrong people –i.e. using a comparison group with no observationally similar counterparts to the experimental group –and comparing the right people in the wrong proportion –i.e. differential representation of observationally similar people in the two groups. They find that bias due to selection on unobservables is less important than other components, although it still represents a substantial fraction of the impact estimates. They also conclude that using propensity score matching to balance observable characteristics of the comparison groups improves the performance of the estimators. A key point of their analysis is that the correct estimation of a treatment impact using non-experimental data requires both a strong data set and the right methods. The high quality of their data explains why their estimators performed better than those of LaLonde (1986).

Bloom, Michalopoulos, Hill, & Lei (2002) conducted their comparative analysis using experimental data from the National Evaluation of Welfare-to-Work Strategies (NEWWS). Like Friedlander and Robins (1995), they drew on experimental control group members from other sites to establish the non-experimental comparison groups.  Like Heckman et al. (1998), they estimated the selection bias by comparing the outcomes of the experimental control groups to their non-experimental comparison groups after statistical adjustments.  For each experimental group, they created an in-state comparison group, an out-of-state comparison group, and a comparison group drawn from multiple states.  They estimated the bias of the non-experimental estimators for a short run time frame comprised of the first two years after randomization and a medium run time frame comprised of the third through fifth years after randomization.

The authors tested variations of propensity score matching, OLS regression, fixed-effects models, and random-growth models.  They drew upon a rich data set of participant background characteristic, employment information, and quarterly earnings data.  They found that biases for non-experimental methods are positive for some applications and negative for others.  The bias in the non-experimental estimates was consistently larger in the medium run comparisons than in the short run.  In some cases, the medium-run bias was three to five times larger than in the short run.  Of the three comparison groups (in-state, out-of-state, multi-state), the in-state comparison group produced the smallest mean bias.  The authors did not find that one statistical adjustment method was able to consistently reduce bias.  They found that using a simple difference of means performed as well as OLS, PSM, and fixed-effects models.  The random-growth model tended to increase the bias regardless of the comparison group used or the time frame of the analysis.

Glazerman et al. (2003) synthesized the results of 12 within-study comparisons of non-experimental impact estimates of welfare, job training, and employment services programs on annual earnings.  The authors found that eight of the 12 studies in the analysis demonstrated that the non-experimental estimates tended to understate the impacts, while four tended to overstate the impacts.  While some of the bias estimates were close to zero, some were very large – over-

or under-estimating annual earnings impacts by as much as 100%. For the entire sample of studies, the un-weighted average of the absolute value of the bias associated with using non-experimental methods was about $1,500, or about 15% of participants' annual earnings. Their analysis did not find that matching methods, such as PSM, performed uniformly better than traditional regression modeling. Of the matching methods, they found that one-to-one matching had less bias than other matching methods.

Wilde & Hollister (2002) present one of few studies to evaluate non-experimental methods in an education setting. They apply propensity score matching to estimate the effect of class size reduction on achievement test scores using experimental data for Kindergarteners from schools in Tennessee's Project STAR. For each of their 11 schools with 100 or more kindergartner, they construct comparison groups using out-school units; that is, they combine treatment children from a given school with control children from all other schools. They conclude that propensity score matching estimates of the treatment effect differ substantially from the experimental estimate. Of the 11 schools, they find that in four cases, the non-experimental estimate would lead to the wrong decision about whether to invest in class size reduction.

Rouse (1997) evaluated the Milwaukee private school vouchers program. She compared the lottery participants that were randomly selected for a voucher to those who were not selected to establish an experimental estimate. She also used a random sample of students from the Milwaukee public schools and used student fixed-effects to conduct a non-experimental comparison group. She found the results of her analysis using the random control group to the non-experimental comparison group to be similar. Her experimental impact estimate of vouchers on math was between 2 and 3 percentage points. Similarly, she found that private school voucher students experienced greater math gains scores than the non-experimental comparison group of about 1.6 to 1.9 percentage points a year.  In reading, she found both the experimental and non-experimental methods failed to find an impact estimate that was statistically different from zero.

# CHAPTER IV

## THE RANDOM ASSIGNMENT EVALUATION OF THE IMPACT OF MAGNET SCHOOL ATTENDANCE ON STUDENT ACHIEVEMENT

This chapter presents the findings from the random assignment evaluation of an academically selective magnet school in an urban district of a mid-sized Southern city. The data and analysis presented in this chapter stem from a three year project led by Dale Ballou and Ellen Goldring, whose original findings were reported in Ballou, Goldring, and Liu (2006) and Ballou (2007). This approach uses the results of the magnet schools' admissions lotteries to create randomized control groups (lottery winners and lottery losers). Students' lottery status is then used as an instrumental variable (IV) for magnet school enrollment in order to estimate the impact of the magnet schools on student achievement. The "experimental IV" findings presented herein serve as the basis for the comparative evaluation of the performance of the non-experimental estimators.

### *Magnet Schools in the U.S.*

Magnet schools originated in urban school districts during the late 1960s in response to "white flight" –i.e. the rapidly increasing withdrawal of non-minority families to the suburbs – and school desegregation efforts. Magnet schools provided an alternative to involuntary racial integration policies, which were on the rise since a federal court ordered the Charlotte-Mecklenburg school district in North Carolina to use forced busing to desegregate schools in 1969. Urban districts hoped they could concurrently retain white families and create racially balanced schools by creating selective magnet schools with specialized curricula and innovative programs that attracted students from across the district.

The federal Emergency School Aid Act (ESAA) of 1972 fueled magnet school growth by targeting funds towards voluntary racial integration programs (U.S. Department of Education, 2001). President Nixon advocated for the law as a means to support districts "…that wish to undertake voluntary efforts to eliminate, reduce or prevent de facto racial isolation" (Nixon, 1970).

ESAA was terminated in 1981, by which time there were over 1,019 magnet schools in operation (Rossell, 2005).[2] In 1984 the federal Magnet School Assistance Program (MSAP) picked up where ESAA left off by providing funds to districts under court-ordered desegregation to create new magnet schools (Steele & Eaton, 1996).

Over time, the policy objectives of magnet schools shifted away from school desegregation and towards the expansion of public school choice. This shift is partly due to the fact that forced desegregation orders, which made the magnet school alternative attractive to families who did not want their children bused across town, were lifted by federal courts. School districts such as Kansas City, Charlotte-Mecklenburg, and Boston were no longer obligated to continue cross-town busing and many more families were free to attend their neighborhood school. At the same time, the growing public demand for more choice in public schooling – evidenced by the rise in inter-district transfer policies and public charter schools – compelled districts to create magnet schools in order to diversify the set of public school options in the district and retain their student population.

According to the National Center on Education Statistics (Hoffman, 2006), 2.1 million students attended 2,736 public magnet schools in 31 states during the 2005-2006 school year.

[2] ESAA was eliminated under the Omnibus Budget Reconciliation Act of 1981.

Fifty-nine percent of these magnet schools were elementary, 16% were middle, 21% were high school, and the remaining 4% had other grade configurations.

Today's magnet school universe is diverse, but most schools share the following features: (1) they have a curriculum and/or instructional program that is catered to a certain student subgroup, academic interest, or content-area; (2) they are schools of choice open to all students in the district – unless they require students to meet specific admissions criteria; (3) they aim to draw students from across the entire district enrollment zone and not just from certain assigned neighborhood zones; (4) they attempt to maintain a racially and economically diverse student population.

### *Review of the Literature on Magnet School Impacts on Student Achievement*

The results of three decades of research do not provide a definitive answer as to whether or not magnet schools are more effective than their traditional public school counterparts.[3] In part, this ambiguity is due to variation in the rigor of methods that have been used to evaluate magnet schools. The impact of magnet schools on student achievement is particularly challenging to estimate because families self-select into the magnet schools for reasons that are unobserved.  If the factors that lead families to select magnet schools have an independent effect on student achievement, then comparing magnet school students to non-magnet school students may lead to biased estimates of the magnet effect.  This "selection bias" is of particular threat to the validity of the many studies that compared the achievement of students in magnet schools to students in non-magnet schools without controlling for differences in prior achievement or other characteristics (see for example, Blank, 1989; Musuneci & Szcypkowski, 1993; Poppell & Hague, 2001).

[3] See Ballou (2009) for a comprehensive review of the literature on magnet school outcomes.

The standard practice for addressing selection bias is to exploit the fact that oversubscribed magnet schools typically use random lotteries to determine who gets admitted. The random assignment via the admissions lottery creates a natural experiment, which allows researchers to evaluate the impact of the magnet school by comparing the achievement of lottery winners to lottery losers. Selection bias is resolved in because any differences between these two groups arrive solely by chance and the characteristics of the two groups will be probabilistically the same as long as the samples are sufficiently large.

Crain, Heebner, & Yiu-Pong (1992) used the admissions lottery design in an evaluation of 59 of New York City's career magnet schools. They analyzed ninth grade outcomes of a single cohort of 9th graders that included 3,272 average readers and 968 below-average readers. The researchers estimated an "Intent-to-Treat" (ITT) effect by comparing the ninth grade outcomes of those who won the lottery to those who lost the lottery, naïve of which school they actually ended up attending. Among average readers, the authors found statistically significant differences in reading scores and credits earned toward graduation. Among below-average readers, there were no significant differences in reading gains or credits earned, but lottery winners had higher pass rates on the Regent's math test. A second evaluation (Crain, Allen, & Thaler, 1999) followed the 9th grade cohort for an additional four years. They found no statistically significant difference between the reading and math results of lottery winners and lottery losers that were administered in the spring of the students' second and third years in high school.

Kemple & Snipes (2000) reported the findings of an investigation into career academies at nine sites that served over 1,700 students. The students were followed from 8th or 9th grade through the end of their scheduled 12th grade year. The researchers conducted an ITT analysis by comparing high school outcomes of those who won the lottery to those who lost the lottery. The students were divided into the three groups based on their risk of dropping out of high school. They found that the career academies substantially improved outcomes among students at high risk of dropping out, but had little effect in the aggregate for those with low or moderate drop out risk. Of those with high dropout risk, they found statistically significant positive effects of the academies on dropout rates, attendance, academic course-taking, and the likelihood of earning enough credits to graduate on time.

Kemple and Scott-Clayton (2004) followed-up on the sample used in Kemple and Snipes (2000) to evaluate their post-high school and labor market experiences four years following their scheduled graduation from high school. Their follow-up sample included more than 1,400 subjects. Among males, they found academy lottery winners had average earnings that were 18% higher than lottery losers. This difference was not found for females. Lottery winners did not have statistically different levels of educational attainment (high school graduation, enrollment in college) in the follow-up analysis, despite differences noted in the original study in course-taking and credits earned toward high school graduation.

Betts et al. (2006) also exploited admissions lotteries to assign students to treatment and control groups. They examined four years of data from a single cohort (2000-2001) that spanned all grade levels. The researchers ran separate analyses for elementary, middle, and high school magnets in each year following the lottery (2001-2002, 2002-2003, 2004-2004). They did not find a statistically significant effect of magnet schools on reading when they controlled for prior test scores. However, they did find a positive and statistically significant effect of the high school magnet program in math in the second and third year.

The latest reported experimental findings come from Ballou's 2007 study of a magnet school program in an urban school district in the South. Unlike the previous randomized studies, which conducted ITT analyses, Ballou used lottery status as an instrumental variable for magnet school attendance. The author found positive effects in the district's academically selective magnet and in a composite of four non-selective magnets. The author estimated the academically selective magnet to have a 3.5 scale score point effect in fifth grade math, but a 1.6 point loss in the sixth grade. The analysis in this chapter replicates the findings from the Ballou study, although it focuses exclusively on the academically selective magnet school.

### *The Study Setting*

This research focuses on students in an urban school district in a mid-sized Southern city. The school district serves approximately 70,000 students in Kindergarten through 12th grade. During the 2005-2006 school year, 46% of students were black, 40% were white, and 10.5% were Hispanic. Sixty-four percent qualified for free or reduced price lunch (FRL) and 10% had limited English proficiency (LEP).

At the time of the study, the district operated 13 magnet schools, 3 of which had academic admissions criteria that included performing above average on standardized achievement tests. Students who apply and meet the criteria for the academically selective magnets were selected at random via an admissions lottery. The other ten magnet schools were theme based. Most applicants to the thematic magnet schools were selected by lottery, but there were three other ways to be admitted to a thematic magnet school: (1) live within the enrollment zone of the magnet school; (2) get admitted under a sibling preference rule; (3) be admitted as a "walk-in" if the school was not over-subscribed. These alternative routes to admission do not apply to the academically selective magnet schools, where admission is restricted to only those students meeting the academic criteria.

### *The Academically Selective Magnet School*

Five of the 13 magnet schools in the district are middle schools (serving grades 5-8). Of the five magnet middle schools, four are theme-based and one is academically selective. This analysis focuses exclusively on the effect of the one academically selective magnet middle school relative to all other non-selective middle schools in the district, including the non-selective magnets.

To attend the selective magnet middle school, students must apply during the fall of their fourth grade year and have their application selected through a lottery that is held in the winter of their fourth grade year. To qualify for the lottery, a student must have a minimum grade average of 85 for the spring semester of 3rd grade, no failing grades in the first grading period of the 4th grade year, and achieved a composite 3rd grade score on a standardized test that falls in or above the seventh stanine.[4]

### *Data*

The sample used in this analysis is limited to 5th and 6th grade students. I restrict the sample to only those students who were enrolled in the school district in 4th grade. This follows the approach used by Cullen, Jacob, & Levitt (2006) in their study of magnet high school programs in Chicago. It is a necessary restriction because I do not have 4th grade achievement data on students who applied from outside the district and therefore cannot control for their prior achievement in the estimation of the magnet school effect. Excluding non-district students does

[4] Stanine is a method of scaling test scores on a nine-point standard scale with a mean of five and a standard deviation of two. Students who are at above the seventh stanine are at or above the 77th percentile of the scale score distribution.

not threaten the equivalence of the randomly assigned comparison groups because the lottery ensures they are equally represented among lottery winners and lottery losers. Moreover, it helps reduce the problem of selective attrition because these students are presumably less likely to stay in the district if they lose the lottery.

While I have data to track the students' performance from grades 5 through 8, I restrict the analysis to 5th and 6th grade students because the district operates another academically selective magnet school that begins in 7th grade and serves grades 7-12. Including the latter grades (7th & 8th) confounds a straightforward comparison of the "selective magnet school" treatment condition to the "non-selective district school" control condition because many students who lose the admissions lottery (and thus serve as our experimental control group) end up attending the other academically selective magnet in 7th grade.

The outcome data are student-level mathematics and reading scale scores from the state standardized test, which is administered to third through eighth grade students in the district. These data can be linked to five cohorts of students, with the first cohort enrolling in 5th grade in the fall of 1999 and the last cohort enrolling in 5th grade in the fall of 2003. Admissions lottery data are available from 1999-2000 to 2003-2004, but the students' standardized test scores can be tracked through 2004-2005 if they remained in the district. In addition to test score data, the analysis uses data on students' ethnicity, gender, FRL status, LEP status, and special education status.

Table 1 presents the admissions lottery activity for the sample. There were 2,282 students who enrolled in the district as 4th graders and participated in the selective magnet admissions lottery during the five years of the study. 1,087 students (47.6% of all lottery participants) won admission to the magnet school either outright or through their position on a wait list. Of the 1,087 lottery winners, 747 (68.7%) enrolled in the magnet school as 5th graders in the following fall. Of the 340 lottery winners who did not enroll, 202 remained in the district and attended a non-selective public school in 5th grade and 138 left the district before enrolling in 5th grade.

Table 1.

*Lottery Participation and Magnet School Enrollment*

| School Year | Lottery Participants | # Lottery Winners | % Lottery Winners | Lottery Losers | % Lottery Losers | Fall Enrollees |
|---|---|---|---|---|---|---|
| 1999-2000 | 386 | 202 | 52.3 | 184 | 47.7 | 126 |
| 2000-2001 | 441 | 232 | 52.6 | 209 | 47.4 | 146 |
| 2001-2002 | 514 | 218 | 42.4 | 296 | 57.6 | 154 |
| 2002-2003 | 424 | 216 | 50.7 | 208 | 49.3 | 168 |
| 2003-2004 | 517 | 219 | 42.4 | 298 | 57.6 | 153 |
| Totals | 2,282 | 1,087 | 47.6 | 1,195 | 52.4 | 747 |

*Note.* Lottery winners include both outright winners (i.e. those who won admission on day of lottery, and delayed winners (i.e. those who won admission after a period on a waiting list). Sample is limited to students who were enrolled in the district in 4th grade.

*Estimating the Experimental Effect of Selective Magnet School Enrollment*

To estimate the causal effect of the magnet school on academic outcomes, we must compare the magnet students to a group of non-magnet students who are exactly similar in all ways that would affect their potential achievement independently of the school they attend. Families who send their children to magnet schools have taken voluntary action to seek out, apply to, and enroll in the magnet schools. This self-selection poses a problem for evaluating the causal effect of the magnet school because the reasons that lead some students to enroll in the magnet school, and others not to, are expected to be correlated with students' future academic performance.

Indeed, there is research to support the claim that the typical magnet school family is different from the typical non-magnet school family. Magnet school parents tend to be more involved in their children's education than typical public school parents (Martinez, Godwin, & Kemerer, 1996; Smrekar and Goldring, 1999). In addition, evidence suggests magnet school parents are more likely to come from higher income groups than other parents in the district (Hausman and Goldring, 2000).

Parental involvement and household income are a few of the many dimensions on which magnet school students may differ from public school students in ways that affect their future achievement. If these dimensions were observed and accurately measured, one could account for them in the estimation of the causal effect via multiple regression or matching procedures. However, the fundamental problem for estimating the selective magnet effect is that most of these differences are unobserved.

To illustrate the consequences of this selection on unobservables, consider the estimation of the magnet school effect in model 5.1, where $Y$ is the achievement of student $i$ and $D$ equals 1 for students in the selective magnet school and 0 for those in other non-selective schools in the district. $\boldsymbol{X}$ represents a vector of observed characteristics of the student, including socio-economic status, demographics, and prior achievement.

$$Y_i = \beta_0 + \beta_i \boldsymbol{X}_i + \delta_i D_i + u_i \qquad (5.1)$$

Under a plausible scenario where (a) parents of selective magnet school attendees are more motivated to improve their children's educational opportunities, (b) higher parental motivation leads to better student achievement, and (c) parental motivation is an unobserved trait, the estimate of $\delta_i$ will be biased because the outcomes of the magnet school students would differ from their non-magnet counterparts in absence of treatment. Even after controlling for $\boldsymbol{X}$, the correlation of $D_i$ to $u_\mathrm{i}$ that results from unobserved differences in parental motivation would upwardly bias the estimate of the magnet effect.

The bias resulting from selection on unobservables can be dealt with experimentally if admission to the magnet school is based on a lottery. Most school districts require that magnet schools with more qualified applicants than vacancies hold a lottery to randomly determine which students are offered admission. Researchers with access to the admissions lottery data can exploit the random assignment of students via the lotteries to achieve an experimental design where self-selection is not a problem because all lottery participants seek entry to the magnet school. Randomization guarantees that unobserved dimensions, such as parental motivation, are probabilistically the same for lottery winners and lottery losers if the sample is sufficiently large.

### *Tests on the Random Assignment Process*

In this study, the admissions lottery process is centrally managed by the district and there is little reason to suspect that the outcomes of the lottery are anything but random. Nevertheless, I look for evidence of non-randomness within the results of the lottery outcomes by comparing various characteristics of lottery winners and lottery losers.[5] This test is done by conducting a t-test of the difference in means of seven variables: race (Black=1), FRL status, LEP status, gender (female =1), special education status, and students' 4th grade math and reading scale scores.

[5] Note that lottery winners are defined as those students who won the lottery outright on the day of the lottery or won because of their place on a wait list.

Table 2 presents the results of the t-tests and reveals that none of the differences in means of the lottery winners and lottery losers were statistically significant at a *p*-value below .05. This suggests the admissions lottery was indeed random and students' lottery assignment can be used to yield an unbiased estimate of the magnet effect.

Table 2.

*Test of the Randomization Process by Comparison of Lottery Winners to Lottery Losers*

| | Lottery Winners | Lottery Losers | Difference |
|---|---|---|---|
| % Black | 19.1 | 22.0 | 2.83 |
| % Free/reduced-price lunch | 13.0 | 14.8 | 1.81 |
| %ESL | 9.01 | 7.60 | 1.41 |
| %Female | 52.7 | 53.6 | 0.84 |
| % Special Education | 13.0 | 11.0 | 1.94 |
| 4th Grade Math Scale Score | 668 | 666 | 2.04 |
| 4th Grade Reading Scale Score | 685 | 684 | -.79 |
| Observations | 1,087 | 1,195 | |

* p<.05; ** p < .01; ***p < .001
*Note.* Sample restricted to 5th grade students who were enrolled in the district in 4th grade. Similar results were found for the 6th grade sample.

*Non-Compliance with Lottery Assignment*

The experimental design created by the admissions lottery is threatened by participants' non-compliance with lottery assignment. Non-compliance with lottery assignment is expected because lottery winners are not forced to enroll in the magnet school. Upon winning the lottery, families still have to exercise their option to enroll in the magnet school. Many lottery winners decide not to exercise their option to attend the magnet school and instead enroll in private schools, transfer to public schools in other surrounding districts, or enroll in other middle schools within the district. Those lottery winners who do not attend the selective magnet and rather enroll in a different middle school in the district are considered non-compliers because we observe their status in the control condition (i.e., another district school). Lottery winners who leave the district for private schools are designated as attritors because their treatment status is unobserved – we do not know if they would have complied or not-complied with lottery assignment had they remained in the district.  Non-compliance is not a concern for lottery losers because there is no way for lottery losers to enroll in the magnet school.

Table 3 reveals the extent of the lottery non-compliance in the five-year sample.  Of 1,087 students who won admission to the selective magnet over the five lotteries, 177 (16%) did not enroll in the selective magnet as 5th grade students and instead enrolled in another district middle school.  191 (18%) of lottery winners did not enroll in the selective magnet as 6th graders and instead enrolled in another district school.[6]

If the non-compliance were random and unrelated to students' potential achievement, it could be ignored in our estimation of the ATT.    However, this is an untenable assumption in this study.  We expect students who won the lottery and chose not to attend the magnet school to be systematically different from those who complied with lottery assignment in ways that affect their future achievement.  For example, the families of the 177 non-compliers may simply have less motivation to improve their children's academic performance than the families of compliers, hence their decision not to attend a perceivably better public school.  This lower family motivation may lead these students to have lower future achievement regardless of the school they attend; in which case non-compliance would lead to an over-estimation of the magnet school's effectiveness if it were not addressed.

[6] Note that a total of 340 lottery winners did not enroll in the magnet as 5th grade students.  177 of these students are non-compliers (those who remained in the district but enrolled in a different school); the remaining 163 students are attritors (students who left the district before 5th grade).

Table 3.

*Non-Compliance of Lottery Winners*

| Lottery Year | Lottery Winners | 5th Grade Fall Magnet Enrollees | 5th Grade Non-Compliers | 5th Grade Non-Compliance Rate | 6th Grade Enrollees | 6th Grade Non-Compliers | 6th Grade Non-Compliance Rate |
|---|---|---|---|---|---|---|---|
| 1999 | 202 | 126 | 34 | 16.8% | 121 | 47 | 23.3% |
| 2000 | 232 | 146 | 51 | 22.0% | 135 | 46 | 19.8% |
| 2001 | 218 | 154 | 37 | 17.0% | 147 | 40 | 18.3% |
| 2002 | 216 | 168 | 25 | 11.6% | 160 | 25 | 11.6% |
| 2003 | 219 | 153 | 30 | 13.7% | 152 | 33 | 15.1% |
| Totals | 1,087 | 747 | 177 | 16.3% | 715 | 191 | 17.6% |

*Note.* Lottery winners include both outright winners (i.e., those who won admission on day of lottery, and delayed winners (i.e., those who won admission after a period on a waiting list). Non-compliers are defined as those who won the lottery, but enrolled in a different school in the district. Those lottery winners who won the lottery, but enrolled in a private school are considered attritors because their observed “treatment” status had they remained in the district is unknown.

Another plausible scenario is that many of the families of the 177 students were unable or unwilling to provide transportation for their children to the selective magnet school. Parents must provide their own transportation to the district’s magnet schools or their children must utilize the city’s busing services. Inability to provide transportation may be an indicator of other home life circumstances that affect academic achievement, such as the amount of slack in parental time and resources available to support student learning outside of school. Refusal to provide transportation when the time and resources are available may be an indicator of parent’s value of education. Under both scenarios, the lottery winners who do not enroll in the magnet for transportation reasons will be different from the lottery winners that do enroll in the magnet in ways that may independently impact their achievement regardless of their school choice.

To find suggestive evidence of the presence of selection bias stemming from non-compliance, I conduct a test with the observed data. In table 4, I look to see if the 4th grade reading and math performance and demographic characteristics of non-compliers are on average different than compliers.[7] The rationale for this test is that the presence of statistically significant differences in observed data may suggest there are statistically significant differences in unobserved data that will bias the estimate of the magnet effect.

Table 4.

*Comparison of Characteristics of 5th grade Compliers and Non-Compliers*

| | Compliers | Non-Compliers | Difference | T-Statistic |
|---|---|---|---|---|
| 4th Grade Math Scale Score | 667.9 | 663.7 | 4.2 | 1.80 |
| 4th Grade Reading Scale Score | 684.7 | 678.8 | 5.9 | 2.57** |
| Black | 20.9% | 22.7% | 1.8% | 0.55 |
| FRL | 12.1% | 19.9% | 7.8% | 2.69** |
| Female | 53.7% | 47.3% | 6.6% | 1.57 |
| ESL | 10.3% | 5.1% | 5.2% | 2.13* |
| Disabled | 13.8% | 10.2% | 3.6% | 1.25 |
| Observations | 747 | 177 | | |

* p<.05; ** p < .01; ***p < .001

*Note.* Sample restricted to 5th grade students who were enrolled in the district in 4th grade. Non-compliers are only those students who remained in the district, but did not enroll in the selective magnet school. Lottery winners who left the district before or during 5th or 6th grade are considered attritors. Findings similar to those in table 3 were found for the 6th grade sample of lottery winners.

Table 4 reveals a few important differences between compliers and non-compliers. The 4th grade math and reading scores of non-compliers are lower than those of compliers, although

[7] These samples exclude those with missing test scores.

only the difference in reading is statistically significant (*p* = .031). The proportion of free and reduced price lunch students in the non-complier sample was larger and statistically different than the proportion of FRL students in the complier sample (*p* =.023). These statistically significant differences in observed characteristics suggest there may also be differences in unobserved characteristics between compliers and non-compliers.

This non-random compliance signals that the estimation of the ATT by comparing magnet attendees to non-magnet attendees will be biased because the randomization created by the admissions lottery is not preserved in the groups of magnet attendees and non-magnet attendees.

### *Intent to Treat Effects*

One methodological solution to non-compliance is to conduct an ITT analysis and estimate the average causal effect of being offered admission to the selective magnet school by comparing the academic performance of those who won the lottery to the academic performance of those who lost the lottery.

The ITT may be of interest to policymakers since it provides a realistic measure of the impact of an intervention that is implemented in the real world, where all participants will not take up and complete a treatment as intended. The limitation of the ITT estimate in the context of the magnet school evaluation, however, is that it does not inform policymakers of the causal effect of actually attending the magnet school. By focusing on lottery assignment, the ITT captures the causal effect of being offered a spot in the selective magnet school. It does not capture the effect of actually attending the selective magnet school.

*Instrumental Variables Regression*

In the presence of non-random non-compliance, it is possible to estimate an unbiased ATT if certain assumptions hold. This is done by a two-stage least squares (2SLS) regression of student achievement on magnet school enrollment (*D*), using the lottery assignment (*Z*) as an instrumental variable for *D*. Using *Z* as an instrumental variable (IV) is a standard technique for addressing the non-compliance problem (see, for example, Angrist, Imbens, and Rubin, 1996; Hoxby, 2000; Heckman, LaLonde, and Smith, 1999). The IV estimator will be a consistent (asympotically unbiased) estimate of the ATT as long as the admissions lottery is random.

Imbens and Angrist (1994) characterize an effect estimated by IV regression as a local average treatment effect (LATE) because inferences are restricted to the subsample of participants whose treatment status (or probability of treatment) is affected by the IV. In many cases this subsample is not of interest to the researcher and the data do not allow for inferences to a meaningful sample without strong assumptions on the effect of the IV on treatment status. However, in this evaluation the LATE is the desired parameter; the admissions lottery is the sole intended path to enrollment in the academically selective magnet and the IV estimate yields inferences for those who enroll in the magnet because of the outcome of the admissions lottery.

To estimate the effect of the academically selective magnet school on student achievement, I specify the following 2SLS IV estimator:

<u>1st Stage:</u>

$$D_{igt} = \pi_0 + \pi_x X_{igt} + \pi_c C_i + \theta_1 Z_{igt} + \lambda_g + \gamma_t + \eta_{gt} + \upsilon_{igt}$$

<u>2nd Stage:</u>

$$Y_{igt} = \beta_0 + \delta_g \hat{D}_{igt} + \beta_x X_{igt} + \beta_c C_i + \lambda_g + \gamma_t + \eta_{gt} + \varepsilon_{igt} \qquad (5.2)$$

The first-stage predicts enrollment in the selective magnet school using lottery status and all other observed covariates expected to influence achievement. *D* is equal to one if student *i* in grade *g* (5th or 6th) was enrolled in the selective magnet in year *t*. ***X*** is a vector of student characteristics that includes special program participation (LEP, FRL, special education), student attributes (female =1, Black =1), and student fourth grade test scale scores in reading and math. ***C*** indicates the year the student participated in the lottery and is included to account for effects that are constant for all students in a cohort, but vary across cohorts. *Z* is the instrument and equals one if the student won the admissions lottery either outright or after spending time on a wait list. In addition, the model includes grade fixed effects ($\lambda_g$) , year fixed effects ($\gamma_t$) , as well as an interaction of grade and year ($\eta_{gt}$) that captures changes in the test across years and grades. Standard errors in the model are adjusted for the clustering of students' observations over time.

The second-stage equation regresses the math (or reading) score of student *i* in grade *g* and year *t* on the predicted values of magnet attendance along with all other regressors, again using least squares. Note the effects of the covariates ***X*** are held constant for across years and grades, while the effect of the magnet school is allowed to vary by grade. The key point of the 2SLS regression is that the randomization created by the lottery assignment is preserved through our restriction of the inference of $\delta$ to those who complied with lottery assignment.

### *Sample Attrition*

The IV estimator addresses the problem of non-compliance, but it does provide a solution to selection bias introduced by attrition from the randomized samples of lottery winners and lottery losers. Selective attrition will lead to bias in the IV estimate of the causal effect. This threat arises primarily because it is not possible to track lottery participants who enrolled outside of the district (or in a private school) in 5th or 6th grade. A substantial number of lottery winners and lottery losers left the district after the 4th grade lottery, but prior the beginning of the 5th grade year. This is likely because the transition from 4th to5th grade is a normal transition year in the district; most students are moving from a K-4 elementary school to a 5-8 middle school and it is a natural time for parent's to shop for new schools inside and outside of the district.

Table 5 reports the attrition rates of lottery winners and lottery losers during the study years. Fifteen percent of lottery winners were missing over the five years of the study. The attrition rate among lottery losers was 38% higher, with 25% of lottery losers missing outcomes.

Manski (1995) asserts that sample attrition makes it is impossible to yield an unbiased point estimate of the causal effect without making strong assumptions on the nature of the attrition. One strong assumption is that the sample attrition yields outcomes that are missing completely at random (MCAR). Outcomes are said to be MCAR when the probability that an outcome is missing ($S =1$) is unrelated to the value of the potential outcome ($Y$) or any other variables in the model ($X$): $Pr(S|Y,X) = Pr(S)$. If the MCAR assumption holds, the missing observations can be thought of as a random subsample of the observed data and the point estimate via the IV estimator would not be biased, even in the presence of differential attrition rates (Little & Rubin, 2002).

A less restrictive assumption is that the attrition results in outcomes that are missing at random (MAR). Data are said to be MAR when the probability that an outcome is missing is unrelated to $Y$ after controlling for $X$: $Pr(S|Y,X) = Pr(S|X)$. Under an MAR assumption, we assume the regressors in the IV model adequately control for the non-random differences between the lottery winners and lottery losers that results from the sample attrition. If the MAR assumption holds, the IV point estimate of the magnet effect will still be unbiased.

Table 5.

*Attrition Rates among Lottery Winners and Losers*

| | Lottery Winners | | | Lottery Losers | | |
|---|---|---|---|---|---|---|
| | *5th Grade* | *6th Grade* | *Total* | *5th Grade* | *6th Grade* | *Total* |
| 1999-2000 | 20.8% | | 20.8% | 20.7% | | 20.7% |
| 2000-2001 | 15.1% | 16.8% | 15.9% | 26.8% | 22.5% | 24.8% |
| 2001-2002 | 12.4% | 20.3% | 16.4% | 15.9% | 37.0% | 24.6% |
| 2002-2003 | 10.6% | 13.8% | 12.2% | 31.0% | 21.2% | 25.8% |
| 2003-2004 | 16.4% | 14.4% | 15.4% | 20.8% | 33.3% | 25.9% |
| 2004-2005 | | 14.4% | 14.4% | | 27.6% | 27.6% |
| Total Attritors | 163 | 172 | 335 | 268 | 331 | 599 |
| Total Non-Attitors | 924 | 906 | 1830 | 927 | 848 | 1775 |
| Total Observations | 1,087 | 1,078 | 2,164 | 1,195 | 1,179 | 2,370 |
| Total Attrition Rate | 15.0% | 15.9% | 15.4% | 22.2% | 28.1% | 25.2% |

*Note.* Sample restricted to 5th and 6th grade students who were enrolled in the district in 4th grade.

While more tenable than MCAR, a MAR assumption is still a strong assumption in this particular study because it assumes the sample attrition does not stem from unobserved factors that relate to future achievement. The different rates of attrition suggest that lottery losers are leaving the district for reasons that do not apply to lottery winners. One plausible scenario is that families are more likely to leave the district in pursuit of better educational options if they are denied entry to the selective magnet school. Those families that leave the district may have more motivation to improve their child’s education or more resources to find other schooling options. Parental motivation and resources are two unobserved factors that will likely have independent positive effects on future achievement. If this is the case, then those who left the district would not be representative of the full sample of lottery losers and the experimental comparison of lottery losers to lottery winners would be biased. This situation would be defined as one where the data are missing not at random (MNAR).

There is no way to empirically test if the data are MAR or MNAR, but it is possible to garner evidence on the nature of the missing outcomes by examining the observed data. To do this, I conduct a test that investigates whether the sample attrition introduces additional differences between lottery losers and lottery winners that were not present in the initial randomized comparison groups. If the sample attrition introduces additional differences in observed characteristics, we have reason to suspect it may also introduce unobserved differences that the randomized lottery assignment effectively balanced between lottery winners and lottery losers.

This test takes the form of the following regression model:[8]

$$X_{it} = \beta_0 + \beta_1 Z_{it} + \beta_2 S_{it} + \beta_3 Z_{it} S_{it} + \beta_c C_{it} + \gamma_t + e_{it} \qquad (5.3)$$

Where the dependent variable *X* is one of the seven student characteristics that are used in the IV model (female, Black, FRL status, LEP status, special education status, 4th grade math and reading scores). The extent to which attrition introduces differences between lottery losers and lottery winners in X is found by the coefficient of the interaction of lottery status and missing status ($\beta_3$).

Table 6 presents the results of these models. None of the seven models revealed a statistically significant interaction of lottery status and missing status. This indicates that the

[8] In cases where the dependent variable is continuous (math and reading scale scores) least squares regression is used. In cases where the dependent variable is dichotomous (female, Black, FRL status, LEP status, special education status ) logistic regression is used.

sample attrition did not introduce additional differences in observed characteristics between lottery winners and lottery losers. This is a positive sign that selective attrition is not a concern. Nevertheless, it is possible the selective attrition still introduced unobserved differences between lottery winners and losers that may result in biased estimates.

Table 6.

*Evidence of Selective Attrition: Predicting student covariates based on lottery status and attrition status*

| | 4th Grade Math | 4th Grade Reading | Black | FRL | ESL | Female | Special Education |
|---|---|---|---|---|---|---|---|
| Lottery Winner (Z) | .959 | .1026 | -.026 | -.0241 | .012 | -.004 | .023 |
| | (1.275) | (1.314) | (.019) | (.0161) | (.012) | (.023) | (.015) |
| Missing (S) | .424 | 2.44 | -.106*** | -.068** | -.022 | .021 | .023 |
| | (1.902) | (1.960) | (.028) | (.024) | (.019) | (.034) | (.022) |
| Winner*Missing (Z*S) | 4.285 | 4.676 | -.037 | .026 | .004 | -.009 | -.032 |
| | (3.011) | (3.102) | (.044) | (.038) | (.030) | (.055) | (.035) |
| Observations: 2282 | | | | | | | |

*Note.* Sample is limited to 5th grade students who were enrolled in the district in 4th grade. All models are estimated with Huber-White robust standard errors to account for the correlation of the errors of lottery losers who attend the same non-selective magnet school

*Bounds on Estimates of the Magnet Effect under Worst-Case Assumptions on Selective Attrition*

Manski (1995) underscores that any analysis that includes missing outcome data rests on untestable assumptions. While it may be possible to assume these data are MCAR or MAR, these are strong assumptions that are likely to be violated given our hypothesis on the nature of attrition in the sample. Accordingly, Manski (1995) argues for developing bounds on the treatment effect under weak assumptions rather than the point estimation of the treatment effect under strong assumptions (e.g. MCAR, MAR). The weak assumptions implied by Manski are to assume the “worst-case” scenario on selective attrition and impute the missing outcomes of attritors with either the largest or smallest values possible given the scale of the outcome variable. This produces the respective largest and smallest possible estimates of the treatment effects that are consistent with the observable data. Manksi’s worst-case bounds on local average treatment effects (those restricted to compliers) are derived as follows:

$$\delta_{upper} = P(S=0 \mid Z=1, X)E(Y \mid D=1, Z=1, X) + P(S=1 \mid Z=1, X)y^{UB}$$
$$- P(S=0 \mid Z=0, X)E(Y \mid D=0, Z=0, X) + P(S=1 \mid Z=0, X)y^{LB}$$

$$\delta_{lower} = P(S=0 \mid Z=1, X)E(Y \mid D=1, Z=1, X) + P(S=1 \mid Z=1, X)y^{LB}$$
$$- P(S=0 \mid Z=0, X)E(Y \mid D=0, Z=0, X) + P(S=1 \mid Z=0, X)y^{UB} \qquad (5.4)$$

Where $y^{UB}$ is the upper bound (maximum value) of the distribution of *Y*, and $y^{LB}$ is the lower bound (minimum value) of the distribution of *Y*.

In the magnet school evaluation, the application of Manski’s worst-case bounds would estimate the magnet impact under the two worst-case scenarios: (1) the lottery losers with missing data would have had the highest possible math (or reading) scores had they been observed and the lottery winners with missing data would have had the lowest possible math (or reading) scores had they been observed; (2) the lottery losers with missing data would have had the lowest possible math (or reading) scores had they been observed and all the lottery winners with missing data would have had the highest possible math (or reading) scores had they been observed.

Manski's procedure is designed for situations where *Y* is a binary outcome and its utility is limited when the outcome has a continuous distribution. This is because it produces bounds that are so wide as to be uninformative. For example, when this procedure is applied to the magnet school evaluation data it produces an upper bound magnet impact estimate of 24.7 scale score points for 5th grade math and a lower bound estimate of -20.3 scale score points. The range covered within these bounds represents over two grade level differences in math performance and thus provides no useful information on whether magnet students perform better, worse, or the same on average as the non-magnet students. This is because the 22% of the 5th grade sample of lottery losers who are missing outcomes are imputed with the maximal math score and the 15.0% of the sample of lottery winner outcomes who are missing are imputed with the minimal math score, and vice versa.

Lee (2008) adapted Manski's procedure for cases where outcomes are continuous. Rather than imputing the missing data with maximal and minimal values of *Y*, Lee's procedure balances the proportion of missing outcomes between treatment and control groups by trimming maximal or minimal outcomes of the group that has fewer missing outcomes such that the proportion of missing outcomes are balanced for the treatment and control groups.

Lee's procedure also yields an upper and lower bound for the treatment effect. The exposition of these bounds when estimating a local average treatment effect is as follows:

$$\delta^{LB} = E(Y \mid D = 1, Z = 1, S = 0, X, Y \leq y_{1-p}) - E(Y \mid D = 0, Z = 0, S = 0, X)$$

$$\delta^{UB} = E(Y \mid D = 1, Z = 1, S = 0, X, Y \geq y_{p}) - E(Y \mid Z = 0, S = 0, X) \qquad (5.5)$$

Where *y* is the conditional distribution of *Y* when Z =1 and S =0.[9] *p* is the proportion of the distribution that must be trimmed, which is found as the difference in the proportion of missing outcomes between the treatment and control groups over the proportion of non-missing treatment observations:

$$p = \frac{P(S=0 \mid Z=1) - P(S=0 \mid Z=0)}{P(S=0 \mid Z=1)} \qquad (5.6)$$

Lee's procedure rests on two assumptions. The first is that individual's "potential" for attrition given their future treatment assignment is independent of their actual treatment assignment. Lee (2008) explains this assumption by denoting $S_0$ and $S_1$as "potential" sample selection indicators for the treatment and control groups. $S_0$ is the future attrition status of individual *i* if assigned to the control group and $S_1$ is the future attrition status of individual *i* if assigned to the treatment group. For example, if a student intends to attrit if assigned to the control group, but will remain if assigned to the treatment group, the corresponding values would be $S_0 = 1$, $S_1 = 0$. Random assignment of individuals to treatment and control conditions ensures this assumption holds. Formally, the independence assumption can be expressed as follows:

$$E(S_1 \mid Z=1) - E(S_1 \mid Z=0) = 0$$

$$E(S_0 \mid Z=1) - E(S_0 \mid Z=0) = 0 \qquad (5.7)$$

For each individual, we observe only $S_1$ or $S_0$. However, random assignment allows us to assume the average values of $S_1$ and $S_0$ are equivalent for the two comparison groups.

[9] This assumes the treatment group (i.e. lottery winners) has fewer missing outcomes than the control group (i.e. lottery losers), which is the case in all years of our sample.

In the magnet school study, this requires us to assume those who will leave the district upon losing the lottery (where $S_0 =1$) do not disproportionately end up in the sample of lottery losers for non-random reasons. The random assignment created by the admissions lottery allows us to assume this is not the case. While lottery status is clearly associated with increased odds of attrition, we do not expect the two comparison groups to differ in their average propensity towards attrition prior to the actual lottery assignment.

The second assumption required of Lee's trimming method is *monotonicity,* which requires the effect of Z on $S_0$ and $S_1$ to be unidirectional. This allows the study sample to be comprised of those who will always have observed outcomes regardless of treatment status ($S_0$ = 0, $S_1$ =0), those who will always have missing outcomes regardless of treatment status ($S_0$ = 1, $S_1$ =1), and those who will be observed because of the treatment status ($S_0$=1, $S_1$ =0). The monotonicity assumption does not allow the simultaneous presence of individuals for whom selection into treatment causes them to leave the sample ($S_0$=0, $S_1$=1) and individuals for whom selection into treatment causes them to stay ($S_0$=1, $S_1$=0).

In the magnet study, the monotonicity assumption requires us to assume that winning the lottery does not cause some to leave the sample while causing others to stay. Conversely, it requires us to assume that losing the lottery does not cause some to leave the sample and others to stay. These are reasonable assumptions; while we expect individuals who do not win admission to the lottery will have incentive to leave the district in search of better schooling options, we have no reason to believe that losing the lottery will create additional incentive to stay in the district. Similarly, while we expect that individuals who win the lottery have incentive to stay in the district because their desired schooling option is available, we do not expect winning the lottery to create incentive for students to leave the district.

I use Lee's trimming procedure to estimate upper and lower bounds for the estimate of the magnet effect that are robust to the concern of selective attrition. The trimming is done separately for each year and grade combination to account for different imbalances in proportions of missing treatment and control outcomes across years and grades. Following Lee (2008) and Cullen et al. (2006) I first run a regression of math and reading scores on the student covariates and then apply the trimming procedure to the conditional distribution of the predicted values of the outcomes. I trim the maximal or minimal predicted scores from the distributions in each grade and year, such that the proportion of missing outcomes is balanced between the samples of lottery winners and lottery losers in each year and grade.

Tables 7 and 8 show the proportion of missing outcomes in the comparison groups for 5th and 6th grade respectively and the number and proportion (*p*) of outcomes that were trimmed in each year. Note that in all years the proportion of missing outcomes was greater in the sample of lottery losers, which necessitated only trimming observations from the sample of lottery winners.

The bounds on the magnet effect are estimated by running the original 2SLS IV model (model 5.2) on the two trimmed samples. The lower bound sample is created by trimming the right tail of the lottery winners' test score distributions. The upper bound sample is found by trimming the left tail of the lottery winners' test score distributions. The results of this procedure yield bounds on the magnet effect that are based on the weakest possible assumptions that are still consistent with the observed data and inform the range within which one can have reasonable confidence an unbiased estimate of the magnet effect lies.

Table 7.

*Lee Trimming Calculations for 5th Grade Sample: Identifying the number of 5th grade lottery winners to remove from sample*

| | Lottery Winners | | | | | Lottery Losers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # Missing | All | % Missing | # Trimmed | *p* trimmed | # Missing | All | % Missing | # Trimmed | *P* Trimmed |
| 2000 | 42 | 202 | 20.8% | 0 | .00 | 38 | 184 | 20.7% | 0 | .00 |
| 2001 | 35 | 232 | 15.1% | 27 | .12 | 56 | 209 | 26.8% | 0 | .00 |
| 2002 | 27 | 218 | 12.4% | 8 | .04 | 47 | 296 | 15.9% | 0 | .00 |
| 2003 | 23 | 216 | 10.6% | 44 | .20 | 65 | 208 | 31.0% | 0 | .00 |
| 2004 | 36 | 219 | 16.4% | 10 | .04 | 62 | 298 | 20.8% | 0 | .00 |
| Totals | 163 | 1,087 | 15.0% | 88 | .07 | 268 | 1,195 | 22.3% | 0 | .00 |

Table 8.

*Lee Trimming Calculations for 6th Grade Sample: Identifying the number of 6th grade lottery winners to remove from sample*

| | Lottery Winners | | | | | Lottery Losers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # Missing | All | % Missing | # Trimmed | *p* trimmed | # Missing | All | % Missing | # Trimmed | *p* trimmed |
| 2001 | 34 | 202 | 16.8% | 11 | .06 | 40 | 178 | 22.5% | 0 | .00 |
| 2002 | 46 | 227 | 20.3% | 38 | .17 | 77 | 208 | 37.0% | 0 | .00 |
| 2003 | 30 | 217 | 13.8% | 18 | .08 | 64 | 292 | 21.2% | 0 | .00 |
| 2004 | 31 | 216 | 14.4% | 41 | .19 | 69 | 207 | 33.3% | 0 | .00 |
| 2005 | 31 | 216 | 14.4% | 29 | .13 | 81 | 294 | 27.6% | 0 | .00 |
| Totals | 172 | 1,078 | 15.9% | 130 | .12 | 331 | 1,179 | 28.1% | 0 | .00 |

*Results*

Table 9 presents the experimental IV estimates of the effect of enrollment in the selective magnet on 5th and 6th grade math and reading achievement.[10] In addition to the point estimates it presents the upper and lower bound estimates from Lee's trimming procedure.

Similar to Ballou's original findings, I find a positive effect of magnet school attendance on 5th grade math and 5th grade reading achievement. In math, the 5th grade estimate of magnet attendance is 5 scale score points and statistically significant (*p*. =.001). The 5th grade estimate in reading is 3.8 (*p*=.006). In 6th grade, we estimate small positive effects in both math and reading, but they are not statistically different from zero. **[11]** This implies that the selective magnet school leads to a boost in academic performance in the students' first year, but the positive effect is not sustained in 6th grade.[12]

Note the explanatory power of the IV model; the model explained 74% of the variance in 5th and 6th grade math achievement and 80% of the variance in 5th and 6th grade reading achievement after adjusting for the number of regressors in the model.

[10] Tests of the validity of the instrument are presented in the appendix.

[11] ITT estimates are similar to the 2SLS estimates and are presented in the appendix.

[12] To explore the possibility that the magnet effect is heterogeneous within the sample, I run the IV estimator on subgroups of the full population. These results are presented in the appendix. Black and economically disadvantaged students appear to benefit more from the selective magnet than their non-Black and non-FRL counterparts. In addition, the effect is largest for students who fell within the bottom quartile of the distribution of lottery participants' 4th grade math and reading scores. This suggests that the lowest performing students that gain entry to the magnet via the lottery benefit more than those who had higher 4th grade achievement levels.

Table 9.

*Impact of Selective Magnet Program on Student Achievement using Experimental Sample and Two-Stage Least Squares Estimation with Lottery Assignment as Instrument for Magnet Attendance*

| | Math | | | Reading | | |
|---|---|---|---|---|---|---|
| | Point Estimate | Lower Bound | Upper Bound | Point Estimate | Lower Bound | Upper Bound |
| Selective Magnet 5th Grade | 4.985*** | 4.712*** | 5.356*** | 3.760** | 3.618** | 3.923** |
| | (1.520) | (1.482) | (1.535) | (1.361) | (1.392) | (1.375) |
| Selective Magnet 6th Grade | 0.685 | 0.420 | 1.254 | 1.015 | 1.030 | 1.134 |
| | (1.605) | (1.623) | (1.618) | (1.438) | (1.469) | (1.458) |
| 4th Grade Reading Test | 0.169*** | 0.166*** | 0.164*** | 0.494*** | 0.489*** | 0.501*** |
| | (0.017) | (0.018) | (0.017) | (0.015) | (0.016) | (0.016) |
| 4th Grade Math Test | 0.424*** | 0.415*** | 0.443*** | .093*** | 0.094*** | 0.090*** |
| | (0.017) | (0.018) | (0.019) | (0.016) | (0.016) | (0.016) |
| Black | -7.341*** | -7.635*** | -7.508*** | -7.591*** | -8.113*** | -7.718*** |
| | (1.127) | (1.183) | (1.118) | (1.009) | (1.058) | (1.002) |
| Free and Reduced-Price Lunch | -7.072*** | -7.279*** | -6.915*** | -6.389*** | -6.134*** | -6.181*** |
| | (1.294) | (1.381) | (1.287) | (1.158) | (1.240) | (1.153) |
| ESL | 5.725*** | 6.055*** | 5.481*** | 1.626 | 1.684 | 1.769 |
| | (1.582) | (1.618) | (1.639) | (1.416) | (1.463) | (1.428) |
| Female | -3.275*** | -3.187*** | -3.528*** | -0.372 | -0.072 | -0.504 |
| | (0.877) | (0.908) | (0.885) | (0.785) | (0.808) | (0.792) |
| Special Education | 6.125*** | 6.728*** | 5.742*** | 3.539** | 3.573*** | 4.268*** |
| | (1.386) | (1.411) | (1.444) | (1.239) | (1.252) | (1.278) |
| Constant | 284.716*** | 290.395*** | 273.523*** | 270.6*** | 273.163*** | 267.943*** |
| | (13.657) | (15.578) | (15.517) | (13.370) | (13.782) | (13.881) |
| Lottery Year (Cohort Effects) | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Year Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade*Year (Test Effects) | Yes | Yes | Yes | Yes | Yes | Yes |
| R-Squared | 0.735 | 0.730 | 0.739 | 0.798 | 0.795 | 0.801 |
| Observations | 3,605 | 3,387 | 3,387 | 3,605 | 3,387 | 3,387 |

*p<0.05, ** p<0.01, *** p<0.001

*Note.* The sample is limited to 5th and 6th grade students who participated in the magnet school lottery and were enrolled in the district in 4th grade. The model is estimated with Huber-White robust standard errors to account for the correlation of errors across years within a single student (i.e. a student's 5th and 6th grade observations) and the correlation of the errors of lottery losers who attend the same non-selective magnet school.

The “worst-case” estimates of the magnet effect using Lee’s trimming procedure to address selective attrition reveal that selective attrition does not appear to be a major threaten to bias the experimental IV point estimates.  In all cases, the bounded IV estimates fall within seven percent of the IV point estimate.  This suggests the IV point estimates do not suffer from attrition bias and therefore can serve as unbiased experimental estimates to use in the comparative evaluation of non-experimental estimators.

To assess the practical importance of the positive 5th grade estimates, I transformed them into standardized effect sizes.  The average math scale score among 5th grade lottery participants was 686, with a standard deviation of 30.1.  Accordingly, an estimated effect of five scale score points translates to a Cohen’s *d* effect size of 0.17.  In reading, the average scale score among lottery participates in the sample was 697, with a standard deviation of 28.8, which translates to a Cohen’s *d* effect size of 0.13.

These effect sizes are considered small by conventional standards, but the story changes when one considers the average gains of lottery participants from 4th to 5th grade. Following the approach recommended by Kane (2004) and Hill, Bloom, Black, & Lipsey (2008), I assess the practical importance of the 5th grade effects by benchmarking them to the normal expectations of growth during one school year in absence of treatment.  I ascertain this normal growth expectation by looking at the 4th to 5th grade gains of lottery losers (17 scale score points in math and 12 scale score points in reading).  These annual gains can be translated into annual effect sizes by dividing them by the pooled standard deviations of the lottery loser’s 4th and 5th grade scale scores.  Table 10 reports these average annual gains next to the average 4th to 5th grade gains of a national norm sample published by Hill et al. (2008).

Table 10 reveals that when the estimated effects of the magnet school on 5th grade achievement are compared to the average annual 4th to 5th grade gain of lottery losers, as well as to national norms, the effect of magnet school attendance is practically large. In both math and reading, the 5th grade estimates of the magnet school effects represent around a 30% increase in the mean 4th to 5th gain in scale scores for lottery losers. They would represent similarly large effects for the national normative sample.

Table 10.

*5th Grade Estimates of the Magnet Effect on Student Achievement as Percent of Average 4th to 5th grade gain and as approximate weeks of instruction*

| | Math | | Reading | |
|---|---|---|---|---|
| | Lottery Losers | National Norms | Lottery Losers | National Norms |
| Grade 4-5 Mean Annual Gain as Effect Size | 0.58 | 0.56 | 0.40 | 0.40 |
| 5th Grade Magnet Estimate as % of Mean Annual Gain | 29.3% | 30.4% | 32.5% | 32.5% |
| Magnet Effect translated into weeks of instruction | 10.3 | 10.6 | 11.4 | 11.4 |

*Note.* National norms used from MDRC technical report released in 2007 by Carolyn J. Hill; Annual gain for reading is calculated from seven nationally normed tests: CAT5, SAT9, TerraNova-CTBS, MAT8, TerraNova-CAT, SAT10, and Gates-MacGinitie. Annual gain fo math is calculated from six nationally normed tests: CAT5, SAT9, TerraNova-CTBS, MAT8, Terra Nova-CAT, and SAT10; average weeks of instruction is found by multiplying the figures in row two by 35 (number of weeks in typical school year).

# CHAPTER V

## COMPARATIVE ANALYSIS OF NON-EXPERIMENTAL ESTIMATORS OF THE MAGNET SCHOOL IMPACT

The objective of this chapter is to investigate the accuracy of the non-experimental estimators of the selective magnet school's effect on student math and reading achievement. The non-experimental methods investigated herein include: multiple regression with observed covariates, analysis of covariance with student fixed effects, and propensity score matching. The accuracy of these estimators is evaluated by comparing their respective estimates of the magnet school's effect to the experimental IV estimates presented in the previous chapter.

This analysis seeks to illustrate the consequences of a typical situation where a researcher wants to determine the causal effect of a program, but is unable to leverage experimental data to do so. Therefore, the researcher is compelled to use a non-experimental method on observations of program participants and non-participants. If the researcher expects self-selection into the program causes participants to differ from non-participants in unobserved ways that associate with their potential outcomes, he is left with a high degree of uncertainty as to the accuracy of the non-experimental estimates. If his non-experimental findings are incorrect, they may lead to incorrect policy decisions and program actions by key decision-makers.

To model this situation in the context of the magnet school evaluation, the non-experimental estimators are run under the assumption the information from the randomized admissions lottery does not exist. I assume I do not know which students participated in the admissions lottery for the selective magnet school and therefore I do not know which students won or lost the lottery. I only know which students enrolled in the selective magnet school in 5th and 6th grade.[13]

With the exception of the data on lottery participation, all other data that were available for the experimental analyses are available for the non-experimental analyses, including student demographics and 4th grade achievement in math and reading. The non-experimental estimators use data from the same years as the experimental estimator (2000-2005) and estimate effects of the selective magnet for the same grades (5th and 6th). Table 11 presents the number of student observations in each year for the selective magnet school students and the students enrolled in non-selective schools in the district. These are the observations that the non-experimental estimators may utilize in the construction of their respective comparison groups.

It is important to note that the same 1,462 observations of the magnet students that were used in the experimental evaluation are used as the treatment group for each non-experimental estimator. Therefore, any bias in the non-experimental estimates stems from differences in their non-experimental comparison group and the experimental comparison group. As with the experimental evaluation, this analysis is restricted to students who were enrolled in the district in 4th grade.

[13] It is important to note that while the non-experimental estimators do not have indicators of lottery status at their disposal, the observations of the lottery losers, as well as the lottery winners who did not enroll in the magnet, are still present in the data and may be used in the comparison groups of the non-experimental estimators.

Table 11.

*Student Observations Available to Non-Experimental Estimators*

| | 5th Grade Students | | 6th Grade Students | |
|---|---|---|---|---|
| | Non-Selective Schools | Selective Magnet | Non-Selective Schools | Selective Magnet |
| 2000 | 3,548 | 126 | | |
| 2001 | 4,289 | 146 | 3,315 | 121 |
| 2002 | 4,359 | 154 | 3,944 | 135 |
| 2003 | 4,288 | 168 | 4,115 | 147 |
| 2004 | 4,609 | 153 | 4,027 | 160 |
| 2005 | | | 4,095 | 152 |
| Total | 21,094 | 747 | 19,496 | 715 |

*Note*. Sample limited to 5th and 6th grade students who were enrolled in the district in 4th grade.

*Pre-Specification of Non-Experimental Estimators*

I made a priori specifications of the non-experimental estimators and did not revise the specifications based on how their results compared to the experimental results. This approach is critical to the validity of this analysis because in a real world situation a researcher would not have the luxury of comparing the performance of the non-experimental estimators to experimental estimates and revising their analysis accordingly. This analysis would be uninformative to the accuracy of the non-experimental estimators if we were to test various specifications until we found the one that performed best (Bloom et al., 2002).

It is equally important to note that my specification of the non-experimental efforts is not aimed at demonstrating their weaknesses for estimating causal effects. One might be inclined to do so if their agenda was to advocate for experimental designs and discourage observational studies. As previously discussed, there is high demand from the education research community for empirical evidence on the merits of non-experimental methods versus experimental methods. I seek to demonstrate the accuracy of non-experimental methods that are appropriately specified given the nature of the data. Therefore, the bias (or lack thereof) in the non-experimental estimators can be fairly attributed to the required assumptions of the estimators and not incorrect specifications.

*The Fundamental Evaluation Problem*

It is helpful to frame our discussion of the various estimators using Rubin's Causal Model, where each student is assumed to have two possible outcomes, $Y_1$ and $Y_0$. $Y_1$ is observed if the student attends the magnet school (D=1) and $Y_0$ is observed if the student does not (D=0). The fundamental evaluation problem arises because we cannot jointly observe $Y_1$ and $Y_0$ for a given student, and consequently we cannot directly observe the magnet school's effect on each student as $\delta = Y_1 - Y_0$.

Given the magnet effect cannot be observed for individual students, we are compelled to estimate the average effect of the magnet school on a population of students. In this study we are interested in the average effect of the magnet school on the performance of those students who attend it. This parameter is known as the effect of treatment on the treated (ATT), defined as: $ATT = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)$

The fundamental problem of causal inference is still present in the estimation of the ATT. We are able to observe the achievement of magnet students when they are enrolled in the magnet school: $E\ (Y_1|D=1)$, but we do not know the achievement of these students had they not attended the magnet school: $E(Y_0|D=1)$.

To overcome this problem, we must employ a method that uses data on non-magnet students to estimate the achievement of the magnet students had they not attended the magnet school. These methods rest on the assumption that student's selection into the magnet school is independent of a student's future achievement.

Our "gold standard" method for satisfying this assumption is to use the experimental conditions created by the admissions lottery, where lottery participants are randomly assigned as lottery winners (Z=1) or lottery losers (Z=0). The lottery randomization ensures that lottery assignment is independent of future outcomes: $E(Y_0|Z=1) = E(Y_0|Z=0)$, but it does not ensure magnet enrollment is independent of Y because not all lottery winners enroll in the magnet school . Therefore, we use Z as an instrument for D and estimate the ATT for the subset of students who complied with their lottery assignment, known as the local average treatment effect (LATE). This randomization allowed us to assume $E(Y_0|D=1|Z=1) = E(Y_0|D=0|Z=0)$, and subsequently estimate an unbiased LATE as: $E(Y_1|D=1,Z=1,X) – E(Y_0|D=0,Z=0,X)$, where X is a vector of pre-existing observed covariates included to improve the precision of the LATE estimate, although they are theoretically unnecessary because they are independent of Z in large samples.

Our non-experimental methods do not have the advantage of random assignment to allow for the independence of D and Y, consequently they have to impose additional assumptions to allow for an unbiased estimation of the ATT. If these assumptions fail, the non-experimental estimates will suffer from selection bias, defined as:

$$B(ATT) = E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X) \qquad (6.1)$$

What follows is a discussion of the assumptions that must hold for each non-experimental estimator to produce unbiased estimates of the ATT as well as our formal specification for each estimator. A summary of the key concepts underlying each non-experimental estimator is presented in table 12. After I run the non-experimental estimators, I am able to empirically test if these assumptions hold by comparing their estimates of the magnet effect to those of the experimental IV estimator.

Table 12.

*Summary of Key Concepts of Non-Experimental Methods*

| Non-Experimental Method | Summary of Estimation Strategy | Inference of Causal Effect | Description of Comparison Group | Strengths | Limitations |
|---|---|---|---|---|---|
| Multiple Regression with Observed Covariates | Linear regression of achievement on indicator of magnet school attendance, while controlling for student demographics and 4th grade achievement as well as cohort effects, grade effects, year effects, and test effects (grade by year). | Average Treatment Effect (ATE) for 5th and 6th grade students in the district | All 5th and 6th grade students in non-selective schools in the district who were enrolled in the district in 4th grade | Most efficient estimator of the magnet effect (least variance) if unbiased; Strong statistical power via large sample properties | Estimates will be biased if selection to magnet is based on unobserved differences between magnet students and non-magnet students in district |
| Analysis of Covariance with Student Fixed Effects | Linear regression of achievement on indicator of magnet school attendance that includes individual student indicators (student fixed effects) as well as cohort effects, grade effects, year effects, and test effects (grade by year). The magnet effect is found as the average difference between each student's achievement when enrolled in the magnet school and achievement when enrolled in a non-selective school in the district. | Average effect of Treatment on the Treated (ATT) | The 4th, 5th, and/or 6th grade observations of magnet students when they were enrolled in a non-selective school in the district. | Will resolve selection bias if selection bias stems from unobserved student factors that do not vary over time. | Estimates will be biased if selection to magnet is based on unobserved factors that vary within a student over time. Standard errors can be large because estimates are based on variation within individuals rather than across the sample. |
| Propensity Score Matching with 1 to 1 Nearest Neighbor Matching with Heckman Difference-in-Difference Estimator (PSM NN) | Each 5th grade student's propensity for enrolling in the magnet is predicted based on student characteristics and 4th grade achievement (models run separately for each year). Then each magnet student is matched to one non-magnet student in the district with the closest propensity score. Effect is estimated as the difference in 4th to 5th and 5th to 6th grade gains of magnet students to the matched comparison group. | Average effect of Treatment on the Treated (ATT) | Comparison group with same number of 5th and 6th grade observations as the magnet school sample in each year; student characteristics and prior achievement of comparison group are balanced with the sample of magnet school students | Will resolve selection bias if selection to the magnet school is due to observed characteristics that can be balanced in the comparison group | Will not resolve selection bias if it stems from unobserved differences between magnet students and non-magnet students; One to one matching is inefficient in that it uses only one observation in the comparison group to estimate the potential outcome of a treatment participant. |
| Propensity Score Matching with Local Regression Matching with Heckman Difference-in-Difference Estimator (PSM LRM) | Each 5th grade student's propensity for enrolling in the magnet is predicted based on student characteristics and 4th grade achievement (models run separately for each year). Then each magnet student is matched to a support group of non-magnet students whose combined observation weight sums to one. The weight is determined via a kernel density estimation. Effect is estimated as the difference in 4th to 5th and 5th to 6th grade gains of magnet students to the matched comparison group. | Average effect of Treatment on the Treated (ATT) | Comparison group with the same number of weighted 5th and 6th grade observations in each year, where student characteristics and prior achievement of comparison group are balanced with the sample of magnet school students | Will resolve selection bias if selection to the magnet school is due to observed characteristics that can be balanced in the comparison group; more efficient than one to one estimate because it allows a weighted composite of multiple observations of comparison group members to estimate the potential outcome of a treatment participant. | Will not resolve selection bias if it stems from unobserved differences between magnet students and non-magnet students. |

*Multiple Regression with Observed Covariates*

The first non-experimental method I evaluate is multiple regression with observed covariates (MR). The MR estimator is specified as follows:

$$Y_{igt} = \beta_0 + \delta_g D_{igt} + \beta_x X_{igt} + \beta_c C_i + \lambda_g + \gamma_t + \eta_{gt} + \varepsilon_{igt} \qquad (6.1)$$

Where Y is the standardized test scale score in math or reading for student *i* in grade *g* (5th or 6th) in year *t* (2000-2005). *D* is equal to one when student *i* is enrolled in the selective magnet in grade *g* and year *t* and 0 otherwise. The effect of the selective magnet school on achievement, $\delta_g$, is allowed to vary for 5th and 6th grade. *X* a vector of explanatory variables indicating student participation in special programs (FRL, ESL, special education), 4th grade achievement in reading and math, race (Black =1), and gender (female =1). We impose constant effects of *X* over grades and time. *C* is an indicator of student *i*'s cohort. $\lambda_g$ are grade fixed-effects, $\gamma_t$ are year fixed-effects, and $\eta_{gt}$ is a year by grade interaction to control for changes in the test scale from one year to the next and across grades. $\varepsilon_{igt}$ is the error term, which we decompose as: $\varepsilon_{igt} = \alpha_i + u_{it}$ . Where $\alpha_i$ are unobserved factors of students that are time-invariant and $u_{it}$ are unobserved factors that vary within an individual over time.

The known advantage of least squares multiple regression is that it is the most efficient estimator if certain assumptions hold (linearity, error homoskedasticity, and zero-conditional mean) and the independent variables are exogenous. For MR to produce an unbiased estimate of the magnet effect, we must assume there are no unobserved factors that explain enrollment in the magnet school after accounting for X (in addition to the other regressors). Formally, this assumption is: $E(\varepsilon_{igt}|D_{igt},X_{igt}) = 0$. This assumption can be decomposed into two assumptions that correspond with the two components of the error term:

(1) $E(\alpha_i|D_{igt},X_{igt}) = 0$ states that enrollment in the magnet school is independent of unobserved student factors that associate with future achievement and are time-invariant

(2) $E(u_{it}|D_{igt},X_{igt}) = 0$ states that enrollment in the magnet school is independent of unobserved student factors that associate with future achievement and are time-variant.

I suspect $E(\alpha_i|D_{igt},X_{igt}) = 0$ is untenable in the MR model and the estimates of magnet effect will be upwardly biased due to selection on time-invariant unobservables. The magnet school sample is composed of students who voluntarily sought admission to the magnet. In contrast, the MR comparison group is mainly composed of individuals who did not seek admission to it. Intuitively, this points to fixed differences between the magnet students and the MR comparison group on unobserved dimensions that will affect future achievement. For example, students who seek out the selective magnet may be more likely to have parents who place attach a high value on education, as evidenced by their pursuit of better educational options. Parents' value of education is an unobserved factor that likely does not change over time and has an independent positive effect on student achievement. If this is the case, the MR estimate will be upwardly biased because we are ascribing the effect of the differences in parental value of education between the magnet and non-magnet students to the magnet school "treatment".

$E(u_{it}|D_{igt},X_{igt}) = 0$ is a weaker assumption, but there are still plausible circumstances under which it will not hold in the magnet school sample. For example, it may be that students who seek out the magnet school were more likely to have experienced negative “shocks” to their 4th grade achievement. These abnormal dips in 4th grade achievement may stem from a variety of idiosyncratic factors, such as a particularly bad experience with a teacher or changes in home life circumstances. The key commonality is that these events do not lead to permanent changes in students’ learning trajectories. Families of 4th graders who experience these shocks may be more likely to seek out the magnet school as a remedy to the sudden performance dip. If this is the case, the estimate of the magnet effect via MR would be upwardly biased, because we would expect students who experienced a 4th grade “shock” to regress to their mean performance trajectory in future years.

Bias due to negative (or positive) shocks may be negligent because 5th grade is a normal transition year in the district; most students are moving from K-4 elementary schools to 5-8 middle schools. Therefore, we can expect the decision to seek out the selective magnet for most families does not stem directly from something that happened in 4th grade, rather it is part of the normal decision-making process that families go through as their children transition from elementary to middle school and they seek out the best available schooling options.

The MR model is run on 42,052 observations of 5th and 6th grade students who were enrolled in the district in 4th grade. The non-magnet comparison group is comprised of all 5th and 6th grade students in the district who are enrolled in non-selective middle schools.[14] Note the

[14] This MR estimator could be improved by restricting the sample to those students who met the magnet school admission criteria, but I do not have the 3rd grade test scores or student report card grades that were the basis for admission for roughly 30% of the students in the sample. As an alternative to this approach, I ran models that restricted the sample to only those non-magnet students who had 4th grade scores above the lowest 4th grade score in the magnet school. This restriction only eliminated 12% of the non-magnet students and did not substantively alter the MR estimates of the magnet effect.

observations are from the same years as those used in the experimental analysis and the sample of magnet students is the same as that used to estimate the experimental IV effects.

*Analysis of Covariance with Student Fixed Effects*

The second non-experimental estimator I evaluate is the analysis of covariance with student fixed-effects. This estimator is also referred to as the "within" student estimator (Baltagi, 1995) as it includes a separate intercept term for each student in the sample.

The student fixed-effects model takes the following form:

$$Y_{igt} = \alpha_i + \delta_g D_{igt} + \lambda_g + \gamma_t + \eta_{gt} + u_{it} \qquad (6.2)$$

Where $\alpha_i$ are the student fixed-effects that capture all factors that are time-invariant for an individual. $\alpha_i$ captures the observed characteristics that are fixed over time, which explains the absence of the vector of time-invariant student characteristics, *X*, and the cohort indicators, C, found in the MR model. The advantage of the fixed-effects estimator is that $\alpha_i$ also captures the unobserved factors that are time-invariant for an individual and associate with magnet enrollment. This offers a solution to the problem of self-selection if the unobserved factors that determine selection into the magnet school are time-invariant and therefore controlled by $\alpha_i$.

It is possible to include $\alpha_i$ in the estimator because magnet status, *D*, is time-variant and thus not subsumed by $\alpha_i$. That is because there are observations of students from when they were enrolled in the magnet school and when they were not enrolled in the magnet school.[15] Specifically, we have observations of 747 students who went from "non-magnet" status in 4th

[15] The FE method has been used in a number of magnet and charter school evaluations (see for example, Rouse, 1998; Bifulco & Ladd, 2006; Sass, 2006; Hanushek et al. 2006).

grade to enrolling in the selective magnet school in 5th grade. In addition, there are 31 students who enrolled in the magnet school in 5th grade, but then transferred back to a non-selective school in the district in 6th grade. This "within-student" variation in magnet school enrollment allows us to difference out each student's fixed achievement from the effect on achievement that is due to their enrollment in the magnet school.

It is useful to compare the assumptions of the student fixed-effects estimator to those of the MR estimator. Recall MR requires two assumptions: (1) selection into the magnet does not stem from unobserved time-invariant factors: $E(\alpha i|D,X) =0$; (2) selection into the magnet does not stem from unobserved time-variant factors: $E(uit|D,X) =0$. Including student fixed-effects, $\alpha i$, only requires the second assumption to hold in order to yield a unbiased estimate of the causal effect of magnet school enrollment.

Therefore, the student fixed-effects estimator is able to ignore bias stemming from unobserved differences between magnet school students and non-magnet school students that remain constant over time. This eliminates our previously discussed concern over differences in parental motivation or value of education if these factors do not change over time. However, bias due to unobserved factors that vary within an individual over time remains a concern.

The student fixed-effect estimator uses observations of 64,265 students in grades 4 through 6. The number of observations in this model is substantially greater than the number of observations in the MR model because students' 4th grade achievement scores enter as outcomes, whereas they were used as covariates in the OLS estimator. The important point is that the same information on student achievement is used in both regression models.

*Propensity-Score Matching*

The third non-experimental estimator uses propensity score matching (PSM) to establish a comparison group for the magnet school students. PSM has gained popularity in recent years as a non-experimental method for program evaluation (See, for example, Diaz & Handa, 2006; Dehejia & Wahba, 2002; Smith & Todd, 2005; Agodini & Dynarski, 2004). The technique, which was first proposed by Rosenbaum and Rubin (1983), improves upon conventional matching methods because subjects can be matched on one “propensity score”, which represents the probability of selection into treatment given a set of variables, rather than on multiple variables.

The objective of PSM is to establish a comparison observation for each treatment observation that has the same predicted probability for selection into treatment. The outcomes of the treatment group are then compared to the matched comparison group to estimate the ATT.

To yield an unbiased estimate of the ATT, PSM requires the conditional independence assumption to hold: $E(Y_0 \mid D = 1, X) = E(Y_0 \mid D = 0, X)$. Put differently, it requires that the factors that predict selection into treatment and affect future outcomes are observed.

Note this is the same assumption required of the OLS estimator. However, in the PSM estimator, we restrict the causal inference to a comparison sample with the same observed propensity for selection into the magnet school as the actual sample of magnet school students. This property is the main advantage of the PSM estimator over the OLS estimator. While both require conditional independence, PSM only requires conditional independence within a sample that includes a comparison group that has the same predicted probability of selection into the magnet school. By balancing the observed covariates that predict selection into the magnet school, PSM assumes we are also able to balance the unobserved factors that predict selection into the magnet school. In contrast, MR uses a sample of 5th and 6th grade students that likely includes some students who have no likelihood of enrolling in the magnet school. Therefore, conditional independence is a stronger assumption in MR.

Nevertheless, if PSM is unable to balance unobserved differences between magnet and non-magnet students, the assumption of conditional independence will be violated and the PSM estimate of the ATT will be biased.[16]

An additional required assumption of PSM is that the probability of treatment is not identified by any single conditioning variable *x*: $0 < \Pr(D = 1 \mid x) < 1$. In the magnet evaluation, this requires that at each level of *x*, it is possible to observe both magnet attendees and non-magnet attendees. If the probability of magnet attendance is equal to one at certain levels of *x*, it will not be possible to observe a control group. This problem would arise if, for example, there were no ESL students in the non-magnet sample, in which case Pr(D=1|x) =1 and it would not be possible to find a non-magnet student to match with an ESL magnet student.

If the above condition is met, matching can performed on a single index, the “propensity score”, which represents the probability of treatment given a vector of conditioning variables X: *P(X)* = Pr(D=1|X). This is the benefit of PSM versus other matching techniques.

*Selection of the PSM conditioning variables.* The first step in our specification of the PSM estimator is the selection of conditioning variables (X). PSM only requires one to match the treatment and comparison groups on characteristics that affect probability of treatment. Therefore, the conditioning variables used to estimate the propensity score were identified as those that were statistically significant ($p$ < 10) predictors of enrollment in the magnet. Table 13

[16] Pearl (2009) argues that PSM may actually exacerbate selection bias in situations where treatment assignment is ignorable in the comparison of unadjusted means of a treatment group and comparison group. The process of balancing the observed covariates within each stratum of the treatment and comparison groups leads to the unobserved factors that were originally balanced in the raw samples to be shifted out of balance.

reports the seven variables that were statistically significant predictors of magnet enrollment in any of the six years of data (2000-2005).[17] The averages of these seven variables for the magnet school students are presented alongside the averages for the non-magnet students in the district to exhibit the baseline differences between the magnet schools students and the full sample of non-magnet students in the district that are available as potential matches.

Table 13.

*Differences in Conditioning Variables between Magnet and Non-Magnet 5th Grade Students*

| | Magnet | Non-Magnet | Difference |
|---|---|---|---|
| 4th Grade Math | 668 | 619 | 49.0*** |
| 4th Grade Reading | 685 | 631 | 53.4*** |
| FRL | 0.121 | 0.541 | -0.420*** |
| ESL | .103 | .136 | .033** |
| Black | 0.210 | 0.510 | -0.30*** |
| Female | 0.542 | 0.491 | 0.051** |
| Disabled | .137 | .167 | .030* |
| Observations | 747 | 21,094 | |

*p<0.05, ** p<0.01, *** p<0.001

Note. The sample reported here is limited to 5th grade students. Similar differences are found with the 6th grade sample.

*Specification of the propensity score model.* In this study, propensity scores $\hat{P}(X_{i,j})$ are calculated for each 5th grade magnet and non-magnet student in the district using a logistic regression model where the dependent variable *D* is equal to 1 if the student attends the selective magnet school and 0 otherwise. The propensity score for a given student is therefore equal to

[17] These are the same seven student-level variables included in the estimation of the IV and OLS models.

the predicted log of the odds of attending the selective magnet school to attending a non-selective magnet school.

The logistic regression model for generating the propensity score is shown in model 6.3. My notation in the PSM exposition departs slightly from the notation above by designating magnet students as *i* and non-magnet students as *j*. In the logistic regression *X* consists of seven conditioning variables: (1)  the 4th grade math test score for each magnet student *i* or non-magnet student *j*,(2) student's 4th grade reading score, (3) an indicator if the student is black, (4)  and indicator if the student is female, (5) an indicator if the student participates in the free/reduced lunch program, (6) an indicator if the student receives ESL support, and (7) and an indicator if the student receives special education services.

$$\hat{P}(X_{i,j}) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{i,j} + e_{i,j} \qquad (6.3)$$

Model 6.3 is run separately for each 5th grade cohort using a cross-section of data from their 4th grade school year.  This allows us to match each 5th grade magnet student to a 5th grade non-magnet student who was in the same grade in the same year, which in turn allows us to balance cohort effects, grade effects, year effects, and test effects (grade by year interactions) when we pool the six years of data to estimate the magnet school effect.

*PSM Method #1: One to One Nearest Neighbor Matching.*  After the propensity score calculation, I use two distinct methods for matching magnet school students to non-magnet school students and estimating the magnet effect.  The first method is referred to as one-to-one nearest neighbor matching.  One-to-one nearest neighbor matching involves matching each treatment subject to the comparison group subject with the closest propensity score, such that $\hat{P}(X_i) - \hat{P}(X_j)$ is minimized.  We allow for "replacement", which means a comparison group member may serve as a nearest neighbor match for more than one magnet school student if it allows $\hat{P}(X_i) - \hat{P}(X_j)$ to be minimized.

The one-to-one nearest neighbor matching is done under two different restrictions on the samples of non-magnet students that are available for support. Under the first restriction, magnet students are allowed to match to any non-magnet student in the district such that $\hat{P}(X_i) - \hat{P}(X_j)$ is minimized. I term this *inter-school* matching because I allow magnet students to be matched to non-magnet students outside of their 4th grade school.

Under the second restriction, we restrict matching to within students' 4th grade schools. Each 5th grade magnet student is matched to a 5th grade non-magnet student who attended the same 4th grade school. I refer to this sample restriction as *intra-school* matching. The first restriction (inter-school matching) increases the likelihood the seven conditioning variables will balance between treatment and comparison groups because the support available for a given magnet student is larger. However, it ignores the information signaled by a student's 4th grade school, which may help balance the unobserved characteristics between magnet and non-magnet students. We use both restrictions to see which produces the least amount of bias in the estimate of the magnet effect.

After the inter- and intra-school nearest-neighbor matching procedures are done for the 5th grade students, we estimate the magnet school impact on achievement using a variation of the difference-in-differences method proposed by Heckman et al. (1998). The estimate of the magnet effect on 5th grade achievement is found as the average difference between the change in achievement (mathematics or reading) from 4th to 5th grade for the magnet students and the change in achievement from 4th to 5th grade for the matched comparison group. Likewise, the estimate of the magnet effect in 6th grade is found as the average difference between the change in achievement from 4th to 6th grade for the magnet students and the change in achievement from 4th to 6th grade for the matched comparison group.

The equation for this estimator is shown below, where *g* indicates either 5th or 6th grade:

$$\hat{\delta}_g = \frac{\sum_i \{(Y_{ig} - Y_{ig=4}) - (Y_{jg} - Y_{jg=4})\}}{n_{ig}} \tag{6.4}$$

*PSM Method #2: Local Regression Matching.* The second PSM method uses local regression matching (LRM) to match magnet students to comparison group students. A number of researchers have posited that LRM is a more efficient matching method than nearest-neighbor because it enables each treatment participant to be matched to multiple observations in the control group, whereas nearest-neighbor restricts matching to one-to-one (see for example, Heckman, Ichimura, and Todd 1997, 1998; Black and Smith, 2004.) The one-to-one restriction may lead to incomplete or inexact matching if there are no subjects in the support group with propensity scores close to those of the treatment subjects. This may cause treatment observations to be excluded for lack of a good match or cause observations to be matched poorly.

The same logistic regression for estimating $\hat{P}(X_{i,j})$ is used for LRM matching. Under LRM, each magnet student is matched to all non-magnet students with propensity scores that fall within a designated window of their own score, this is referred to as a common-support group. The outcomes of the magnet student are then compared to a weighted average of the common-support group, where the weights of each common-support group member are set based on the proximity of their propensity score from the propensity score of the magnet student. The amount each member contributes to the mean of the common-support group is based on their propensity score –the closer $\hat{P}(X_j)$ is to $\hat{P}(X_i)$ the greater the weight the non-magnet student contributes. The weight *W* for each non-magnet student *j* is calculated using a kernel function, where *G* is a kernel function and θ is a bandwidth parameter (i.e. the window). This is shown in model 6.5:

$$W_{ij} = \frac{G\left(\frac{P_j - P_i}{\theta_n}\right)}{\sum G\left(\frac{P_k - P_i}{\theta_n}\right)} \tag{6.5}$$

As with the nearest neighbor procedure, I create two comparison groups with the LRM. The first allows for inter-school matching and the second is restricted to intra-school matching.

After the weighted mean is calculated for each magnet student's common support group, the 5th grade magnet effect is calculated as the average difference between the change in the achievement scores from 4th grade to 5th grade for the magnet students *i* and the change in the weighted achievement scores from 4th to 5th grade for the common-support groups. Likewise, the 6th grade effect is calculated as the difference in the change in achievement from 4th to 6th grade for magnet students and the change in 4th to 6th grade achievement for the common-support groups. The equation for this estimator is presented below:

$$\hat{\alpha}_g = \frac{\sum_i \{(Y_{ig} - Y_{ig=4}) - \sum_j W_{ij}(Y_{jg} - Y_{jg=4})\}}{n_{ig}} \tag{6.6}$$

The LRM inter-school sample includes 36,872 non-magnet attendees compared to the 1460 magnet students. The LRM intra-school sample includes 19,390 non-magnet attendees because the sample is limited to only those students within the same 4th grade schools as the magnet school students.

*Balancing tests of PSM samples.* Rosenbaum and Rubin (1985) decompose the bias that may arise in matching estimators into three components: (1) bias due to incomplete matching, which results when some treatment subjects are discarded because adequate matches do not exist in the comparison group; (2) bias due to inexact matching, which results when the characteristics of the treatment group differ from those of the matched comparison group, and (3) bias due to selection on unobservables.

The first two components of this bias are under the control of the researcher via their specification of the PSM estimator, while the third bias – selection on unobservables – is unknown to the researcher and thus a required assumption of the PSM estimator. In this analyses, we are only interested in the bias that arises because of selection on unobservables. Therefore, it is important to test how well our matching procedures have controlled for the other two types of bias. If our methods are unable to account for the first two types of bias, it indicates a misspecification of the PSM method and signals to the researcher that PSM estimates will be biased. A careful researcher would then decide to either re-specify the PSM model or utilize a different analytic approach. Our interest is in understanding how well the estimators perform when they are used appropriately. Therefore, it is critical to test to determine that the bias we find in the PSM estimates of the magnet effect are only due to selection on unobservables and not due to the inappropriate use of PSM.

The first two components of bias both stem from the inability of the PSM method to find suitable matches for the treatment subjects. In this study I make the a priori decision not to exclude magnet school students from the analyses if I cannot find suitable matches because my objective is to use the same treatment group for all estimators and only vary the comparison group. Therefore we are concerned only with the second form of bias that results from poor matching. To measure this bias we conduct a series of post-matching balancing tests. The common goal of the balancing tests is to determine if the propensity score serves to balance the distribution of the covariates in the treatment and comparison groups. Formally, it seeks to confirm that: pr(X|D=1, P(X)) = pr(X|D=0,P(X)). If this condition is satisfied, we can assume any bias we estimate via our comparison with the experimental estimates is due to selection on unobservables and not due to misspecification of the PSM estimator.

For each of the four matched samples (Nearest Neighbor Inter-School, Neighbor Intra-School, LRM Inter-School, LRM Intra-School), we conduct two balancing tests that are common in the PSM literature: (1) test for equality of means after matching (Rosenbaum and Rubin, 1985), and (2) test of joint equality of means in the matched sample (Smith and Todd, 2005). The *Test for Equality of Means after Matching* was proposed by Rosenbaum and Rubin (1985) and tests for the equality of each covariate across the comparison groups. If statistically significant mean differences are found, it signals that the samples are unbalanced and requires the re-specification of the PSM estimators. The *Test of Joint Equality of Means in the Matched Sample* was proposed by Smith and Todd (2005) and conducts a test of the null hypothesis that the vector of means among the seven covariates are equal for the magnet and matched comparison sample. This is a test of the joint balance of the conditioning variables that takes the form of an F-test, or Hotelling test.

The two tests were run on the separate matched samples of 5th and 6th grade. The results are presented in tables 14 and 15. Both the intra- and inter-school Nearest Neighbor (NN) matching methods satisfy the two balancing tests. However, neither LRM method was able to achieve balance when all seven conditioning variables are used in the model. Note the significant differences in mean 4th grade math and reading scores. Consequently, the propensity score logistic regression model for the LRM estimators is re-specified to match on fewer conditioning variables. I prioritize balancing 4th grade math and reading scores in the LRM model and find I am only able to achieve balance in the LRM matched samples when I base the propensity score exclusively on 4th grade math and reading performance.[18]

Table 14.

*Balancing Tests of PSM Methods: Differences between Magnet School Mean and Matched Comparison Group Mean for 5th Grade Sample*

| | NN Intra | NN Inter | LRM Intra | LRM Inter |
|---|---|---|---|---|
| ESL | -0.003 | -0.004 | -0.010 | -0.000 |
| Black | -0.018 | -0.017 | 0.024 | 0.038** |
| Female | 0.005 | -0.005 | -0.003 | -0.005 |
| Free or Reduced Price Lunch | -0.017 | -0.012 | 0.0491*** | 0.062*** |
| 4th Grade Reading Score | -1.64 | -0.496 | -6.886*** | -6.65*** |
| 4th Grade Math Score | -1.82 | 0.110 | -7.400*** | -5.875*** |
| Disabled | -0.026 | -0.021 | -0.025* | -0.019 |
| Hotelling Test F- Statistic | 1.565 | 1.5477 | 4.83*** | 3.841*** |
| Weighted Observations | 1494 | 1494 | 1494 | 1494 |
| Unweighted Observations | 1394 | 1396 | 20,158 | 10,999 |

*p<0.05, ** p<0.01, *** p<0.001

[18] I ran the results with the imbalanced samples matched on all seven conditioning variables as well as the samples that achieve balance with 4th grade math and reading scores. The results are almost exactly the same, falling within three-tenths of a scale score in all cases.

Table 15.

*Balancing Tests of PSM Methods: Differences between Magnet School Mean and Matched Comparison Group Mean for 6th Grade Sample*

| | NN Intra | NN Inter | LRM Intra | LRM Inter |
|---|---|---|---|---|
| ESL | 0.008 | 0.011 | 0.005 | 0.012 |
| Black | 0.02 | -0.001 | -0.05 | -0.047 |
| Female | -0.015 | -0.008 | 0.01 | 0.01 |
| Free or Reduced Price Lunch | -0.017 | -0.014 | -0.102 | -0.095 |
| 4th Grade Reading Score | 1.95 | 0 | 7 | 12 |
| 4th Grade Math Score | 1.81 | -1 | 5 | 10 |
| Disabled | 0.012 | 0.014 | 0.016 | 0.038 |
| Hotelling Test F- Statistic | 1.5923 | 0.7790 | 5.913*** | 5.53*** |
| Weighted Observations | 1426 | 1426 | 1426 | 1426 |
| Unweighted Observations | 1290 | 1288 | 18,174 | 9851 |

*p<0.05, ** p<0.01, *** p<0.001

*Estimates of Bias in Non-Experimental Estimators*

Tables 16 and 17 present the non-experimental estimate alongside the experimental IV estimate. With the non-experimental estimates of the magnet impact in hand, the next step is to estimate the amount of bias that results from each non-experimental estimator *k*. Bias in the non-experimental estimators can be defined as the difference between the true causal effect $\delta$ and the non-experimental estimate of the causal effect $\hat{\delta}_k$. This bias cannot be directly observed because $\delta$ is unknown. However, the bias can be estimated when one has access to estimates from an experimental estimator *z*, if the experimental estimates of the treatment effect are themselves unbiased estimates of $\delta$: $E(\hat{\delta}_z) = \delta$.

Using the experimental IV estimates $\hat{\delta}_z$, I estimate the bias in the non-experimental estimates as the difference in the non-experimental estimates and the experimental IV estimate:

$\hat{b}(\delta_k) = \hat{\delta}_k - \hat{\delta}_z$. These estimates will be unbiased conditional on the assumption that the experimental IV estimates are not biased due to selective attrition. Table 18 presents the bias in the non-experimental estimates in three metrics. The first metric presents the bias in test scale scores. The second form presents a standardized measure of the bias by translating it into an effect size – found by dividing $\hat{b}(\delta_k)$ by the pooled standard deviation of the annual test score gain among lottery participants. The final form presents the approximate number of weeks of instruction that the bias estimates represent. This is done by dividing the average annual scale score gain among lottery losers by the conventional number of instructional weeks in a school year (35). The bias estimate is then divided by the measure of learning per week to find the total number of weeks of learning that the bias represents. This is a relevant indicator for education decision makers who want to consider the practical importance of a program. It helps ascertain the extent to which the non-experimental methods may lead an education decision-maker to incorrectly change policy based on a biased estimate.

Table 18 reveals that most of the non-experimental estimates have substantial bias. Bias is largest in the MR estimates. The magnitude of the MR estimates relative to the experimental estimate suggests that MR has failed to adequately control for selection on unobservables. The bias in the MR estimates represents close to a full academic year's worth of learning in reading and over half a year's worth of learning in math.

Propensity score matching techniques performed better than MR, suggesting that creating a comparison group with balanced covariates reduces the threat of selection bias better than using the full sample of students via MR. Nevertheless, the PSM estimates still have practically important levels of positive bias. The fact that all estimates are large and positive suggests the PSM is unable to effectively address the form of selection bias that exists between magnet and non-magnet students by balancing observable characteristics. The similarity of the estimates corroborates the claims of many researchers that the choice of matching method makes little difference in the impact estimates (Bloom, 2002). In general, restricting matching to within student's 4th grade school produced less biased estimates than allowing matching across 4th grade schools, although neither intra- or inter-school matching performed well enough to substitute for the experimental IV estimator. Notice from the standardized effect sizes that the bias in the MR and PSM estimates is smaller in math than in reading.

The student fixed-effect estimator performed the best of all estimators, which indicates that some of the bias in the MR and PSM estimators are due to time-invariant unobserved factors. This is most evident in reading, where both the 5th and 6th grade estimates were not meaningfully different from the experimental IV estimates. This allows for the conclusion that the fixed-effects estimator could serve as an unbiased substitute for an experimental design, if the outcome of interest was only reading achievement. The fixed effects math estimates had less bias than the other non-experimental estimators (with the exception of the 6th grade PSM NN Intra-School). Nevertheless, the magnitude of the bias in the 5th grade fixed-effect estimate equates to over a month of instruction and the 6th grade bias equates to over two months.

Table 16.

*Results: Non-Experimental Estimates of Magnet Impact on Math Outcomes*

| | Experimental IV Estimates | Multiple Regression | Student Fixed Effects | PSM NN Intra-School | PSM NN Inter-School | PSM LRM Intra-School | PSM LRM Inter-School |
|---|---|---|---|---|---|---|---|
| 5th Grade Magnet | 4.985*** | 13.181*** | 7.497*** | 8.582*** | 10.925*** | 9.841*** | 11.0717*** |
| | (1.520) | (0.951) | (1.006) | (1.491) | (1.516) | (1.507) | (1.528) |
| 6th Grade Magnet | 0.685 | 12.791*** | 6.850*** | 6.199*** | 6.857*** | 5.527*** | 7.137*** |
| | (1.605) | (1.142) | (1.024) | (2.342) | (2.312) | (1.597) | (3.365) |
| 4th Grade Reading | .169*** | 0.226*** | | | | | |
| | (.017) | (0.007) | | | | | |
| 4th Grade Math | .424*** | 0.568*** | | | | | |
| | (.017) | (0.007) | | | | | |
| Black | -7.341*** | -3.869*** | | | | | |
| | (1.127) | (0.329) | | | | | |
| FRL | -7.071*** | -4.410*** | 0.507 | | | | |
| | (1.294) | (0.315) | (0.371) | | | | |
| ESL | 5.725*** | 2.040*** | 0.559 | | | | |
| | (1.581) | (0.509) | (1.953) | | | | |
| Female | -3.275*** | -1.017*** | | | | | |
| | (.877) | (0.295) | | | | | |
| Disabled | -6.125*** | -6.246*** | -3.754*** | | | | |
| | (1.386) | (0.507) | (0.912) | | | | |
| Constant | 284.7*** | 56.22*** | 592.31*** | | | | |
| | (13.6) | (6.113) | (11.046) | | | | |
| Cohort Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Year Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Test Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 3,616 | 42,052 | 64,265 | 2,916 | 2,916 | | |

*p<0.05, ** p<0.01, *** p<0.001

Table 17.

*Results: Non-Experimental Estimates of Magnet Impact on Reading Outcomes*

| | Experimental IV Estimates | Multiple Regression | Student Fixed Effects | PSM NN Intra-School | PSM NN Inter-School | PSM LRM Intra-School | PSM LRM Inter-School |
|---|---|---|---|---|---|---|---|
| 5th Grade Magnet | 3.760** | 13.118*** | 3.726*** | 9.501*** | 8.984*** | 10.333*** | 10.785*** |
| | (1.361) | (0.850) | (0.982) | (1.340) | (1.336) | (1.355) | (1.368) |
| 6th Grade Magnet | 1.015 | 10.762*** | 1.562 | 8.240** | 8.393*** | 7.278*** | 7.883*** |
| | (1.438) | (0.953) | (1.000) | (3.333) | (3.332) | (3.546) | (2.376) |
| 4th Grade Reading | .494*** | 0.596*** | | | | | |
| | (.015) | (0.007) | | | | | |
| 4th Grade Math | .093*** | 0.172*** | | | | | |
| | (.016) | (0.006) | | | | | |
| Black | -7.591*** | -4.381*** | | | | | |
| | (1.009) | (0.316) | | | | | |
| FRL | -6.389*** | -5.443*** | 0.222 | | | | |
| | (1.158) | (0.298) | (0.362) | | | | |
| ESL | 1.626 | -1.221* | 2.291 | | | | |
| | (1.416) | (0.492) | (1.911) | | | | |
| Female | -.372 | 1.361*** | | | | | |
| | (.785) | (0.281) | | | | | |
| Disabled | 3.539** | -3.515*** | 2.302* | | | | |
| | (1.239) | (0.479) | (0.894) | | | | |
| Constant | 270.6*** | 63.027*** | 604.4*** | | | | |
| | (13.36) | (5.713) | (10.76) | | | | |
| Cohort Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Year Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Test Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 3,616 | 42,251 | 64,265 | 2,916 | 2,916 | | |

*p<0.05, ** p<0.01, *** p<0.001

Table 18.

*Estimates of Bias in the Non-Experimental Estimates of the Effect of Magnet school Attendance on Math and Reading Achievement*

| | Multiple Regression | Student Fixed Effects | PSM NN Intra-School | PSM NN Inter-School | PSM LRM Intra-School | PSM LRM Inter-School |
|---|---|---|---|---|---|---|
| **Math** | | | | | | |
| ***5th Grade*** | | | | | | |
| Scale Scores | 8.20 | 2.51 | 3.60 | 5.94 | 4.86 | 6.09 |
| Effect Size | 0.28 | 0.08 | 0.12 | 0.20 | 0.16 | 0.20 |
| Instructional Weeks | 16.9 | 5.2 | 7.4 | 12.3 | 10.0 | 12.6 |
| ***6th Grade*** | | | | | | |
| Scale Scores | 12.11 | 6.17 | 5.51 | 6.17 | 4.84 | 6.45 |
| Effect Size | 0.45 | 0.23 | 0.20 | 0.23 | 0.18 | 0.24 |
| Instructional Weeks | 18.4 | 9.4 | 8.4 | 9.4 | 7.3 | 9.8 |
| **Reading** | | | | | | |
| ***5th Grade*** | | | | | | |
| Scale Scores | 9.36 | -0.03 | 5.74 | 5.22 | 6.57 | 7.03 |
| Effect Size | 0.35 | 0.00 | 0.22 | 0.20 | 0.25 | 0.26 |
| Instructional Weeks | 27.5 | -0.1 | 16.8 | 15.3 | 19.3 | 20.6 |
| ***6th Grade*** | | | | | | |
| Scale Scores | 9.75 | 0.55 | 7.23 | 7.38 | 6.26 | 6.87 |
| Effect Size | 0.37 | 0.02 | 0.28 | 0.28 | 0.24 | 0.26 |
| Instructional Weeks | 31.2 | 1.8 | 23.1 | 23.6 | 20.0 | 22.0 |

The fixed-effect estimator performed better in reading than in math. A possible explanation for this discrepancy is that unobserved factors that affect reading achievement are time-invariant, such as parental literacy, exposure to reading material outside of school, and students' general enjoyment of reading and writing. These factors would be absorbed by the student fixed-effect and differenced-out of the magnet impact estimate. The unobserved factors affecting math achievement may be time-invariant, such as time spent on homework, parent involvement, and overall student motivation. The fixed-effect would fail to capture these factors and the differences between the magnet and non-magnet students in these dimensions would be attributed to magnet school attendance.

With the exception of the fixed-effect estimator in reading, all the non-experimental estimators predicted a positive effect of magnet school attendance in 6th grade, whereas the experimental IV estimate showed an effect that was not statistically different from zero. In some cases the 6th grade estimates were larger than the 5th grade estimates. This is an important finding, as a decision-maker may incorrectly conclude that the magnet school has a consistently positive effect on students. The experimental data show that most of the 5th grade magnet effect is given back in 6th grade.

*Discussion*

This study has important implications for researchers interested in the causal effect of magnet schools on student achievement. The only estimator that performs well enough to substitute for an experimental design is the student fixed-effects estimator of the magnet effect in reading. All the PSM and MR estimates overstate the causal effect of magnet schools in both reading and math. The bias in these estimates is not trivial, representing months of instruction. Collectively, these findings caution against making policy decisions based on non-experimental evidence. With the exception of the fixed-effects estimator in reading, all estimators may lead a decision-maker to incorrectly conclude the magnet school effect is noteworthy and be compelled to expand the program.

The generalization of these findings beyond this sample should be done with caution. Magnet school policies and practices vary substantially across districts, as do the general contexts and dynamics of public schooling. These differences will influence the effectiveness of the magnet schools relative to the rest of the schools in the district.

One important consideration is the gap in the average student performance of the magnet school and the average student performance in the non-magnet schools in the district. In this study the comparison group's "treatment" is a composite of the non-selective schools in the district. In districts with relatively homogenous schools, one might not find a significant magnet school advantage because the peers and academic programs found in the non-magnet schools may be very similar to those found in the magnet school. This is not the case with the district I studied. In 2004, the average 5th grade math scale score in the selective magnet school was 692, making it the highest performing middle school in the district. In contrast, the "average" school in the district (i.e. the school at the median of the district's distribution in average 5th grade school scores) had an average score of 623. The 69 scale score points that separates these two schools equates to over three grade levels of learning. In other words, the students in the selective magnet school are learning at more than three grades levels from the students in the average school in the district. Such a profound difference between the magnet schools and the non-magnet schools is probably not found in most districts.

It is particularly important to note that these findings may not generalize to other types of schools of choice, namely charter schools. The nature of self-selection into charter schools is probably very different than magnet schools. Unlike charter schools, magnet schools have been normal educational options within urban public school systems for decades. Most parents are aware of the magnet schools in their district and recognize them as viable educational options if their students qualify for admission. Consequently, when a parent submits a magnet school application it does not signal they are particularly different from the other parents in the district. In contrast, charter schools are still a novel educational option in most cities. Many families do not understand them, are unaware of their presence, or view them with skepticism because they are misinformed and assume they charge tuition or have special admissions requirements. Therefore the parents who submit an application to a charter school are likely to be very different from the average family in the district, potentially more so than the parents of magnet school students. This may suggest a greater threat of selection bias for charter school studies that use non-experimental methods, although the direction and magnitude of this bias is not immediately clear.

# CHAPTER VI

# EVALUATING THE EFFECTS OF ATTRITION ON THE EXPERIMENTAL AND NON-EXPERIMENTAL ESTIMATORS OF THE MAGNET IMPACT

## *Selective Attrition and School Choice Research*

Sample attrition poses a methodological challenge to school choice program evaluations that use admissions lotteries to create random comparison groups. Attrition is a problem because participants who do not win entry to their school of choice often seek other alternatives to their residentially-zoned district school. Researchers often do not have the resources or agreements to gather data on these students. If this attrition is non-random and systematically related to the outcomes (i.e. "selective"), the integrity of the experimental design is jeopardized and bias may result. A number of recent school choice evaluations have had to deal with high levels of attrition (see for example, Abdulkadiroglu, Che, & Yasuda, 2009; Buckley & Schneider, 2008). Kemple & Scott-Clayton's (2004) investigation of career academies found roughly 40% of the sample left during the study years.

School choice researchers' decisions on how to deal with attrition have important consequences for their research conclusions. The conflicting findings on Milwaukee's voucher program are a good example of this; three research teams reached three different conclusions primarily because of their different methods for dealing with attrition. Witte, Sterr & Thorn (1996) determined the attrition among voucher lottery losers was so severe as to render them useless as an "experimental" control group and instead created a non-experimental comparison group by drawing a random sample of Milwaukee public school students. Using the non-experimental comparison group, Witte and colleagues did not find a positive effect of private school vouchers on math or reading achievement.

Green, Peterson, and Du (1997) argued Witte's approach did not adequately control for selection bias and conducted a random-assignment evaluation by comparing the achievement of voucher students to students who applied but did not get into a private school. They found statistically positive effects of the voucher program in both reading and math for students in their 3rd and 4th year in the program. Witte challenged these findings on the grounds that 52% of the unsuccessful applicants left the district. He argued this attrition leads to an unfair comparison group for voucher students because the unsuccessful applicants who remained in the district were from less educated, lower income families than the full sample of voucher participants. A subsequent analysis by Rouse (1998) revealed the test scores of unsuccessful applicants who left the district were actually lower on average than those who remained. Using lottery status as an instrument for private school enrollment as well as a student fixed-effect model, Rouse found a positive effect of the voucher program in math, but not in reading.

The Milwaukee voucher research calls attention to the need for more empirical guidance on how to deal with attrition from school choice evaluations. There is a broad consensus among researchers that lottery-based experiments are the gold-standard for estimating causal effects of school choice programs. Nevertheless, it is unclear how robust these experimental estimators are to the effects of selective attrition and when (if ever) the experimental comparison should be abandoned in favor of a non-experimental comparison.

The purpose of this chapter is to provide some empirical guidance on this topic by evaluating the bias in the experimental and non-experimental estimators of the magnet effect under various rates and forms of selective and random attrition. Using simulated data, I examine the performance of the estimators under 30 unique samples that have different rates and forms of attrition among the samples of lottery winners, lottery losers, and lottery non-participants. For parsimony, I focus exclusively on the effect of the magnet school on 5th grade math achievement.

*Sample Set-Up*

The sample used in this exercise consists of all 5th grade students who were enrolled in the district in 4th grade. The real missing outcomes in the 5th grade dataset were simulated to create a dataset that has zero attrition, thus allowing us to assume the experimental IV estimator is unbiased when run on the full sample because there is no threat of attrition bias.[19] Inducing artificial attrition into this sample, where 80% of the observations are real, allows for more realistic inferences than one using a completely artificial sample.

The sample consists of three subgroups that are important to the simulation: (1) lottery winners, (2) lottery losers, (3) lottery non-participants. I examine how the estimators perform under different forms and rates of attrition among these three groups. The first two groups are important because they represent the experimental comparison groups, and attrition within these groups represents a threat to the random assignment.[20] The last group is important because attrition among students who did not participate in the lottery may bias the non-experimental estimates.

*Specifying the Form of Simulated Attrition in the Three Subgroups*

The performance of the estimators is evaluated under six characterizations ("scenarios") of the form of attrition in the sample. I specify different forms of attrition within the three subgroups (lottery winners, lottery losers, non-participants) to fit with plausible scenarios that school choice researchers

[19] To create a full sample of 5th grade students who were enrolled in the district in 4th grade, I had to make some assumptions on the actual attrition in the sample. First, I assume 85% of lottery winners with missing 5th grade outcomes actually would have attended the selective magnet school, while the other 15% would have enrolled in another district school. This is the approximate proportion that was observed among non-attritors in the sample. Second, that the actual attritors' 5th grade outcomes would fall at the same percentile within their cohort's distribution of math scores as was observed in 4th grade. For example, if a student had a 4th grade math score that fell at the 65th percentile of the 4th grade math scores in the district in 2001, that student's missing 5th grade outcome would be equivalent to the 65th percentile of the observed 5th grade math scores in 2002. By extension, this assumes attritors made average growth from 4th to 5th grade, thus allowing them to maintain their normative status in the distribution of Y.

may encounter. In some scenarios, a subgroup's attrition results in outcomes that are missing completely at random (MCAR), outcomes that are missing not at random (MNAR) where high achieving students are more likely to leave, or outcomes that are MNAR where low-achieving students are more likely to leave.

Table 19.

*Coding of the Characterizations of Subgroup's Missing Outcomes*

| | |
|---|---|
| R: | Missing Completely at Random (MCAR) |
| H: | Missing Not at Random (MNAR) where high achieving students are more likely to leave |
| L: | Missing not at Random (MNAR) where low achieving students are more likely to leave |

Table 20 presents the six scenarios. It is followed by a brief discussion of the rationale for each specification on the nature of attrition in each subgroup.

Table 20.

*Six Scenarios on Nature of Attrition in Three Subgroups*

| | Lottery Winners | Lottery Losers | Non-Participants |
|---|---|---|---|
| Scenario R-H-R | MCAR | MNAR<br>(High Achievers Leave) | MCAR |
| Scenario H-H-R | MNAR<br>(High Achievers Leave) | MNAR<br>(High Achievers Leave) | MCAR |
| Scenario R-H-H | MCAR | MNAR<br>(High Achievers Leave) | MNAR<br>(High Achievers Leave) |
| Scenario H-H-H | MNAR<br>(High Achievers Leave) | MNAR<br>(High Achievers Leave) | MNAR<br>(High Achievers Leave) |
| Scenario R-H-L | MCAR | MNAR<br>(High Achievers Leave) | MNAR<br>(Low Achievers Leave) |
| Scenario H-H-L | MNAR<br>(High Achievers Leave) | MNAR<br>(High Achievers Leave) | MNAR<br>(Low Achievers Leaver) |

*Simulated attrition of lottery winners.* The attrition of lottery winners is specified in one of two forms: (1) MCAR; or (2) MNAR with high achievers more likely to leave. The attrition of lottery winners may be MCAR if students left for idiosyncratic reasons such as residential moves, parent job transfers/changes, disciplinary problems, etc. This specification assumes the potential outcomes of the lottery winners who left the sample are expected to be the same on average as those who remained.

The attrition of lottery winners may also be MNAR with high achievers more likely to leave for reasons that are unobserved and associated with future performance. This would be the case if the lottery winners who left the sample had greater motivation to seek out better schooling options, even after winning entry to the district's highest performing school. If this motivation has a positive association with student achievement, then the estimates of the magnet effect will be downwardly biased because the best students will have left the sample.

*Simulated attrition of lottery losers.* All six characterizations assume the attrition of lottery losers results in outcomes that are MNAR, where the students that are most likely to leave are those who would have had higher future performance than what the observed covariates would have predicted. I maintain this sole assumption for the lottery losers because of substantial anecdotal evidence that lottery losers were leaving the district in pursuit of better education in private schools. We expect those who left were more likely to be high achievers because their parents had more resources and/or more motivation to improve their schooling than those lottery losers who stayed. Given that parental motivation is unobserved in the data, we assume that this attrition is missing not at random (MNAR) rather than missing at random (MAR).

*Simulated attrition of non-participants.* I specify the missing outcomes of non-participants in three forms. First, I specify the missing outcomes as MCAR, which would be the case if families left the district for idiosyncratic reasons such as residential moves, parent job transfers, etc.

Second, I specify the missing outcomes as MNAR with high-achievers more likely to leave. This would be the case if families were satisfied with the district's elementary schools, but uncomfortable sending their child to one of the district's middle schools. Parents with the resources and motivation may decide to exercise another schooling option such as an inter-district transfer or a private school. If there is a positive association between parental resources and motivation and student achievement, this would downwardly bias non-experimental estimates that use non-participants in their comparison group.

Third, I specify the missing outcomes as MNAR with low-achievers more likely to leave. This would be the case if mobility was a signal of an unstable home life and an unstable home life associated with poor future school performance.

*Specifying the Attrition Rates among Three Subgroups*

For each of the six characterizations on the form of attrition, I specify five different attrition rates (zero, low, moderate, high, severe). Table 21 shows the induced attrition rates for each subgroup under these five specifications. The first specification is that the attrition rate is zero for all subgroups, meaning the estimators are run on the complete sample. The experimental IV estimator generated from this sample represents the unbiased estimate of the causal effect of magnet school attendance because it suffers from zero attrition bias as well as zero selection bias (as a result of using randomly assigned comparison groups). The non-experimental estimators are also free of attrition bias, however they may still suffer from selection bias because they use non-experimental comparison groups. Specifications 2-5 induce increasing amounts of attrition into the sample. In order to maintain a pattern of attrition in the simulated data that was similar to that observed in the actual data, I impose attrition in the lottery loser sample that is 50 % greater than the attrition rates of lottery winners and non-participants. [21]

Table 21.

*Simulated Attrition Rates among Three Subgroups*

| | Lottery Winners | Lottery Non-Participants | Lottery Losers |
|---|---|---|---|
| Zero Attrition | 0% | 0% | 0% |
| Low Attrition | 10% | 10% | 15% |
| Moderate Attrition | 20% | 20% | 30% |
| High Attrition | 30% | 30% | 45% |
| Severe Attrition | 40% | 40% | 65% |

### *Inducing Attrition in the Samples*

With six different characterizations on the form of attrition in the sample and five different assumptions on the attrition rates, I have 30 unique characterizations of sample attrition. Attrition is induced into the complete sample to create 30 unique samples, one for each attrition characterization.

To artificially induce attrition that matches the MCAR specification, I randomly flag student outcomes and drop them from the sample in accordance with the specified attrition rate of each subsample. For example, under characterization 2 (low-attrition) I randomly drop 10% of the outcomes of lottery winners and 10% of the outcomes of non-participants.

[21] Lottery losers had a 46% higher attrition rate than lottery winners. This difference was statistically significant. Whereas the attrition rates of lottery winners and lottery non-participants were not statistically different. For simplicity, I round the different attrition rate of lottery losers to lottery winners and non-participants up to 50%.

To artificially induce attrition that is MNAR, I drop outcomes based on the values of their residuals from a least squares regression of 5th grade math scores on the observed covariates, using the complete sample. Using the residuals allows us to ensure the attrition we simulate is MNAR because we base attrition on the unobserved factors that explain Y. This regression takes the following form:

$$Y_{it} = \beta_0 + \beta_x X_{it} + \beta_c C_i + \gamma_t + \varepsilon_{it} \qquad (7.1)$$

Where $Y$ is the math score for student $i$ in year $t$. $X$ is a vector of explanatory variables indicating student participation in special programs (FRL, ESL, special education), 4th grade achievement in reading and math, race (Black =1), and gender (female =1). We impose constant effects of $X$ over time. $C$ is an indicator of student $i$'s cohort. $\gamma_t$ are year fixed-effects. $\varepsilon_{it}$ is the error term.

I drop the outcomes of the three subgroups (lottery winners, lottery losers, non-participants) based on whether or not their predicted residuals $\hat{\varepsilon}_i$ falls above or below the median of the residuals for the complete sample. Students with positive values of $\hat{\varepsilon}_i$ are those who performed better than predicted based on the regressors in model 7.1. Students with negative values of $\hat{\varepsilon}_i$ are those who performed worse than predicted. In simulations where the attrition is characterized as MNAR where those who left were more likely to be high achievers, I disproportionately drop outcomes for those whose value of $\hat{\varepsilon}_i$ falls above the median. For example, in characterization 2 (low attrition), I drop a total of 15% of the outcomes of lottery losers. To characterize these outcomes as MNAR with high achievers more likely to leave, two-thirds of these outcomes (10% of the full sample of lottery losers) are for students with values of $\hat{\varepsilon}_i$ above the median and one-third of the outcomes are for those with values of $\hat{\varepsilon}_i$ below the median.

Table 22 presents the distribution of the simulated missing outcomes above and below the median of the distribution of $\hat{\varepsilon}_i$ for each of the 30 characterizations of sample attrition. Note that in situations where the attrition is MCAR, half of the missing outcomes will have values of $\hat{\varepsilon}_i$ that are above the median and half will be below the median.

The random selection of outcomes within each subgroup is done 1,000 times for each of the 30 attrition simulation samples. Each estimator is then run on 1,000 samples and the average of the 1,000 estimates is used as the final estimate of the magnet effect.

*Estimate of the Causal Effect of Magnet School Enrollment under the Complete Sample*

In this analysis, the experimental IV estimate run on the complete sample is the unbiased estimate of $\delta$. I refer to this estimate as $\hat{\delta}_z^c$, where *c* stands for "complete sample" and *z* indicates the experimental IV estimator.

$\hat{\delta}_z^c$ equals 3.57 with standard error of 1.28 (p=0.005). This serves as our unbiased estimate of the causal effect of magnet enrollment on 5th grade math achievement. We can use this baseline to estimate to estimate the bias that results under the 30 characterizations of sample attrition.

Table 22.

*Distribution of missing outcomes above and below the 50th percentile of residuals for the 30 simulated attrition samples*

| | | **Scenario R-H-R** | | **Scenario H-H-R** | | **Scenario R-H-H** | | **Scenario H-H-H** | | **Scenario R-H-L** | | **Scenario H-H-L** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Attrition Rate** | **Subgroup** | **Above Median** | **Below Median** | **Above Median** | **Below Median** | **Above Median** | **Below Median** | **Above Median** | **Below Median** | **Above Median** | **Below Median** | **Above Median** | **Below Median** | **Total Attrition** |
| **Zero** | Lottery Winners | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| | Lottery Losers | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| | Non-Participants | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| **Low** | Lottery Winners | 5% | 5% | 6.7% | 3.3% | 5% | 5% | 6.7% | 3.3% | 5% | 5% | 6.7% | 3.3% | 10% |
| | Lottery Losers | 10% | 5% | 10% | 5% | 10% | 5% | 10% | 5% | 10% | 5% | 10% | 5% | 15% |
| | Non-Participants | 5% | 5% | 5% | 5% | 6.7% | 3.3% | 6.7% | 3.3% | 3.3% | 6.7% | 3.3% | 6.7% | 10% |
| **Moderate** | Lottery Winners | 10% | 10% | 13.3% | 6.6% | 10% | 10% | 13.3% | 6.7% | 10% | 10% | 15% | 5% | 20% |
| | Lottery Losers | 20% | 10% | 20% | 10% | 20% | 10% | 20% | 10% | 20% | 10% | 20% | 10% | 30% |
| | Non-Participants | 10% | 10% | 10% | 10% | 15% | 5% | 13.3% | 6.7% | 6.7% | 13.3% | 6.7% | 13.3% | 20% |
| **High** | Lottery Winners | 15% | 15% | 20% | 10% | 15% | 15% | 20% | 10% | 15% | 15% | 20% | 10% | 30% |
| | Lottery Losers | 30% | 15% | 30% | 15% | 30% | 15% | 30% | 15% | 30% | 15% | 30% | 15% | 45% |
| | Non-Participants | 15% | 15% | 15% | 15% | 20% | 10% | 20% | 10% | 10% | 20% | 10% | 20% | 30% |
| **Severe** | Lottery Winners | 20% | 20% | 26.6% | 13.3% | 20% | 20% | 26.6% | 13.3% | 20% | 20% | 26.6% | 13.3% | 40% |
| | Lottery Losers | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 60% |
| | Non-Participants | 20% | 20% | 20% | 20% | 26.6% | 13.3% | 26.6% | 13.3% | 13.3% | 26.6% | 13.3% | 26.6% | 40% |

*Estimating the Bias in the Estimators under the 30 Simulated Attrition Samples*

Using $\hat{\delta}_z^c$, I estimate the bias from each non-experimental estimator $k$ for each of the 30 simulated samples *s*, as: $\hat{b}(\delta_k^s) = \hat{\delta}_z^c - \hat{\delta}_k^s$. The non-experimental estimators investigated herein are those presented in the previous chapter, with the same specifications.[22] Likewise, I estimate the bias in the experimental estimator *z* when run on *s* as: $\hat{b}(\delta_z^s) = \hat{\delta}_z^c - \hat{\delta}_z^s$. To comparatively evaluate the performance of the non-experimental estimators against the experimental estimator for each sample *s,* I find the difference in absolute values of the two bias estimates: $\hat{\sigma}_k^s = |\hat{b}(\delta_k^s)| - |\hat{b}(\delta_z^s)|$. If $\hat{\sigma}_k^s < 0$ it indicates that non-experimental estimator *k* yields estimates with less bias than the experimental estimator when attrition is of the form and rate *s*. Conversely, if $\hat{\sigma}_k^s > 0$ it indicates the experimental estimator remains the least biased estimate under *s*.

*Results*

The bias estimates from the simulations are presented below in six graphs (figures 1-6) corresponding to the six attrition scenarios.[23] In each graph the amount of bias is presented in a standardized effect size measure. This measure is the difference between the unbiased experimental IV estimate from the complete sample ($\hat{\delta}_z^c$) and the experimental IV or non-experimental estimate from the sample suffering from attrition ($\hat{\delta}_z^s$ or $\hat{\delta}_k^s$) divided by the standard deviation of gains in math test scores for the complete sample of 5th grade students:

[22] The PSM estimators evaluated in this chapter allow “inter-school” matching – meaning the potential matches for a given student are not restricted to a student’s 4th grade school.

[23] The actual estimates of the magnet effect for each estimator under the 30 attrition characterizations are presented in the appendix.

$ES\left(\hat{b}(\delta_k^s)\right) = \dfrac{\hat{\delta}_z^c - \hat{\delta}_k^s}{s_{\Delta Yz}}$. This allows us to evaluate the magnitude of the bias in relation to annual achievement gains.

The experimental IV estimator is clearly most sensitive to sample attrition; whereas the other methods produce relatively stable estimates (with similar amounts of bias) under all attrition rates, the bias in the experimental IV estimates trends upward as the attrition rates increase. Nevertheless, the experimental IV estimates are less biased than the non-experimental estimates for all samples with low or moderate attrition rates and most scenarios with high attrition rates.

The MR estimates are the most biased in every one of the 30 characterizations of attrition. There never reaches a point where the MR estimates are less biased than the experimental IV estimates. Even when the lottery loser sample suffers 60% non-random attrition, where high-achieving lottery losers are more likely to attrit, MR yields estimates with bias that is 0.20 greater than the the experimental IV estimate. This is evidence that multiple regression with observed covariates using non-participants is not a defensible substitute for an analysis that uses the experimental comparison groups, regardless of the severity of attrition.

The student fixed-effects estimator was found to be most robust to the effects of selective attrition and to perform relatively well in samples with moderate, high, and severe attrition rates. In fact, when attrition rates reached high levels (45% among lottery losers, 30% among lottery winners and non-participants) the fixed-effects estimator performed better than the experimental IV estimator in the two scenarios where the attrition form of non-participants was specified as non-random with low-achievers more likely to leave. When attrition rates reach the severe level (60% among lottery losers, 40% among lottery winners and non-participants), the fixed-effect estimates were least biased of all estimates. This finding lends affirmation to the approach used by Celia Rouse in her investigation of the Milwaukee voucher program, where Rouse used the fixed-effects estimator to address the fact that 52% of lottery losers attrited.

The propensity score matching estimators are robust to the effects of attrition, as evidenced by similarity of PSM estimates from samples with different attrition rates. This is to be expected given PSM methods adapt to sample attrition by either re-weighting the comparison group (the case with LRM) or seeking the best one-to-one matches among those who are left (the case with NN).  Nevertheless, the PSM estimates have substantial bias due to the fact they cannot control for selection on unobservables and, like the MR estimator, must rely on observed covariates to control for self-selection into the magnet school.  This analysis suggests the PSM estimators should be ruled out as a method for evaluating magnet school programs.  Only when attrition reaches the severe level do the PSM estimators perform as well as the experimental IV. However, when attrition is severe both the PSM and experimental IV estimators produce estimates with bias greater than a standardized effect of 0.20, suggesting neither should be used to estimate a treatment effect.

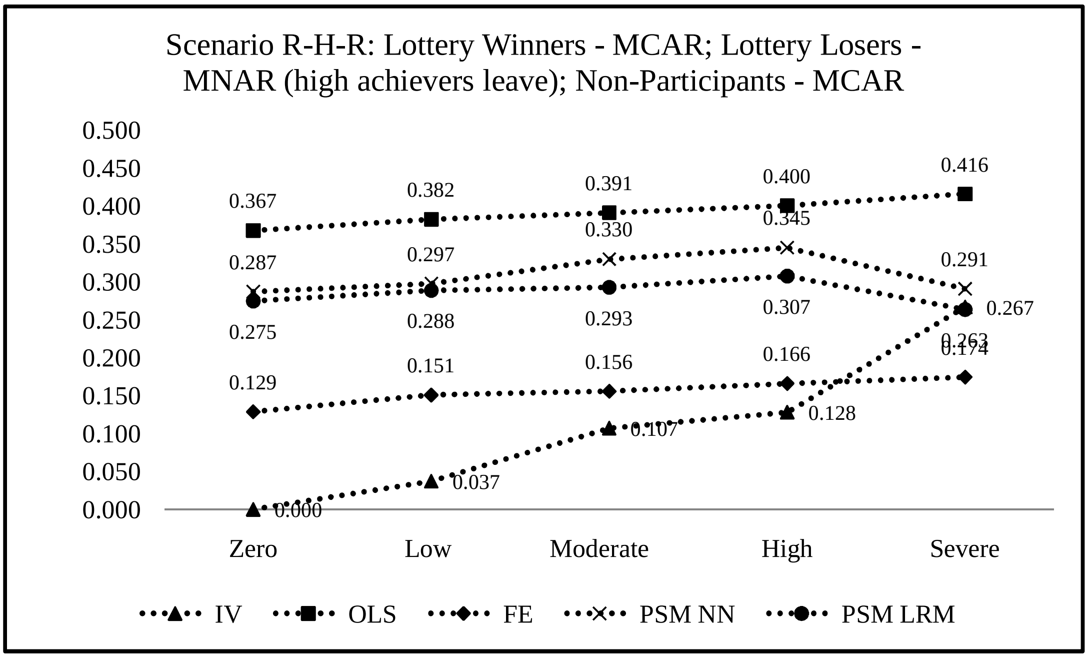

Figure 1.Bias in Experimental and Non-Experimental Estimators under Scenario R-H-R

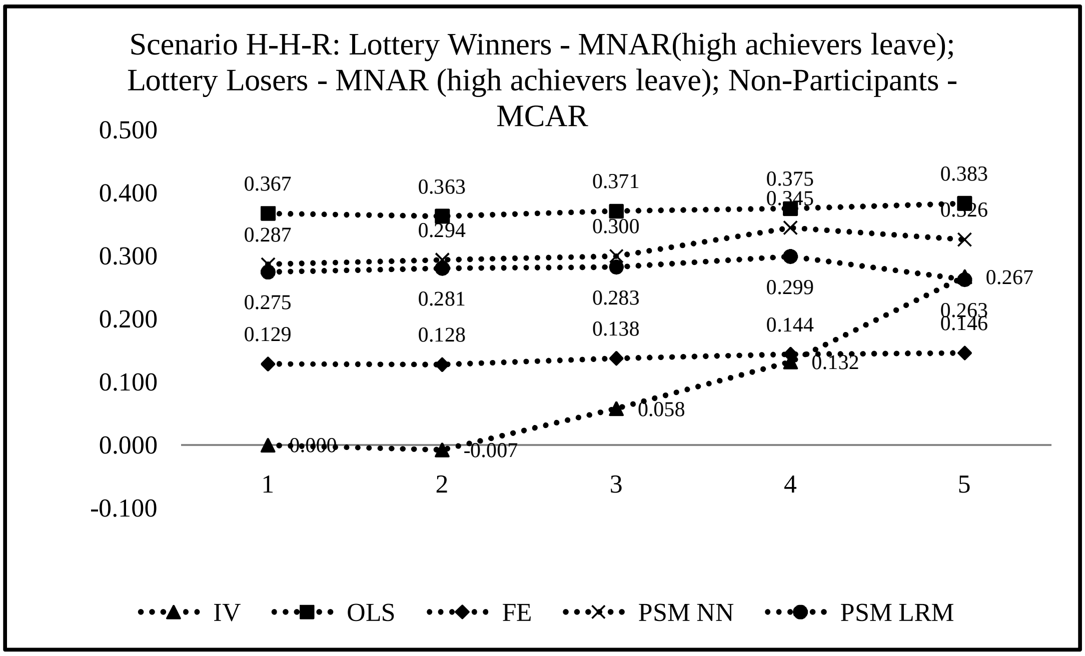

Figure 2. Bias in Experimental and Non-Experimental Estimators under Scenario H-H-R

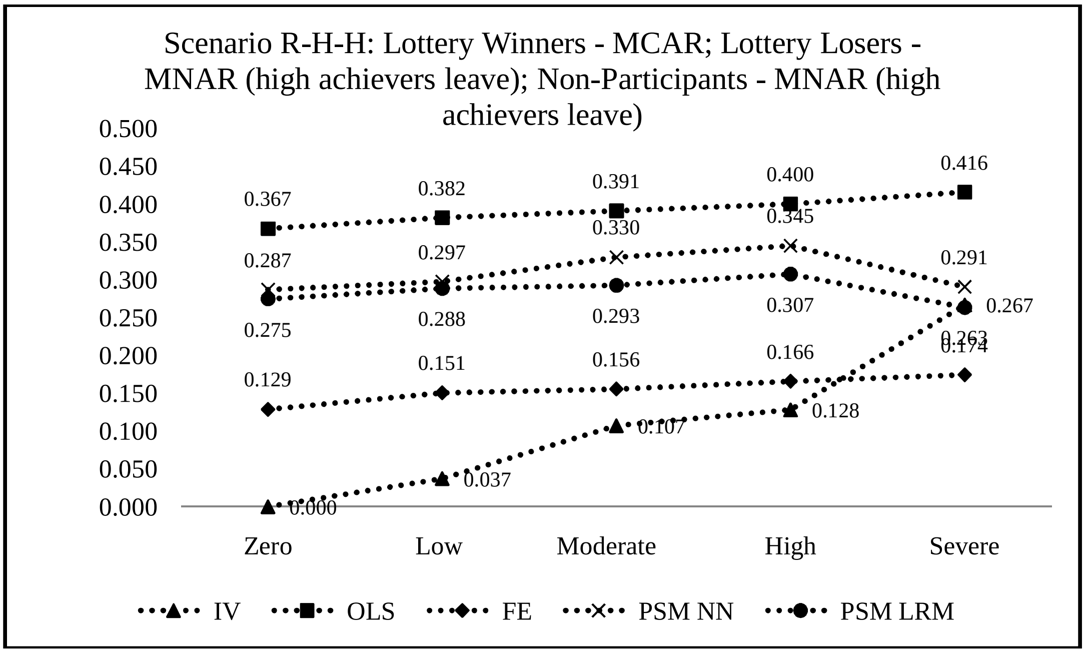

Figure 3. Bias in Experimental and Non-Experimental Estimators under Scenario R-H-H

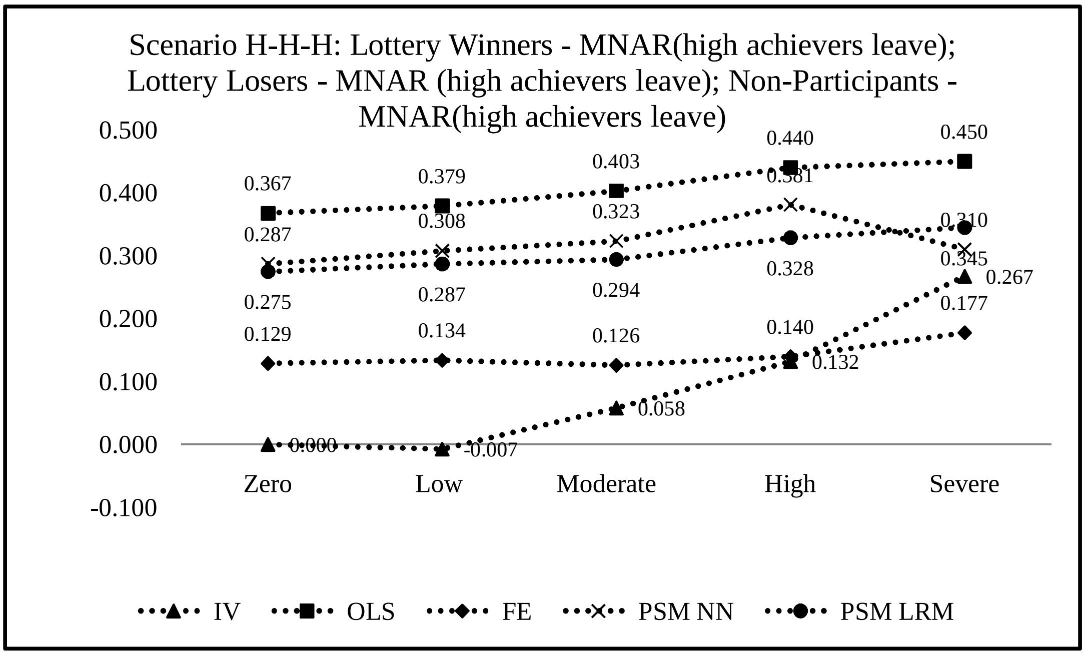

Figure 4. Bias in Experimental and Non-Experimental Estimators under Scenario H-H-H

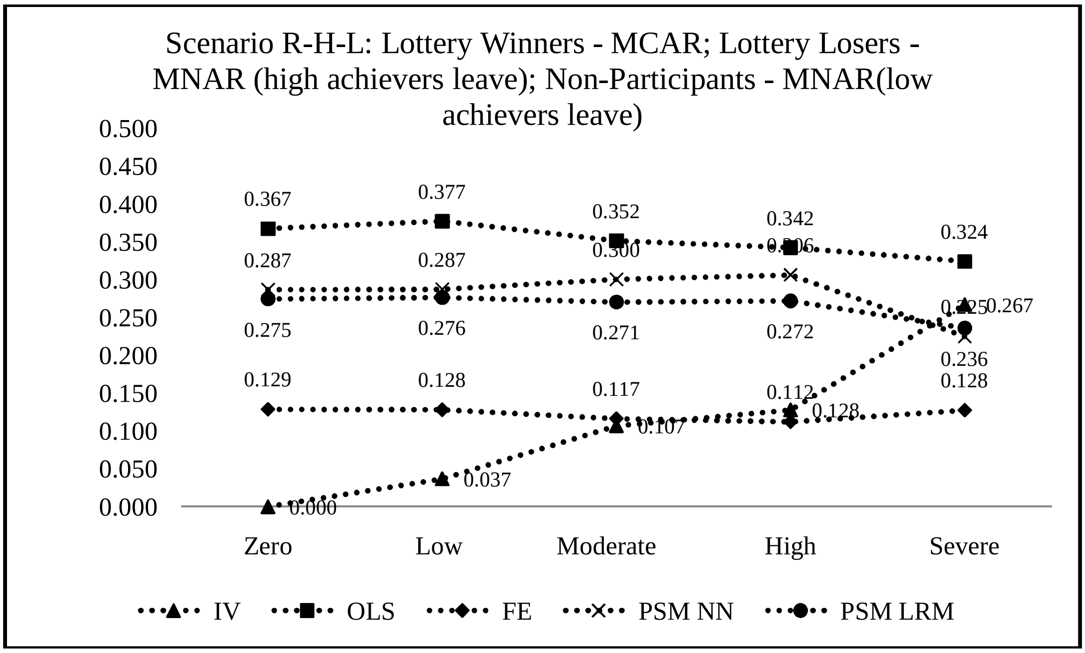

Figure 5. Bias in Experimental and Non-Experimental Estimators under Scenario R-H-L

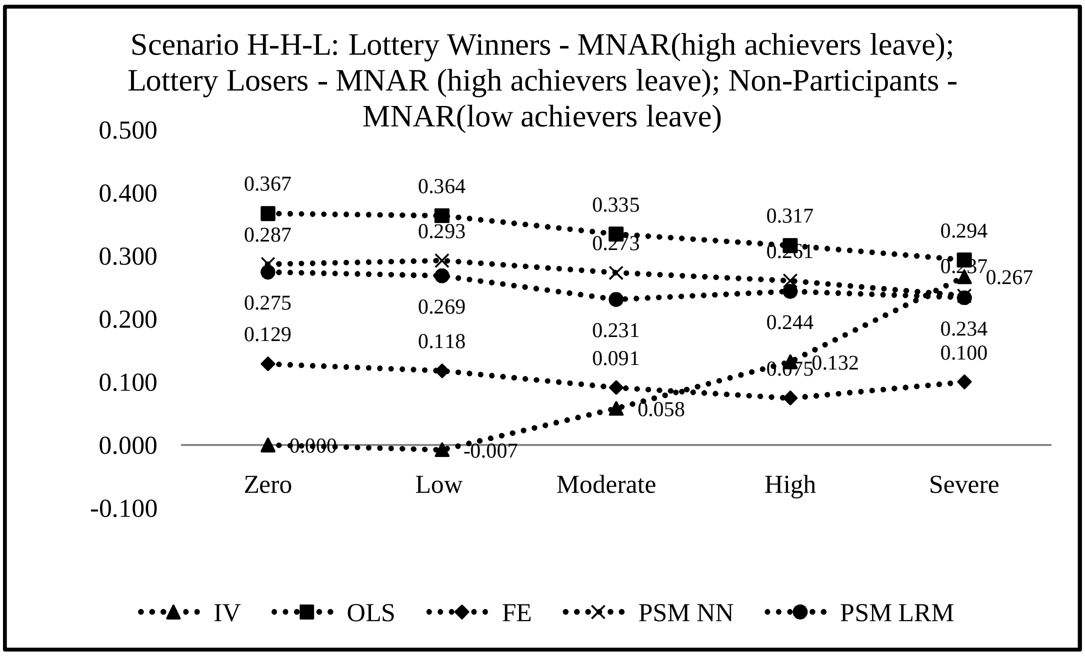

Figure 6. Bias in Experimental and Non-Experimental Estimators under Scenario H-H-L

## APPENDIX

Table A-1.

*First Stage Results of Experimental IV 2SLS Regressions*

| | Magnet Enrollment |
|---|---|
| Magnet Lottery Winner 5th Grade | 0.805*** |
| | (0.010) |
| Magnet Lottery Winner 6th Grade | 0.001 |
| | (0.009) |
| 4th Grade Reading | 0.001*** |
| | (0.000) |
| 4th Grade Math | 0.000 |
| | (0.000) |
| Black | 0.024* |
| | (0.012) |
| Free and Reduced-Price Lunch | -0.084*** |
| | (0.014) |
| ESL | 0.067*** |
| | (0.017) |
| Female | 0.022 |
| | (0.010) |
| Constant | -0.459 |
| | (0.918) |
| Lottery Year | Yes |
| Cohort | Yes |
| Grade | Yes |
| Observations | 3616 |
| R-Squared | 0.759 |

*p<0.05, ** p<0.01, *** p<0.001

*Note.* The F-statistic from the first-stage is 3135.88; this allows me to reject the null hypothesis that the independent variables in the first stage equation weakly identify the instrument. Stock and Yogo (2005) indicate that the critical first-stage F-statistic value for indicating a weak instrument is 16.38.

Table A-2.

*Intent to Treat Experimental Estimates of Magnet Effect*

| | Math | Reading |
|---|---|---|
| Magnet Lottery Winner 5th Grade | 4.016*** | 3.027*** |
| | (1.227) | (1.097) |
| Magnet Lottery Winner 6th Grade | -.528 | -.790 |
| | (1.259) | (1.127) |
| 4th Grade Reading | 0.179*** | 0.499*** |
| | (0.017) | (0.015) |
| 4th Grade Math | 0.430*** | 0.096*** |
| | (0.018) | (0.016) |
| Black | -7.702*** | -7.843*** |
| | (1.129) | (1.008) |
| Free and Reduced-Price Lunch | -7.407*** | -6.695*** |
| | (1.292) | (1.154) |
| ESL | 5.668*** | 1.64 |
| | (1.585) | (1.415) |
| Female | -3.326*** | -0.369 |
| | (0.879) | (0.785) |
| 5th Grade | 3.338 | 28.449*** |
| | (6.646) | (5.933) |
| Constant | 281*** | 270 |
| | (14.9) | (13.4) |
| Cohort Effects | Yes | Yes |
| Grade Effects | Yes | Yes |
| Year Effects | Yes | Yes |
| Test Effects | Yes | Yes |
| Observations | 3613 | 3613 |
| R-Squared | 0.738 | 0.797 |

*p<0.05, ** p<0.01, *** p<0.001

Table A-3.

*Tests of Validity of Instruments*

| Test | Test Statistic | Description of Test(s) |
|---|---|---|
| Weak Identification Tests | | |
| Cragg-Donald Wald F Statistic | 3135*** | Test of correlation of endogenous regressors with excluded instruments; Ho: equation is weakly identified |
| Underidentification Tests | | |
| Anderson canon. corr. N*CCEV LM statistic | 2297*** | Test of relevance of excluded instruments; Ho: the model is underidentified |
| Cragg-Donald N*CDEV Wald statistic | 6305*** | |
| Weak-Instrument/Robust Inference Tests | | |
| Anderson-Rubin Wald test | 5.81** | Tests of joint significance of endogenous regressors in main equation; Ho: B1=0 and overidentifying restrictions are valid |
| Anderson-Rubin Wald test | 11.7** | |
| Stock-Wright LM S statistic | 11.6** | |

Table A-4.

*Reduced Form Math IV Estimates by student characteristic*

| | Black | Non-Black | FRL | Non-FRL | Bottom Quartile 4th Grade Math | Inter Quartiles 4th Grade Math | Top Quartile 4th Grade Math |
|---|---|---|---|---|---|---|---|
| Selective Magnet5th Grade | 10.191* | 3.510** | 10.508 * | 4.180*** | 5.648** | 5.461*** | 2.954 |
| | (2.662) | (1.780) | (4.164) | (1.635 | (2.6925) | (2.067) | (3.322) |
| Selective Magnet 6th Grade | 4.148 | -.387 | 8.433 | -.321 | .5708 | .101 | -1.373 |
| | (2.721) | (1.918) | 5.301 | (1.689 | 2.932 | 2.157 | 3.534 |
| 4th Grade Reading Test | 0.124* | .183** | 0.112 | .17646*** | 0.139** | 0.209*** | 0.135** |
| | (0.031) | (.020) | (0.072) | (.0186 | (0.052) | (0.037) | (0.049) |
| 4th Grade Math Test | 0.425*** | .419*** | 0.440*** | .4220*** | 0.560*** | 0.457*** | 0.309*** |
| | (0.031) | (.020) | (0.073) | (.0190 | (0.117) | (0.105) | (0.079) |
| Black | -- | -- | -4.002 | -8.100*** | -5.031* | -6.830*** | -11.314** |
| | -- | -- | (3.753) | (1.296 | 2.705 | (2.539) | (4.150) |
| Free and Reduced-Price Lunch | -4.736** | -9.149*** | -- | | -5.492 | -7.674*** | -6.080 |
| | (1.606) | (1.814) | -- | | (3.122) | (2.934) | (4.699) |
| ESL | -2.166 | 7.405*** | -1.766 | 7.550*** | -0.284 | 6.866 | 8.220* |
| | (3.347) | (1.805) | (5.369) | (1.808 | (5.263) | (3.723) | (3.853) |
| Female | -1.253 | -4.052*** | -3.945 | -3.287*** | -0.357 | -5.915 | -1.339 |
| | (2.737) | (1.043) | (3.555) | (.9614 | (2.506) | (1.920) | (2.650) |
| Disabled | 4.566 | 6.141*** | -1.463 | 6.497*** | .854*** | 5.531 | 11.932*** |
| | (4.082) | (1.514) | 5.222 | (1.4514 | 2.878 | (2.019) | 2.633 |
| Constant | 327.3*** | 274.5*** | 331.2*** | 275.376*** | 192.0*** | 260.8 | 405.646*** |
| | (43.1) | (16.1) | (54.2) | (15.008 | (44.9) | (70.4) | (58.023) |
| Lottery Year (Cohort Effects) | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Year Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade*Year (Test Effects) | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| R-Squared | 0.828 | 0.700 | 0.788 | 0.729 | 0.794 | 0.727 | 0.610 |
| Observations | 830 | 2775 | 544 | 3,061 | 957 | 1766 | 882 |

*p<0.05, ** p<0.01, *** p<0.001

*Note.* The specification of the IV model for reduced form estimates shown above is identical to the specification used to generate the estimates for the full sample model. The sample is limited to 5th and 6th grade students who participated in the selective magnet lottery and were enrolled in the district in 4th grade. All models are estimated with Huber White robust standard errors to account for the correlation of errors across years within a single student as well as the correlation of the errors of lottery losers who attend the same non-selective magnet school

Table A-5.

*Reduced Form Reading IV Estimates by student characteristic*

| | Black | Non-Black | FRL | Non-FRL | Bottom Quartile 4th Grade Math | Inter Quartiles 4th Grade Math | Top Quartile 4th Grade Math |
|---|---|---|---|---|---|---|---|
| Selective Magnet5th Grade | 8.388*** | 2.659* | 5.936 | 3.455** | 5.535** | 2.148 | 4.090 |
| | (2.560 | (1.5903 | (3.868 | (1.455 | (2.586 | (1.709) | (3.242) |
| Selective Magnet 6th Grade | 4.191 | .1022 | 4.083 | .7024 | .8703 | 1.037 | -1.016 |
| | (2.631 | (1.694 | (4.954 | (1.504 | (2.824 | (1.797) | (3.327) |
| 4th Grade Reading Test | .478*** | .496*** | .441*** | .499*** | .654*** | .657*** | .262*** |
| | (.0302 | (.018 | (.0403 | (.016) | (.062) | (.054) | (.053) |
| 4th Grade Math Test | .0684 | .099*** | .102*** | .093*** | .109*** | .075*** | .098 |
| | (.0321 | (.018 | (.041) | (.016) | (.030) | (.021) | (.033) |
| Black | | | -9.602** | -7.341*** | -2.246 | -7.442*** | -13.189*** |
| | | | (2.124) | (1.155) | (1.497) | (1.372) | (2.940) |
| Free and Reduced-Price Lunch | -6.376 | -6.294*** | | | -5.333*** | -4.283*** | -10.862*** |
| | (1.543 | (1.602) | | | (1.709) | (1.577) | (3.321) |
| ESL | 1.842 | 1.645 | -1.341 | 2.213 | 3.626* | 1.871 | 1.344 |
| | (3.221 | (1.595) | (3.012) | (1.610) | (2.239) | (1.918) | (3.599) |
| Female | -.44060 | -.196 | .751 | -.534 | -1.683 | -1.324 | 2.177 |
| | (1.461 | (.9217) | (2.004 | (.856) | (1.378) | (1.024) | (1.887) |
| Disabled | 3.755 | 3.538** | -1.647 | 3.894*** | -.896 | 2.248 | 7.888 |
| | (3.929 | (1.337) | (4.859 | (1.291) | (2.863) | (1.708) | (2.367) |
| Constant | 288.0314 | 274.782 | 294.620 | 276.360*** | 152.358*** | 182.136*** | 439.961*** |
| | (23.15444 | (14.249) | (30.474 | (13.363) | (39.344) | (37.484) | (41.323) |
| Lottery Year (Cohort Effects) | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Year Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Grade*Year (Test Effects) | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| R-Squared | 0.843 | 0.778 | 0.812 | 0.792 | 0.8111 | 0.803 | 0.746 |
| Observations | 830 | 2775 | 544 | 3,061 | 936 | 1831 | 838 |

*p<0.05, ** p<0.01, *** p<0.001

*Note*. The specification of the IV model for reduced form estimates shown above is identical to the specification used to generate the estimates for the full sample. The sample is limited to 5th and 6th grade students who participated in the selective magnet lottery and were enrolled in the district in 4th grade. All models are estimated with Huber White robust standard errors to account for the correlation of errors across years within a single student as well as the correlation of the errors of lottery losers who attend the same non-selective magnet school

Table A-6.

*Heckman & Hotz (2002) Specification Test: Running Non-Experimental Models using Pre-Intervention Data (2nd, 3rd, and 4th grade scores)*

| | Math | | Reading | |
|---|---|---|---|---|
| | 3rd Grade | 4th Grade | 3rd Grade | 4th Grade |
| Multiple Regression | 19.876*** | 8.598*** | 16.898*** | 9.689*** |
| | (1.437) | (1.441) | (1.451) | (1.456) |
| Student Fixed Effects | .0307 | -9.468*** | 7.895*** | 3.723* |
| | (1.734) | (1.734) | (1.545) | (1.545) |
| PSM NN | 15.353 | -4.053 | 15.369*** | -.0991 |
| | (2.955) | (2.617) | (2.615) | (2.184) |
| PSM LRM | 15.226 | -4.301 | 15.147*** | -1.628 |
| | (2.953) | (2.576) | (2.599) | (2.161) |

*p<0.05, ** p<0.01, *** p<0.001

Table A-7.

*Estimates of the Magnet effect Under 30 Characterizations of Attrition*

| | Zero | Low | Moderate | High | Severe |
|---|---|---|---|---|---|
| **Scenario R-H-R** | | | | | |
| IV | 3.57*** | 4.58*** | 6.478*** | 7.05*** | 10.81*** |
| MR | 13.55*** | 13.95*** | 14.19*** | 14.44*** | 14.86*** |
| FE | 7.07*** | 7.66*** | 7.80*** | 8.08*** | 8.30*** |
| PSM NN | 11.36*** | 11.65*** | 12.53*** | 12.94*** | 11.46*** |
| PSM LRM | 11.03*** | 11.41*** | 11.52*** | 11.92*** | 10.73*** |
| | | | | | |
| **Scenario H-H-R** | | | | | |
| IV | 3.57*** | 3.37*** | 5.15*** | 7.16*** | 10.83*** |
| MR | 13.55*** | 13.43*** | 13.65*** | 13.76*** | 13.98*** |
| FE | 7.07*** | 7.04*** | 7.31*** | 7.49*** | 7.54*** |
| PSM NN | 11.36*** | 11.56*** | 11.72*** | 12.93*** | 12.43*** |
| PSM LRM | 11.03*** | 11.19*** | 11.25*** | 11.70*** | 10.70*** |
| | | | | | |
| **Scenario R-H-H** | | | | | |
| IV | 3.57*** | 4.30*** | 4.47*** | 6.30*** | 12.95*** |
| MR | 13.55*** | 14.22*** | 14.81*** | 15.99*** | 16.66*** |
| FE | 7.07*** | 7.48*** | 7.67*** | 8.39*** | 9.13*** |
| PSM NN | 11.36*** | 12.07*** | 13.10*** | 14.66*** | 13.93*** |
| PSM LRM | 11.03*** | 11.59*** | 12.12*** | 13.14*** | 13.06*** |
| | | | | | |
| **Scenario H-H-H** | | | | | |
| IV | 3.57*** | 3.37*** | 4.70*** | 6.83*** | 10.83*** |
| MR | 13.55*** | 13.86*** | 14.52*** | 15.52*** | 15.79*** |
| FE | 7.07*** | 7.20*** | 6.99*** | 7.37*** | 8.39*** |
| PSM NN | 11.36*** | 11.93*** | 12.35*** | 13.92*** | 11.56*** |
| PSM LRM | 11.03*** | 11.37*** | 11.56*** | 12.49*** | 12.93*** |
| | | | | | |
| **Scenario R-H-L** | | | | | |
| IV | 3.57*** | 5.54*** | 6.48*** | 10.06*** | 12.95*** |
| MR | 13.55*** | 13.82*** | 13.12*** | 12.87*** | 12.37*** |
| FE | 7.07*** | 7.05*** | 6.74*** | 6.62*** | 7.03*** |
| PSM NN | 11.36*** | 11.37*** | 11.73*** | 11.89*** | 9.68*** |
| PSM LRM | 11.03*** | 11.08*** | 10.92*** | 10.96*** | 9.99*** |
| | | | | | |
| **Scenario H-H-L** | | | | | |
| IV | 3.57*** | 4.66*** | 4.70*** | 6.83*** | 10.83*** |
| MR | 13.55*** | 13.46*** | 12.66*** | 12.17*** | 11.55*** |
| FE | 7.07*** | 6.76*** | 6.05*** | 5.59*** | 6.29*** |
| PSM NN | 11.36*** | 11.52*** | 10.99*** | 10.65*** | 10.01*** |
| PSM LRM | 11.03*** | 10.87*** | 9.84*** | 10.19*** | 9.92*** |

# REFERENCES

Abdulkadiroglu, A., Che, Y., & Yasuda, Y. (2009). R*esolving Conflicting Preferences in School Choice: The 'Boston' Mechanism Reconsidered*. Accessed July 2009 from the Social Sciences Research Network: http://ssrn.com/abstract=1456088.

Agodini, R. & Dynarski, M. (2004). Are Experiments the Only Option? A Look at Dropout Prevention Programs. *Review of Economics and Statistics,* 86(1):180-194.

Angrist, J., Imbens, G. & Rubins, D. (1996) Journal of the American Statistical Association, Vol. 91, 1996

Ballou, D., Goldring, E., & Liu, K. (2006). *Magnet Schools and Student Achievement*. Unpublished.

Ballou, D. (2007). Magnet Schools and Peers: Effects on Mathematics Achievement. Nashville: Vanderbilt University. Unpublished.

Betts, J., Rice, L., Zau, A., Tang, E., Koedel, C. Andrew Zau, Emily Tang, & Koedel, C. (2006). *Does School Choice Work? Effects on Student Integration and Academic Achievement.* Public Policy Institute of California.

Black, D. & Smith, J. (2004). How Robust is the Evidence on the Effects of College Quality? *Journal of Econometrics*, 121(1-2): 99-124.

Blank, R. (1989). *Educational Effects of Magnet High Schools*. National Center on Effective Secondary Schools, Madison, WI. Office of Educational Research and Improvement. ED Washington DC.

Bloom, H., Michalopoulos, C., Hill, C., & Lei, Y. (2002). *Can Non-Experimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?* New York, NY: Manpower Demonstration Research Corporation, June 2002.

Bound, J., Jaeger, D., & Baker, R. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak . *Journal of the American Statistical Association*, 90(430): 443-450.

Buckley, J. & Schneider, M. (2008). *Charter Schools: Hope or Hype?* Princeton, NJ: Princeton University Press.

Crain, R, Heebner, A. & Yiu-Pong, S.(1992). *The Effectiveness of New York*

*City's Career Magnet Schools: An Evaluation of Ninth Grade Performance Using an Experimental Design.* Berkeley, CA: National Center for Research in Vocational Education.

Crain, R., Allen, A., Thaler, R., et al. (1999). *The Effects of Academic Career Magnet Education on High Schools and Their Graduates.* Berkeley, CA: National Center for Research in Vocational Education.

Cullen, J., Jacob, B., Levitt, S. (2006). The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools. *Journal of Public Economics*, 89(5):729-760.

Dehejia, R. & Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics,* 84(1):151-161.

Dehejia, R. & Wahba, S. (1999). Causal Effects in NX Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448), 1053-1062.

Diaz, J. & Handa, S. (2006). An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program. *Journal of Human Resources,* 16: 319-345.

Friedlander, D. & Robins, P. (1995). The Distributional Impacts of Social Programs. *Evaluation Review*, 21(5): 531-553

Goldring E. and Smrekar, C. (2000). Magnet Schools and the Pursuit of Racial Balance. *Education and Urban Society*, 33( 1): 17-35.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The annals of the American Academy of Political and Social Science, 589:* 63–93.

Greene, P., Peterson, P., & Du, Jiangtao. (1997). *The effectiveness of school choice: The Milwaukee experiment.* Harvard University Education Policy and Governance Occasional Paper 97-1.

Hausman, C., & Goldring, E. (2000). Parent Involvement, Influence, and Satisfaction in Magnet Schools: Do Reasons for Choice Matter? *The Urban Review,* 32(2): 105-121.

Heckman, J. & Hotz, J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of Manpower Training. *Journal of The American Statistical Association,* 84(408):862-874.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *NBER Working Paper No. W6699.*

Heckman, J., LaLonde, R., Smith, J. (1999). The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter, O., Card, D., eds., *Handbook of Labor Economics, 3A Volume 3A*:1865-2097.

Heinsman, D. (1993). Effect sizes in meta-analysis: Does random assignment make a difference? (Doctoral Dissertation). Memphis State University.

Hill, C., Bloom, H., Black, A. & Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives,* 2(3): 172-177.

Hoffman, L. (2006). Numbers and Types of Public Elementary and Secondary Schools From the Common Core of Data: School Year 2005-06. National Center for Education Statistics.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association,* 81(396): 945-960.

Hoxby, C. (2000). Does Competition among Public Schools Benefit Students and Taxpayers? *American Economic Review,* 90 (5): 1209-38.

Imbens, G. & Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects *Econometrica*, 62(2), 467-475.

Kane, T. (2004). *The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations.* William T. Grant Foundation Working Paper.

Kemple, J. & Snipes, J. (2000). *Career Academies: Impacts on Students' Engagement and Performance in High School.* Manpower Demonstration Research Corporation.

Kemple, J. & Scott-Clayton, J. (2004). *Career Academies: Impacts on Labor Market Outcomes and Educational Attainment*. Manpower Demonstration Research Corporation.

Lalonde, R. (1986). Evaluating the Econometric Evaluations of Training with Experimental Data. *The American Economic Review*, 76(4), 604-620.

Larson, J., Witte, J., Staib, S. & Powell, M. (1993). A Microscope on Secondary Magnet Schools in Montgomery County, Maryland. In Waldrip, Donald R., Walter L. Marks & Nolan Estes (Eds.), *Magnet School Policy Studies and Evaluations*. Houston: International Research Institute on Educational Choice, 261-435.

Lee, D. (2008). Training,Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. National Bureau of Economic Research Working Paper 11721.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist*, *48:* 1181-1201.

Manski, C. (*1995*), Identification Problems in the Social Sciences, Cambridge, MA: Harvard.

Martinez, V., Godwin, K., & Kemerer, F. (1996). *Public school choice in san antonio: Who chooses and with what effects?* In B. Fuller, R. Elmore, & G. Orfield (Eds.), Who Chooses? Who Loses? Culture, institutions, and the unequal effects of school choice (pp. 50-69). New York: Teachers College Press.

Musumeci, M. & Szczypkowski, R.(1993). New York State Magnet School Evaluation Study. In Waldrip, Donald R., Walter L. Marks & Nolan Estes (Eds.), *Magnet School Policy Studies and Evaluations.* Houston: International Research Institute on Educational Choice, 97-259.

Nixon, R. (1970). “156 - Special Message to the Congress Proposing the Emergency School Aid Act of 1970” *May 21st, 1970* accessed October 25, 2007 from http://www.presidency.ucsb.edu/ws/?pid=2509.

Poppell, J.& Hague, S. (2001). Examining Indicators to Assess the Overall Effectiveness of Magnet Schools: A Study of Magnet Schools in Jacksonville, Florida. Paper presented at the Annual Meeting of the American Educational Association, Seattle, WA April 10-14, 2001.

Quandt, R. (1972). A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association*, 67: 306-310.

Rosenbaum, P, & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika,* 70(1): 41-55.

Rosenbaum, P. & Rubin, D.(1985). The Bias Due to Incomplete Matching. *Biometrics,* 41: 103-116.

Rossell, C. (2005). No Longer Famous but Still Intact. Education Next. Spring(2005): 44-49.

Roy, A. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, 3: 135-146.

Rubin, D. (1973). Matching to Remove Bias in Observational Studies. *Biometrics,* 29(March): 159-183

Rubin, D. (1977). Assignment to a Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*. 2(1): 1-26

Rubin, D. (1980). Discussion of Randomization Analysis of Experimental Data: The Fisher Randomization Test, by D. Basu. *Journal of the American Statistical Association.* 75(371): 591-593

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Shadish, W. & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting in Clinical Psychology*. 64:1290–1305

Smrekar, C. & Goldring, E. (1999). *School choice in urban America: Magnet schools and the pursuit of equity.* New York: Teachers College Press.

Smith, J. & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2): 305-375.

Staiger, D., & Stock, J. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557-586

Steele, L. & Eaton, M. (1996). Reducing, Eliminating, and Preventing Minority Isolation in American Schools: The Impact of the Magnet Schools Assistance Program. Department of Education, Washington, DC. Office of the Under Secretary.

Wilde, E. & Hollister, R. (2007). How Close is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment. *Journal of Policy Analysis and Management*, 26 (3), 455-477.

Witte, J., Sterr, T. & Thorn, C. (1996). *Fifth-year Report: Milwaukee Parental Choice Program.* Madison, WI: Department of Political Science and The Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison.