Essays on the Economics of Education

Jason P. Imbrogno

Doctoral Committee:

- Dennis Epple, Thomas Lord Professor of Economics, Carnegie Mellon University
- Holger Sieg, J. M. Cohen Term Chair in Economics, University of Pennsylvania
- Ron Zimmer, Associate Professor of Public Policy and Education, Vanderbilt University
- John Engberg, PhD, Senior Economist, RAND Corporation
- David Klahr, Walter van Dyke Bingham Professor of Cognitive Development and Education Sciences, Carnegie Mellon University

My dissertation is titled *Essays on the Economics of Education.* It is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Economics) from Carnegie Mellon University.

Overall, education research is conducted at varying grain sizes with experts from many fields, ranging from neuroscientists focused on brain scans of a few students to cognitive psychologists interested in single classrooms to public policy experts evaluating federal standards across the entire country. My dissertation shows some of this variety, and it is organized as follows. Essay 1 is an evaluation of magnet schools in a large urban district in the US. In the essay, we develop a novel technique for dealing with selective attrition, and then apply it to explore the effects of the magnet programs on student behavior and test scores. It is joint work with John Engberg, Dennis Epple, Holger Sieg, and Ron Zimmer, and that essay itself appears as a separate publication in 2014 in the *Journal of Labor Economics*, volume 32, issue 1. Essay 2 is an evaluation of a math cognitive tutor (MCT) system in Chilean and Mexican public middle schools. By exploiting a clean experimental design, we show that students enrolled in schools which were randomly assigned to adopt the MCT significantly improved their standardized math test scores as compared to control group peers. This essay is part of a larger joint project with Ignacio Casas, Luis Quintero, and Paul Goodman. Finally, essay 3 contains an investigation into student hint-seeking behavior while using the MCT. I show that students who ask for the available hints early on in sections realize quicker learning gains, and that male students benefit more than female peers.

Financial support for this research has been provided by the Institute of Education Sciences (IES R305A070117 and R305D090016) and the Inter-American Development Bank (ATN/KK-11117-RS). The work is also supported by Carnegie Mellon University's Program in Interdisciplinary Education Research (PIER), funded by the US Dept of Education (R305B090023). Many employees of Carnegie Learning, including Susan Berman, Steve Ritter, and Steve Fancsali, were extremely helpful in procuing the necessary data for essay 3.

I would like to thank my committee for their feedback and help throughout my pursuit of a PhD in Economics. My parents and brother have been supportive throughout my academic career. My friends in graduate school (Billie Davis, Luis Quintero, Marco Vincenzi, Melanie Zilora, David Bergman, Will Boney, Grace Haaf, Ruth Poproski, Andy Schultz, Dan Walter, and many others) have helped with coursework, research, writing, and everything in between. I am also grateful to Lawrence Rapp for his unending support of the Tepper PhD students.

Essay 1

Bounding the Treatment Effects of Education Programs That Have Lotteried Admission and Selective Attrition

1 Introduction

The purpose of this paper is to estimate sharp bounds on treatment effects of education programs that ration excess demand by admission lotteries when selective attrition cannot be ignored. Many school districts use lotteries to determine access to over-subscribed educational programs. Lottery winners are accepted into the program, with the ultimate choice of attendance left to the student and his family. Lottery losers do not have the option to participate in the program, but have many different outside options. As a consequence, lottery losers often decide to pursue options outside of the traditional public school system and attend charter or private schools. If educational outcomes are not observed for students that leave the school system and attrition rates differ by lottery status, the randomization inherent in the lottery assignment is not necessarily sufficient to identify meaningful treatment effects. Selective attrition may also arise when lottery winners that initially participate in the program drop out because they experience unfavorable outcomes.

The starting point of our analysis is the insight that lotteries can be viewed as experimental designs with multiple sources of non-compliance that arise from parental or student decisions. Since our application focuses on magnet programs, we develop our methods in this context.[1] We focus on two of the most important outside options: parents can send their children to a non-magnet program within the district or they can leave the school district and send their children to a private school or a public school in a different district. We model this behavior as non-compliance with the intended treatment using five latent household types. It is useful to distinguish among these latent types since not all types of non-compliance lead to selective attrition problems. We face different types of missing data problems for different non-compliers.

The first type is a "complying stayer" that chooses the magnet program if it wins the lottery. The second type is a "non-complying stayer" that does not choose the magnet program even if it wins. Both of these types stay in the district regardless of lottery

[1] The methods derived in this paper apply quite broadly to many different educational programs such as charter schools and open enrollment policies.

outcome.[2] The third and fourth types leave the district if they lose the lottery. The third type is a "leaver" and will not enroll its child in the district independently of the outcome of the lottery. The fourth type complies with the lottery and participates in the magnet program if it wins the lottery and leaves if it loses. We denote these households as "at risk," since they are at risk of leaving the district. Given that many urban school districts are experiencing declining enrollment, which affects funding and district programs, this type is important from a policy perspective. Finally, there is a fifth type, the "always takers," that enrolls in the magnet option regardless of the outcome of the lottery.

The household types are latent, i.e. unobserved by both the researcher and the school district administrators.[3] Differential attrition arises in this model due to the presence of "at risk" households for whom we do not observed educational outcomes when they leave. We show how to identify and estimate the proportions of these five latent types. We also characterize differences in observed characteristics among these types. If the households that cause the differential attrition problem differ in observed characteristics from the other latent types, one may also expect that they differ in unobserved characteristics. Our approach thus allows us to characterize the extent of the differential attrition problem.

We then discuss how to estimate sharp bounds on the treatment effect of educational programs. The object of interest is the (local) average treatment effect for complying stayers. It is well-known that the standard IV estimator is only consistent if selective attrition can be ignored.[4]

[2]The district offers a standard education program to all households that do not win the lottery.

[3]Comparing our approach to the one developed in Angrist, Imbens, and Rubin (1996), note that we have two types of "never-takers" that we denote by "noncomplying stayers" and "leavers." Similarly, we have two types of "compliers" that we denote by "complying stayers" and "at risk" households. The main difference arises because individuals have more than one outside option and outcomes are not observed for "at risk" households that leave the district when they lose the lottery. These two assumptions give rise to the differential attrition problem.

[4]If there are two different types of compliers, the IV estimator does not identity a local average treatment effect. A related paper is Heckman, Urzua, and Vytlacil (2006), who also consider multiple unordered treatments with an instrument shifting agents into one of the treatments.

If point identification is not feasible, researchers have typically relied on "worst-case" scenarios to construct bounds for treatment effects. Horowitz and Manski (2000) provide a framework that exploits the assumption that the support of the outcome variable is bounded. Lee (2009) has recently proposed the use of sample trimming rules to construct more informative bounds. The basic idea of his estimator is to assume that the marginal group that only participates because of the treatment is either at the top of the bottom of the observed distribution. Our approach is in the spirit of Lee's, but uses known quantiles of the outcome distribution (test scores at the state level) to create "worst-case" scenarios. Our approach has the advantage that it does not rely on a trimming rule which is helpful when samples are small and power is an issue. Moreover, our estimator allows us to impose all orthogonality conditions that arise from our model simultaneously which can result in significant efficiency gains. This is exhibited by our empirical findings that show that our bound estimates are typically tighter than the ones obtained from the Lee estimator.[5]

Our approach also explicitly deals with heterogeneity in treatment across different schools (or job training centers, as in Lee.) Since estimation is not feasible for each school, researchers often pool data across schools. This creates an aggregation problem in estimation. Our estimators deal with the aggregation problem that is encountered when researchers have to pool among lotteries to deal with small sample problems. We show that flexible weighting schemes can be employed to estimate meaningful weighted averages of the underlying mean treatment effects.

We apply the techniques developed in this paper to study the effectiveness of magnet programs in a mid-sized urban school district. A second contribution of this paper is that we provide new research to understand the causal effects of magnet programs. While debates surrounding the effectiveness of other school choice options such as charter schools and educational vouchers have attracted much attention from researchers and policymakers,

[5]There are two related papers that use bounding methods. Dinardo, McCrary, and Sanbonmatsu (2006) develops a bounding method that requires an instrument for attrition. Blundell, et al (2007) develops bounds for the quantiles of the treatment distribution, rather than using an extreme quantile of the outcome distribution to bound the average treatment effect.

magnet programs have gotten less attention despite the fact that they are much more prevalent than charter schools or educational voucher programs.

Our findings show that magnet programs help the district to attract and retain students. Approximately 25 percent of applicants to magnet programs that serve K-5 students are "at risk." Thus selective attrition poses an important problem for the school district in our application. Households that selectively attrit come from neighborhoods that have higher incomes and are more educated than households that stay in the district regardless of the outcome of the lottery. These "at risk" households have many options outside the public school system, but apparently they view the existing magnet programs as desirable programs for their children. We also find that the market for elementary school education is more competitive than the market for middle and high school education. The fraction of households at risk declines with the age of the students.

Our findings for achievement effects are mixed. While the point estimates of the upper and lower bounds point to positive treatment effects, sample sizes are still too small to provide precise estimates. This is largely the case because standardized achievement tests were only conducted in grades 5, 8, and 11 during most of our sample period. For a variety of behavioral outcomes, we do not face these data limitations. We find that our bounds analysis is informative and demonstrates that magnet programs offered by the district improve behavioral outcomes such as offenses, attendance, and timeliness.

Our paper is related to a growing literature that evaluates educational programs using lottery based estimators.[6] Lotteries were used by Rouse (1998) to study the impact of the Milwaukee voucher program. Angrist, et al (2002) also study the effects of vouchers when there is randomization in selection of recipients from the pool of applicants using data from Colombia. Hoxby and Rockhoff (2004) use lotteries to study Chicago charter schools. Cullen, Jacob, and Levitt (2006) have analyzed open enrollment programs in the Chicago Public Schools. Ballou, Goldring, and Liu (2006) examine a magnet program. Hastings, Kane, Staiger (2008) estimate a model of school choice based on stated preferences for

[6] Angrist (1990) introduced the use of lotteries to study the impact of military service on earnings.

schools in Charlotte. Since school attendance was partially the outcome of a lottery, they use the lottery outcomes as instruments to estimate the impact of attending the first choice school. Abdulkadiroglu, et al (2009) and Hoxby and Murarka (2009) study charter schools in Boston and New York, respectively, and find strong achievement effects. Dobbie and Fryer (2009) study a social experiment in Harlem and show that high-quality schools or high-quality schools coupled with community investments generate the achievement gains. All of these papers focus on applications in which selective attrition is not present and thus do not explicitly deal with the key selective attrition problem discussed in this paper.[7]

The rest of the paper is organized as follows. Section 2 develops our new methods for estimation of treatment effects when program participation is partially determined by lotteries and selective attrition cannot be ignored. We discuss identification and estimation. Section 3 provides some institutional background for our application and discusses our main data sources. Section 4 reports the empirical findings of our paper. Finally, we offer some conclusions and discuss the policy implications of our work in Section 5.

2 Identification and Estimation

2.1 The Research Design

We consider a design that arises when randomization determines eligibility to participate in an educational program. Consider the problem of a parent that has to decide whether or not to enroll a student in a magnet program offered by a school district.[8] We only consider households that participate in a lottery that determines access to an oversubscribed (magnet) program. Let W denote a discrete random variable which is equal to 1 if the student wins the lottery and 0 if it loses. Let w denote the fraction of households that win the lottery.

[7] Angrist, et al (2002) encounter a related issue of selective test participation since students in private schools are more likely to take college entrance exams than public school students.

[8] We use the terms "parent" or "households" to describe the decision maker and "student" to describe the person that participates in the program.

We assume that a student who wins the lottery has three options: participate in the magnet program, participate in a different, non-magnet program offered by the same school district, or leave the district and pursue educational opportunities outside the district. A student who loses and is not an always-taker has only the last two options. Let M be 1 if a student attends the (magnet) program and 0 otherwise. Finally, let A denote a random variable that is 1 if a student attends a school in the district and 0 otherwise.

To model compliance with the intended treatment, we use five latent types to classify households into compliers and non-compliers. We make the following assumption.

Assumption 1

1. *Let s_m denote the fraction of "complying stayers." These households will remain in the district when they lose the lottery. If they win the lottery, they comply with the intended treatment and attend the magnet school.*

2. *Let s_n denote the fraction of "noncomplying stayers." These households will remain in the district when they lose the lottery. If they win the lottery, they will not comply with the intended treatment and instead will attend a non-magnet program in the district.*

3. *Let l denote the fraction of "leavers." These are households that will leave the district regardless of whether they are admitted to the magnet program.*[9]

4. *Let r denote the fraction that is "at risk." These households will remain in the district and attend the magnet program if admitted to the magnet program, and they will leave the district otherwise.*

[9]Parents have incomplete information and need to gather information to learn about the features of different programs. Parents have to sign up for lotteries months in advance. At that point, they have not accumulated all relevant information. Once they have accumulated all relevant information, they may decide to opt out of the public school system if their preferred choice dominates the program offered by the district. In addition, household circumstances may change. For example, parents may obtain a job that requires moving to a different metropolitan area. Note that there are typically no penalties for participating in the lottery and declining to participate in the program.

5. *Let a_t denote the fraction of "always takers." They will attend the magnet school regardless of the outcome of the lottery.*

Since the household type is latent, one key empirical problem is identifying and estimating the proportions of each type in the underlying population. These parameters are informative about the effectiveness of magnet programs in attracting and retaining households that participate in the lottery. Moreover, we will show that households "at risk" cause the selective attrition problem.

The latent types of households are likely to differ in important characteristics and we need to characterize these differences. If households "at risk" differ among observed characteristics from the other latent types, one may also expect that they differ by unobserved characteristics. As a consequence, ignoring the selective attrition problem will be problematic. By characterizing the observed characteristics of all latent types, we can thus gain some important insights into the potential importance of the selective attrition problem.

To formalize these ideas, consider a random variable X that measures an observed household characteristic such as income or socio-economic status. Appealing to our decomposition, let μ_r, μ_{s_m}, μ_{s_n}, μ_l and μ_{a_t} denote the means of random variable X conditional on belonging to group r, s_m, s_n, l, and a_t, respectively. The goal of the first part of the analysis is then to identify and estimate the following eleven parameters $(w, r, s_n, s_m, l, a, \mu_r, \mu_{s_n}, \mu_{s_m}, \mu_l, \mu_{a_t})$.[10]

The next objective is to study the effects of the program on student outcomes. Let T be an outcome measure of interest, for example, the score on a standardized achievement test. Following Fisher (1935), we adopt standard notation in the program evaluation literature and consider a model with three potential outcomes:

$$T \;=\; A\,M\,T_1 \;+\; A\,(1-M)\,T_0 \;+\; (1-A)\,T_2 \tag{1}$$

where T_1 denotes the outcome if the student attends the magnet school, T_0 if he attends a different program in the district, and T_2 if he attends a school outside of the district.[11]

[10] It is straightforward to allow X to be a vector.

[11] This approach shares many similarities with the "switching regression" model introduced into economics

We will later assume that T is not observed for students that do not attend a public school within the district, i.e. iT_2 is not observed. This assumption is plausible since researchers typically only have access to data from one school district. Private schools rarely provide access to their confidential data and often do not administer the same standardized tests as public schools. Attention, therefore, focuses on the individual treatment effect $\Delta = T_1 - T_0$. Note that Δ is unobserved for all students. Conceptually, we can define five different average treatment effects, one for each latent group.[12]

$$ATE_{Type} = E[T_1 - T_0 | Type = 1] \qquad Type \in \{S_n, S_m, R, L, A_t\} \tag{2}$$

The key research question is then whether we can identify and estimate these types of treatment effects when selective attrition is important. To answer this question, we first discuss how to characterize the extent of the selective attrition problem. We then derive bounds estimators for the relevant treatment effects.

2.2 Identification of the Fraction of Latent Types

First we need to establish the information set of the researcher.

Assumption 2 *The researcher observes probabilities and conditional means for the feasible outcomes shown in Table 1.*

Note that only six of the eight outcomes listed in Table 1 are possible since a student attending a magnet program ($M = 1$) must also attend a public school ($A = 1$).

Identification can be established sequentially. First, we discuss identification of the probabilities that characterize the shares of the latent types. We have the following result.

by Quandt (1972), Heckman (1978, 1979), and Lee (1979). Heckman and Robb (1985) and Bjorklund and Moffitt (1987) treated heterogeneity in treatment as a random coefficients model. It is also known in the statistical literature as the Rubin Model developed in Rubin (1974, 1978). See also Heckman and Vytlacil (2007) for an overview of the program evaluation literature.

[12] There are other effects that may also be of interest such as treatment effect on the treated or the marginal treatment effect. For a discussion see, among others, Heckman and Vytlacil (2005) and Moffitt (2008).

Table 1: Observed Outcomes

	W	M	A	$Pr\{W, M, A\}$	$E[X\|W, M, A] = E[X\|M, A]$
I	1	1	1	$w\ (r + s_m + a_t)$	$\frac{r\mu_r + s_m\mu_{s_m} + a_t\mu_{a_t}}{r+s_m+a_t}$
II	1	1	0	not possible	
III	1	0	1	$w\ s_n$	μ_{s_n}
IV	1	0	0	$w\ l$	μ_l
V	0	1	1	$(1-w)a_t$	μ_{a_t}
VI	0	1	0	not possible	
VII	0	0	1	$(1-w)\ (s_n + s_m)$	$\frac{s_n\mu_{s_n} + s_m\mu_{s_m}}{s_n+s_m}$
VIII	0	0	0	$(1-w)\ (r + l)$	$\frac{r\mu_r + l\mu_l}{r+l}$

Proposition 1 *The parameters (w, r, s_n, s_m, l, a) are identified by the six non-degenerate probabilities in Table 1.*

Proof: Parameter w is the fraction that wins the lottery:

$$\begin{aligned} w &= Pr(W=1, M=1, A=1) + Pr(W=1, M=1, A=0) \\ &+ Pr(W=1, M=0, A=1) + Pr(W=1, M=0, A=0) \end{aligned} \tag{3}$$

Given w, s_n is identified from (1,0,1):

$$s_n = Pr(W=1, M=0, A=1)/w \tag{4}$$

l is identified from (1,0,0):

$$l = Pr(W=1, M=0, A=0)/w \tag{5}$$

a_t is identified from (0,1,1):

$$a_t = Pr(W=0, M=1, A=1)/(1-w) \tag{6}$$

Given w and s_n, s_m is identified from (0,0,1):

$$s_m = Pr(W=0, M=0, A=1)/(1-w) - s_n \tag{7}$$

Given a_t, l, s_n, and s_m, r is identified of the identity:

$$r = 1 - l - s_m - s_n - a_t \tag{8}$$

Q.E.D.

Note that there is no over-identification at this stage since the six probabilities in Table 1 add up to one, and the last three non-degenerate probabilities add up to $1 - w$.

Next we discuss identification of the five conditional means of household characteristics. We have the following result.

Proposition 2 *Given (w, r, s_n, s_m, l, a_t), the parameters $(\mu_r, \mu_{s_m}, \mu_{s_n}, \mu_l, \mu_{a_t})$ are identified by the observed conditional expectations observed in Table 1.*

Proof: μ_l is identified from (1,0,0):

$$\mu_l = E(X|W = 1, M = 0, A = 0) \tag{9}$$

Similarly μ_{s_n} is identified from (1,0,1):

$$\mu_{s_n} = E(X|W = 1, M = 0, A = 1) \tag{10}$$

and μ_{a_t} is identified from (0,1,1):

$$\mu_{a_t} = E(X|W = 0, M = 1, A = 1) \tag{11}$$

Given μ_{s_n}, μ_{s_m} is identified from (0,0,1):

$$\mu_{s_m} = [(s_n + s_m)E(X|W = 0, M = 0, A = 1) - s_n\mu_{s_n}]/s_m \tag{12}$$

Given μ_{s_m} and μ_{a_t}, μ_r is identified from (1,1,1):

$$\mu_r = [(r + s_m + a_t)E(X|W = 1, M = 1, A = 1) - s_m\mu_{s_m} - a_t\mu_{a_t}]/r \tag{13}$$

Q.E.D.

There is one over-identifying condition at this stage. This restriction arises due to the condition that W is orthogonal to X.[13] Propositions 1 and 2 then imply that the parameters $(w, r, s_n, s_m, l, a_t, \mu_r, \mu_{s_n}, \mu_{s_m}, \mu_l, \mu_{a_t})$ are identified. We can thus study the effectiveness of magnet programs to attract and retain students. Moreover, the fraction of households that are "at risk" is the key parameter that measures the selective attrition between lottery winners and losers. We show this in the next section.

2.3 Identification of Treatment Effects

We now turn to the analysis of identification of causal treatment effects of magnet programs on educational and behavioral outcomes. We assume that the researcher only observes outcomes, T, for students that remain in the school district, i.e. we do not observe outcomes for "leavers" and "at risk" households that lose the lottery.

It is useful to assume initially that we observe the latent household type. Table 2 provides a summary of the relevant conditional expectations.[14] Conditioning on lottery outcomes, there are ten conditional expectations. Three of these pertain to outcomes that are not observed since students in these latent groups leave the school district (T_2). The remaining seven conditional expectations relate to household types that remain in the district.

From Table 2, it is evident that even if we observed the latent types, there is little hope in identifying ATE_{S_n}, ATE_R, ATE_L, or ATE_{A_t}. For stayers that never attend the magnet program, we cannot identify $E[T_1|S_n = 1]$. For students at risk, we cannot identify $E[T_0|R = 1]$. For leavers, we can neither identify $E[T_1|L = 1]$ nor $E[T_0|L = 1]$. For always-takers we never observe $E[T_0|A_t = 1]$. Without imposing additional assumptions on the selection of students into latent groups, ATE_{S_n}, ATE_R, ATE_L and ATE_{A_t} are not identified. Attention, therefore, focuses on identification of ATE_{S_m}. Note that ATE_{S_m}

[13] The lotteries are assumed to be fair and blind in the sense that the district does not pre-select winners and losers based on beliefs about attendance or any socio-economic or student characteristic found in X.

[14] Note that we are implicitly assuming that the mean performance of stayers who would decline lottery admission is the same whether they win or lose the lottery, i.e. $E[T_0|S_n = 1, W = 1] = E[T_0|S_n = 1, W = 0] = E[T_0|S_n = 1]$.

Table 2: Mean Outcomes Conditional on Type

	Complying Stayers	Non-Complying Stayers	At Risk	Leavers	Always Takers
$W = 1$	$E[T_1 \mid S_m = 1]$	$E[T_0 \mid S_n = 1]$	$E[T_1 \mid R = 1]$	$E[T_2 \mid L = 1]$	$E[T_1 \mid A_t = 1]$
$W = 0$	$E[T_0 \mid S_m = 1]$	$E[T_0 \mid S_n = 1]$	$E[T_2 \mid R = 1]$	$E[T_2 \mid L = 1]$	$E[T_1 \mid A_t = 1]$

Note that T_2 is never observed.

would be identified if types were not latent. Of course, household types are not observed and as a consequence identification of ATE_{S_m} is not straightforward. One key result of this paper is that the local average treatment effect for compliers is not point identified if there is selective attrition.

Proposition 3

If there is selective attrition ($r \neq 0$) and if households that are at risk have different expected outcomes than compliers in the treated case ($E[T_1|S_m = 1] \neq E[T_1|R = 1]$), then the local average treatment effect for compliers, ATE_{S_m}, is not identified.

Proof:

We only observe mean outcomes for the students conditional on W, M and A. For students who win the lottery and attend the magnet school, we observe

$$E[T|W = 1, M = 1, A = 1] \;=\; \frac{s_m E[T_1|S_m = 1] + rE[T_1|R = 1] + a_t E[T_1|A_t = 1]}{s_m + r + a_t} \tag{14}$$

For students who lose the lottery and attend the magnet school, we observe

$$E[T|W = 0, M = 1, A = 1] \;=\; E[T_1|A_t = 1] \tag{15}$$

We also observe mean performance of stayers who lose the lottery:

$$E[T|W = 0, M = 0, A = 1] \;=\; \frac{s_m E[T_0|S_m = 1] + s_n E[T_0|S_n = 1]}{s_m + s_n} \tag{16}$$

Finally, we also observe the mean performance of stayers who win the lottery and decline to enroll in the magnet program:

$$E[T|W = 1, M = 0, A = 1] \;=\; E[T_0|S_n = 1] \tag{17}$$

Equations (16) and (17) imply that we can identify $E[T_0|S_m = 1]$ and $E[T_0|S_n = 1]$, since s_n and s_m have been identified before. Equation (15) implies that we can identify $E[T_1|A_t = 1]$. However, equation (14) then implies that we cannot separately identify $E[T_1|S_m = 1]$ and $E[T_1|R = 1]$. Q.E.D.

Proposition 3 illustrates that attrition *per se* is not the problem. If the fraction of "at risk" households is negligible (i.e., $r = 0$), identification is achieved even if the fraction of leavers is large.[15] The lack of point identification arises from the "at risk" households which cause the selective attrition problem. Selective attrition is only a problem if "at risk" households have different mean outcomes than compliers.[16]

Since point identification is no longer feasible when selective attrition is not negligible, attention focuses on set identification and the construction of bounds.[17]

Proposition 4

i) Suppose we have an upper bound, denoted by T_1^u, for $E[T_1|R = 1]$ i.e. T_1^u satisfies $E[T_1|R = 1] \leq T_1^u$. We can then construct a lower bound for the $E[T_1|S_m = 1]$ and ATE_{S_m}.

ii) Suppose we have a lower bound, denoted by T_1^l, for $E[T_1|R = 1]$, i.e. T_1^l satisfies $E[T_1|R = 1] \geq T_1^l$, we can then construct an upper bound for the $E[T_1|S_m = 1]$ and ATE_{S_m}.

[15] Recall that if $r = l = 0$ our research design simplifies to the one considered in Angrist, Imbens and Rubin (1996).

[16] We can generalize Proposition 3 by assuming that $E[T_1|S_m = 1, X] \neq E[T_1|R = 1, X]$, i,e, by conditioning on some observables X. If controlling for selection on observables is sufficient to deal with the selection problem, a matching approach can be justified. For a discussion of matching estimators, see, among others, Rosenbaum and Rubin (1983), Heckman, Ichimura, and Todd (1997), and Abadie and Imbens (2006).

[17] Point identification cannot be achieved in many econometric applications. In that case, attention naturally shifts to characterizing informative bounds on the parameters of interest. See, for example, Manski (1997), Horowitz and Maski (2000), Imbens and Manski (2004), Chernozhukov, Imbens, and Newey (2006), and Lee (2009).

Proof:

Consider the first part of the statement. Equation (14) then implies that:

$$\begin{aligned} & E[T_1|S_m = 1] \\ = & \frac{s_m + r + a_t}{s_m} E[T|W = 1, M = 1, A = 1] - \frac{rE[T_1|R = 1] + a_t E[T_1|A_t = 1]}{s_m} \\ \geq & \frac{s_m + r + a_t}{s_m} E[T|W = 1, M = 1, A = 1] - \frac{rT_1^u + a_t E[T_1|A_t = 1]}{s_m} \end{aligned} \tag{18}$$

where the last inequality follows from $E[T_1|R = 1] \leq T_1^u$. Since all terms in the last line of equation (18) are identified, we conclude that we can construct a lower bound. Replacing T_1^u with T_1^l and reversing the inequality yields the upper bound. Q.E.D.

There are many ways of constructing both lower bounds or upper bounds depending on the outcome variable and the application. For example, a plausible assumption for the construction of an upper bound of the mean treatment effect is that the "at risk" households are at least as good as the compliers, $T_1^l = E[T_1|S_m = 1] \leq E[T_1|R = 1]$.

A better approach that we explore in this paper is to bound outcomes using known percentiles of the outcome distribution. These type of aggregate distributions are often available in applications in education at the state level, as we discuss in detail in the next section.

Alternatively, we can apply the trimming approach suggested by Lee (2009). This approach is applied in our context by first ordering magnet students from lowest to highest performance on the outcome variable being studied. Then treatment observations are dropped from the sample based both on the proportions of missing data in the control and treatment groups and the distribution of the outcome variable being bounded.

We have thus seen that selective attrition implies that we have to focus on the construction of bounds since point identification is not feasible. It is therefore important to have a simple test to determine whether r is zero. If we cannot reject the null hypothesis that $r = 0$, treatment effects are point identified and can be estimated using standard linear IV estimators. A simple way to estimate r is to regress A_i on W_i. The slope coefficient in that

regression is equal to r. At minimum, researchers that work with lottery data in educational applications should run this regression and test whether one of the key identifying assumptions of the IV estimator is valid. If we reject the null that r is equal to zero, the bounds analysis suggested in this paper is more appropriate than IV estimation.

2.4 A GMM Estimator

Suppose we observe a random sample of N applicants to an education program, indexed by i. We view these as N independent draws from the underlying population of all applicants to this program. Let W_i, M_i, A_i, and X_i now denote the random variables that correspond to observation i. The proofs of identification are constructive. Replacing population means by sample means thus yields consistent estimators for the parameters of interest. Nevertheless, it is useful to place the estimation problem within a well defined GMM framework. This allows us to estimate simultaneously all parameters and compute asymptotic standard errors. We can estimate the fractions of each latent type based on moment conditions derived from the choice probabilities in Table 1. Define:

$$f_1(A_i, M_i, W_i) \quad = \quad \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\Big[W_i M_i A_i - w(r + s_m + a_t)\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[W_i(1-M_i)A_i - w\ s_n\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[W_i(1-M_i)(1-A_i) - w\ l\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[(1-W_i)M_i A_i - (1-w)\ a_t\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[(1-W_i)(1-M_i)A_i - (1-w)(s_n + s_m)\Big] \end{cases}$$

and note that $E[f_1(A_i, M_i, W_i)] = 0$. Similarly we can estimate the mean characteristics of each type. Define:

$$f_2(A_i, M_i, W_i, X_i) \quad = \quad \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\Big[W_i M_i A_i X_i - w[r\mu_r + s_m\mu_{s_m} + a_t\mu_{a_t}]\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[W_i(1-M_i)A_i X_i - w\ s_n\ \mu_{s_n}\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[W_i(1-M_i)(1-A_i)X_i - w\ l\ \mu_l\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[(1-W_i)M_i A_i X_i - (1-w)a_t\ \mu_{a_t}\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[(1-W_i)(1-M_i)A_i X_i - (1-w)[s_n\mu_{s_n} + s_m\mu_{s_m}]\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[(1-W_i)(1-M_i)(1-A_i)X_i - (1-w)[r\mu_r + l\mu_l]\Big] \end{cases}$$

and note that $E[f_2(A_i, M_i, W_i)] = 0$. Finally, we can construct additional orthogonality conditions to construct both upper and lower bounds. Consider first the case of estimating an upper bound for compliers, denoted by $E[T_1^u|S_m = 1]$, by setting the lower bound for $E[T_1|R = 1]$ to the 5th percentile of the observed outcome distribution, denoted by T_1^l. Define:

$$f_3(A_i, M_i, W_i, T_i) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\Big[T_iW_iM_iA_i - w(s_mE[T_1^u|S_m = 1] + rT_1^l + a_tE[T_1|A_t = 1])\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[T_i(1-W_i)M_iA_i - (1-w)a_tE[T_1|A_t = 1]\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[T_i(1-W_i)(1-M_i)A_i - (1-w)(s_mE[T_0|S_m = 1] + s_nE[T_0|S_n = 1])\Big] \\ \frac{1}{N}\sum_{i=1}^{N}\Big[T_iW_i(1-M_i)A_i - ws_nE[T_0|S_n = 1]\Big] \end{cases}$$

and we have $E[f_3(A_i, M_i, W_i)] = 0$. Similarly, we can construct an orthogonality condition for the lower bound if we use the 95th percentile outcome for T_1^u. This value comes from state level data for test scores and from our sample of non-missing data for all other outcomes. Combining all orthogonality conditions, we can estimate the parameters of the model using a GMM estimator (Hansen, 1982). Note that the estimator above easily generalizes to the case in which X is a vector of random variables. We simply stack all orthogonality conditions to obtain a simultaneous estimator. The main advantage of the GMM framework is that we can estimate all parameters jointly by imposing all relevant orthogonality conditions. Moreover it is straightforward to obtain standard errors for the upper and lower bounds using a GMM framework. Many of the parameters of the model – especially all parameters that characterize the fraction of latent types – can be estimated using linear estimators.[18] We find in the application that imposing the additional orthogonality conditions that model the mean characteristics of the types ($f_2(A_i, M_i, W_i, X_i)$ above) yields significant efficiency gains.

Thus far we have considered the problem of estimating causal effects using data from one lottery. In practice, researchers often need to pool data from multiple lotteries to obtain large enough sample sizes. We discuss in detail in Appendix A of this paper the problems that are encountered when aggregating across lotteries. Using a suitable weighting procedure, we show that we can estimate weighted averages of the underlying parameters of

[18] An appendix is available upon request which shows exactly how to set up the linear estimators.

the model. Weights can be chosen in accordance to the objectives of the policy or decision maker.

3 Data

Our application focuses on magnet programs that are operated by a mid-sized urban school district that prefers to not be identified. Magnet schools emerged in the United States in the 1960's. Magnet schools are designed to draw students from across normal attendance zones. In contrast, a feeder school typically only admits students that live inside the attendance zone. As a consequence, the composition of feeder schools reflects residential choices of parents and is largely driven by the composition of local neighborhoods. Magnet schools were thus initially used as a way to reduce racial segregation in public schools.

More recently, magnet programs have been viewed as attractive options to increase school choice, to retain students with better socio-economic backgrounds in public schools, and to increase student achievement. In some cases, magnet programs are housed in separate schools. But they can also be a program within a more comprehensive school. Magnet programs offer specialized courses or curricula. There are magnet programs for all grade levels in our district. We only consider magnet programs that are academically oriented. These magnet programs typically provide specialized education in mathematics, the sciences, languages, or humanities. Other magnet programs have a broader focus on topics such as international studies or performing arts.

Every academic year, interested students submit applications for one magnet program of their choice. Some magnet programs in the district have a competitive entrance process, requiring an entrance examination, interview, or audition. We do not include these magnet programs in this study since the admission procedure does not use randomization. Instead we focus on magnet programs that do not have competitive entrance procedures. If the number of applications submitted during registration for any magnet program exceeds the number of available spaces, the district holds a lottery to determine the order in which

applicants will be accepted.

In the case of over-subscription, a computerized random selection determines each student's lottery number. The lottery is binding in the sense that students with lower numbers are accepted, and higher numbered students are rejected. There is a clear cut-off number that separates the groups. We do not observe students attending magnet schools that lose the lottery, i.e. there are no "always-takers" in our sample.

To preserve racial balance in the magnet programs, separate lotteries are held for black students and other students. Some programs also have preferences for students with siblings already attending the magnet programs or for students who live close to the school. Separate lotteries are held for those students with an acceptable preference category for each magnet program. All in all, each lottery is held for a given program, in a given academic year, separately by race, and, finally, separately by preference code.

Lottery winners (lotteried-in) have the option to participate in the magnet program, with the ultimate choice of participation left to the student and his family. Lottery losers (lotteried-out) do not have this option, and thus must make their schooling choice without the availability of the magnet option. When winners decline admission, the students on the wait list become eligible. Again the rank on the wait list is determined by the original lottery. With a fair and balanced lottery, the winners and losers will be determined by chance, thus creating two groups that are similar to each other both on observable and unobservable characteristics.

The district granted us access to its longitudinal student database. We use data from the 1999-2000 school year through 2005-2006. In addition to demographic data, the database contains detailed information about educational outcomes. This information is linked to each student by a unique ID number. The demographic characteristics for the students include race, gender, free/reduced lunch eligibility, and addresses.[19] Using the addresses, we can assign census tract level variables to each student. We use two community characteris-

[19]The race variable is one if a student is African American and zero otherwise. The gender variable is one for girls and zero for boys.

Table 3: Descriptive Statistics

Variable	*Entire Sample (2054 obs)*	*Elem School (820 obs)*	*Middle School (457 obs)*	*High School (777 obs)*
Gender	0.51 (0.50)	0.51 (0.50)	0.51 (0.50)	0.51 (0.50)
Race	0.75 (0.44)	0.59 (0.49)	0.79 (0.40)	0.88 (0.32)
FRL	0.33 (0.47)	0.33 (0.47)	0.35 (0.48)	0.32 (0.47)
Poverty	0.23 (0.14)	0.22 (0.14)	0.23 (0.14)	0.24 (0.15)
Education	0.29 (0.19)	0.34 (0.22)	0.28 (0.18)	0.25 (0.14)
Offenses	0.99 (2.23)	0.18 (0.99)	1.15 (2.32)	1.67 (2.71)
Suspension Days	1.88 (4.71)	0.29 (1.62)	1.97 (4.39)	3.32 (6.17)
Absences	13.28 (14.56)	8.74 (7.96)	10.30 (8.54)	19.30 (19.30)
Tardies	7.31 (13.10)	3.94 (7.03)	8.66 (12.89)	9.70 (16.55)
Win Percentage	61.8	52.1	53.2	77.1

tics that measure the socio-economic composition of the neighborhoods in which students reside. Poverty is the percentage of adults in the student's census tract with an income level below the poverty line. Education is the percentage of adults in the student's census tract with at least a college degree.

As pertaining to student educational outcomes, the database includes the school of attendance in each year and standardized scores for the state assessment tests. In addition, we observe a variety of behavioral outcome measures such as offenses, suspensions, and absences. The database also contains the outcomes of the magnet lotteries. One of the key features of the database is that it contains unusually good information about students residing in the district that attend private, charter, and home schools. Unfortunately, we do not observe test scores or behavioral outcome measures for students outside of the district. Table 3 shows descriptive statistics for the entire sample used in this study as well as three important sub-samples that we also consider in estimation.[20] We only consider binding lotteries in this research. In total, over the time frame of the data, there are 173 binding lotteries with 1,269 students lotteried-in and 785 students lotteried-out.

Before we implement the estimators, we check whether the lotteries are balanced on student observables. While assignment within lotteries may be random, participation in a lottery is not. To make use of within-lottery randomness and not the between-lottery non-randomness, we perform a check for balance by running a lottery-fixed effect regression for each observable characteristic as a dependent variable with acceptance as the only independent variable other than the fixed effects. Separate lotteries are held by race, so race is left out of the balance analysis. We test every other observable student characteristic in the data set.

Following Cullen, Jacob, and Levitt (2006) we use equation (7) to determine whether

[20] For a small sample of students we imputed absences and tardies. Also note that outcome variables are not observed for students that leave the district. Thus the means of the outcome variables in Table 3 reflect means of stayers.

Table 4: Lottery Balance Result

Variable	*Entire Sample*	*Elem School*	*Middle School*	*High School*
Gender	0.0053	0.0366	-0.0183	-0.0257
	(0.0262)	(0.0384)	(0.0559)	(0.0469)
FRL	0.0056	0.0111	-0.0501	0.0385
	(0.0229)	(0.0322)	(0.0482)	(0.0431)
Poverty	-0.0050	-0.0023	0.0044	-0.0160
	(0.0068)	(0.0092)	(0.0136)	(0.0135)
Education	0.0041	0.0110	-0.0038	-0.0007
	(0.0078)	(0.0127)	(0.0165)	(0.0125)

the lottery is balanced:

$$X_i = \beta_1 W_i + \sum_{j=1}^{J} I_{ij}\beta_{2j} + v_i \tag{19}$$

where X_i is the observable characteristic of interest, W_i is a dummy equal to 1 if student i wins lottery j, I_{ij} is an indicator variable equal to 1 if student i participated in lottery j, and v_i is the error term.[21] We estimate a separate regression for each observable. The coefficient β_1 determines the fairness of the lottery system. If we cannot reject the null hypothesis that it is equal to zero, then acceptance into a magnet is not determined by the value of that particular student observable, X.

The first column of Table 4 shows the results when all students in all binding lotteries are included in the regressions. β_1 is not significant for any tested variable at 10 %. The second and third columns consider the three sub-samples of interest. The second column includes all students in elementary school while the third column focuses on middle school students and the fourth on high school students. We find that the estimates of β_1 are not significantly

[21] Alternatively we could use multivariate Behrens-Fisher type test statistics which require less restrictive assumptions. See, for example, Kim (1992)

different from zero. We thus find that the lotteries are fair, creating separate winner and loser groups that are similar in observed characteristics. Any differences between winners and losers are small and statistically insignificant. This holds for the overall population in binding lotteries and for the smaller sub-samples that were tested.

4 Empirical Results

4.1 Attraction, Retention and Selective Attrition

To study the importance of selective attrition in our sample, we implement a number of different estimators. First, we use a GMM estimator that only imposes the orthogonality conditions that identify the fraction of latent household types. Then we add the orthogonality conditions that capture the mean characteristics of the types. The characteristics include race, gender, free or reduced lunch, poverty, and college education. Recall that the last two measures are based on neighborhood characteristics as reported by the U.S. Census. We report estimates for three samples which include all students that applied to an oversubscribed magnet program that is associated with an elementary school (ES), middle school (MS), and high school (HS), respectively. We pool across all lotteries in each sample and, therefore, use the weighted estimator discussed in Appendix A. Tables 5 and 6 report the point estimates and estimated standard errors for each of the three samples.

Comparing the estimates in the upper and lower panels of Table 5 clearly allows us to evaluate whether there are efficiency gains that arise when using a GMM estimator.[22] We find that there are significant efficiency gains in the estimates of two key parameters, the fraction of compliers and the fraction at risk. Estimated standard errors are up to 50 percent larger when one ignores the additional orthogonality conditions. We thus conclude that our approach of jointly estimating the model using GMM is preferable to simpler methods.

Table 5 reveals some interesting new insights into the importance of selective attrition

[22] This comparison is also interesting since the GMM estimates and associated standard errors in the upper panel are identical to the results that could be obtained using simpler linear estimators.

Table 5: Empirical Results: Selective Attrition

	First Set of Orthogonality Conditions			
	Fraction At Risk	Fraction Stay, Attend	Fraction Stay, Non	Fraction Leave
ES	0.25 (0.04)	0.61 (0.05)	0.06 (0.01)	0.08 (0.01)
MS	0.12 (0.15)	0.60 (0.16)	0.24 (0.04)	0.04 (0.01)
HS	0.15 (0.09)	0.70 (0.09)	0.08 (0.01)	0.06 (0.01)
	First and Second Set of Orthogonality Conditions			
	Fraction At Risk	Fraction Stay, Attend	Fraction Stay, Non	Fraction Leave
ES	0.25 (0.04)	0.61 (0.04)	0.06 (0.01)	0.08 (0.01)
MS	0.12 (0.05)	0.61 (0.06)	0.24 (0.04)	0.04 (0.01)
HS	0.14 (0.06)	0.72 (0.06)	0.08 (0.01)	0.06 (0.01)

Estimated standard errors are reported in parentheses.

in our application. Recall that the fraction of households at risk is the key parameter that captures selective attrition. We find that selective attrition is substantial and ranges between 12 and 25 percent across our three samples. We also find that the majority of students will stay in the district regardless of the outcome of the lottery. The majority, 61 to 71 percent, will attend the magnet program if they win they lottery. The fraction of households that will leave the district regardless of the outcome of the lottery ranges between 4 and 8 percent. Overall, these results suggest that most households consider the magnet programs desirable. We conclude that magnet programs are effective tools for attracting and retaining households and students.

Equally interesting are the observed mean characteristics of the latent types of households reported in Table 6. These and the ones reported in the lower part of Table 5 are the results from the first and second set of orthogonality conditions (f_1 and f_2). For each characteristic, the differences across household types (at risk, leavers, stayers) are statistically

Table 6: Empirical Results: Characteristics

Gender				
	At Risk	Stay, Attend	Stay, Non	Leave
ES	0.57 (0.09)	0.47 (0.03)	0.55 (0.11)	0.47 (0.08)
MS	0.85 (0.34)	0.43 (0.06)	0.50 (0.08)	0.31 (0.13)
HS	0.55 (0.34)	0.57 (0.05)	0.49 (0.08)	0.41 (0.08)
Race				
	At Risk	Stay, Attend	Stay, Non	Leave
ES	0.50 (0.09)	0.70 (0.04)	0.39 (0.11)	0.18 (0.07)
MS	0.99 (0.41)	0.80 (0.05)	0.80 (0.06)	0.28 (0.14)
HS	0.89 (0.41)	0.93 (0.03)	0.85 (0.07)	0.79 (0.06)
FRL				
	At Risk	Stay, Attend	Stay, Non	Leave
ES	0.12 (0.04)	0.43 (0.03)	0.19 (0.07)	0.07 (0.04)
MS	0.26 (0.15)	0.47 (0.06)	0.26 (0.09)	0.07 (0.06)
HS	0.15 (0.11)	0.39 (0.04)	0.25 (0.06)	0.12 (0.05)
Poverty				
	At Risk	Stay, Attend	Stay, Non	Leave
ES	0.21 (0.03)	0.23 (0.01)	0.20 (0.04)	0.14 (0.01)
MS	0.24 (0.10)	0.24 (0.02)	0.23 (0.02)	0.13 (0.02)
HS	0.28 (0.12)	0.25 (0.01)	0.24 (0.02)	0.19 (0.02)
Education				
	At Risk	Stay, Attend	Stay, Non	Leave
ES	0.40 (0.05)	0.29 (0.02)	0.41 (0.05)	0.53 (0.04)
MS	0.20 (0.11)	0.29 (0.02)	0.30 (0.03)	0.55 (0.08)
HS	0.27 (0.14)	0.25 (0.01)	0.21 (0.02)	0.36 (0.03)

Estimated standard errors are reported in parentheses.

significant. We find that "at risk" households are on average less likely to be African American and on free or reduced lunch programs than households that are stayers. Moreover, they come from better educated neighborhoods.[23] These differences are more pronounced at the elementary school level where the fraction of "at risk" households is the greatest. We thus conclude that magnet programs are effective devices for the school district to retain more affluent households. Not surprisingly, the leavers are the most affluent group and come from neighborhoods with the highest levels of education. These households may just apply to the magnet programs as a back-up option in case their students should unexpectedly not be admitted to an independent, charter, or parochial school.[24]

The demographic differences, summarized above, between "at risk" students and "stayers" drive our assumptions on the bounds. Poor minority students are known to perform poorly in school compared to wealthier majority peers (Dobbie and Fryer, 2009). Therefore, our upper bound estimation assumes that the "at risk" students are only as good as the "stayers," while the lower bound estimation assumes that the at risk students are in the 95th percentile of the outcome distribution.

Table 6 also permits interesting comparisons across grade levels. Elementary and middle school lotteries are somewhat more competitive than high school lotteries. The former have average win rates of 52 percent and 53 percent respectively while the latter have an average win rate of 77 percent. Elementary programs attract a clientele from more highly educated neighborhoods. The fraction of African American families is also lower among applicants to elementary school lotteries. Not surprisingly, we find that the fraction of at risk families and the fraction of leavers is also higher among elementary school students. These findings highlight the fact that, among the magnet school applicants, the market for elementary school education is more competitive than the market for high school education.

[23] Note that the differences in household characteristics are statistically significant from zero at all conventional levels.

[24] It could also be that these households left the district because of job transfers or other issues unrelated to schools.

4.2 Treatment Effects

We have seen in the previous section that the fraction of "at risk" households is large and significantly different from zero in our application in all three samples. Moreover, households that are "at risk" of leaving the district have more favorable socio-economic characteristics than other types except for "leavers". As a consequence, we conclude that selective attrition cannot be ignored in this application. Since treatment effects are only set-identified when selective attrition matters, we implement our bounds estimators. We implement our bounds estimators by adding the orthogonality conditions for these variables to the conditions, discussed in Section 4.1, for estimating the proportions of latent types and the demographic characteristics of latent types. For comparison purposes, we also report the IV estimates that ignore selective attrition.

We start our analysis by focusing on achievement effects. The main problem encountered in this part of the analysis arises due to missing data. This is largely the case because standardized achievement tests were only conducted in grades 5, 8, and 11 during most of our sample period. For our middle school sample, there are only 155 observations for which we have test scores. For the high school sample, the reduction is of similar magnitude.[25] Including households that participate in the lotteries but subsequently leave the district gives us with 213 middle school students and 203 high school students. Table 7 summarizes our main findings using standardized test scores in reading and mathematics as outcome variables.

We find that the point estimates of the upper and lower bounds point to positive treatment effects, but sample sizes are too small to provide precise estimates. While few people would advocate the use of the simple IV estimator in the presence of selective attrition, it is useful to compare the results of our bounds analysis with the IV approach. One surprising finding is that the simple IV estimates suggest statistically significant positive treatment effects. Our bounds analysis reveal that this inference is not correct.

[25] Moreover we find some evidence that lower performing students are more likely to drop out of the sample, perhaps because they drop out of school.

Table 7: Empirical Results: Achievement

	Reading			Mathematics		
	Upper Bound ATE_{S_m}	Lower Bound ATE_{S_m}	IV ATE_{S_m}	Upper Bound ATE_{S_m}	Lower Bound ATE_{S_m}	IV ATE_{S_m}
MS	66.25	3.68	139.71	180.89	91.08	138.56
	(118.30)	(172.69)	(77.33)	(124.89)	(183.69)	(63.63)
HS	77.05	-25.09	81.97	87.09	-24.22	94.30
	(64.79)	(136.24)	(47.17)	(57.62)	(148.00)	(40.70)

Estimated standard errors are reported in parentheses.

We next turn our attention to behavioral outcomes measured one year after the lotteries were conducted.[26] The main advantage of studying these outcomes is that we do not face the data limitations that we encounter with test scores. Comprehensive records of four important behavioral measures are available: suspensions, offenses, absences, and tardies.

Table 8 summarizes our main findings. Note that a negative treatment effect is a reduction in undesirable behavior and thus a good outcome. For elementary students, we find that magnet programs significantly reduce offenses and suspensions. There are no measurable effects on tardies and absences. We find that there are few significant treatment effects at the middle school level. The estimates themselves suggest that middle school magnet programs have a negative effect on offenses, no effect on suspensions, and possibly an increase in absences and tardies. Again, however, these estimates at the middle school level are generally not significant. For the high school sample, we find strong evidence that the magnet schools reduce absences and tardies while having no significant effects on offenses or suspensions. Comparing the IV estimates with the bounds, we find that the IV estimates are often of similar magnitude to our upper bound estimates and have smaller estimated standard errors than the bound estimates.

[26] Previously Cullen, Jacob, and Levitt (2006) and Imberman (2010) have studied behavioral outcomes when examining school choice programs.

Table 8: Empirical Results: Behavioral Outcomes

	Offenses			Suspensions		
	Upper Bound ATE_{S_m}	Lower Bound ATE_{S_m}	IV ATE_{S_m}	Upper Bound ATE_{S_m}	Lower Bound ATE_{S_m}	IV ATE_{S_m}
ES	-0.28	-0.26	-0.26	-0.49	-0.45	-0.47
	(0.09)	(0.09)	(0.09)	(0.15)	(0.15)	(0.14)
MS	-0.62	-0.48	-0.66	-0.22	0.00	-0.56
	(0.36)	(0.36)	(0.35)	(1.17)	(1.18)	(0.77)
HS	-0.03	0.28	0.20	-0.47	0.14	0.03
	(0.34)	(0.39)	(0.31)	(0.87)	(0.93)	(0.75)
	Absences			Tardies		
	Upper Bound ATE_{S_m}	Lower Bound ATE_{S_m}	IV ATE_{S_m}	Upper Bound ATE_{S_m}	Lower Bound ATE_{S_m}	IV ATE_{S_m}
ES	-2.26	0.98	-1.70	-0.95	0.52	-0.98
	(0.90)	(1.24)	(0.77)	(0.73)	(0.87)	(0.59)
MS	1.98	4.16	1.82	3.04	4.97	2.32
	(1.60)	(2.02)	(1.36)	(1.82)	(2.07)	(2.07)
HS	-8.64	-5.35	-7.77	-7.90	-6.61	-9.41
	(3.32)	(3.60)	(2.55)	(2.78)	(2.87)	(2.45)

Estimated standard errors are reported in parentheses.

We thus conclude that our bounds analysis is informative and demonstrates that magnet programs offered by the district improve behavioral outcomes. In particular, we find that offenses are signicantly lower for elementary school students, while high school students have significantly better attendance and timeliness records. It is also important to note that the 95th percentile of all the behavioral outcomes is zero. Thus our lower bound estimates for all behavioral outcomes is the most pessimistic possible, since it attributes flawless behavior to all who leave the district.

4.3 Comparison with the Lee Estimator

The main alternative to our estimator is the one proposed by Lee (2009) that relies on trimming to construct an estimator for the lower and upper bounds of the treatment effect. It is, therefore, useful to compare both approaches using the data from our application. Table 9 compares our estimates with those obtained from Lee's trimming method.[27] As we detail in the appendix, weighting is appropriate when estimating bounds using data from multiple lotteries. In implementing Lee's estimator, we do not weight lotteries by number of applicants.[28] Hence, the comparison in Table 9 reflects both a difference in the approach to bounding as well as a difference in weighting, potentially confounding the two effects. For the outcomes considered in Table 9, we have confirmed that the results from our weighted estimator are similar to those when we do not weight by lotteries. This is not always the case, however. For example, for MS reading, weighting by lotteries proves to be quite important.[29] Hence, it would be desirable in future work to extend the Lee estimator to weight lotteries. The two methods could then be compared on a common footing in

[27] The results are similar for other outcomes analyzed in this paper. The four outcomes were chosen for the following reason. We have a large sample for elementary school offenses. Our point estimates suggest that the magnet schools may reduce offenses. For tardies, our estimates suggest no effect. The sample size for high school math is small and our estimates suggest no significant treatment effect. Finally, the sample for middle school math is also small, but our estimates suggest that there may be a positive treatment effect.

[28] Lee's estimator has not yet been extended to estimate bounds when combining data from multiple lotteries, though it is surely possible to do so.

[29] Details are available on request.

applications with multiple lotteries.

Table 9 suggests that the empirical results are similar, but there is at least one noteworthy difference. We find that our estimator provides tighter bounds estimates for the magnet treatment effects than the one proposed in Lee (2009) in this application. Table 9 also reports the trimming proportions $\hat{p}$ for Lee's estimator for all outcomes. Note that $\hat{p}$ is the trimming proportion and is defined just as in Lee's paper. The TE CI is the treatment effect confidence interval.

We find that the trimming rates are much greater in our application than in Lee's application, where $\hat{p} = 0.068$. This is due to the fact that our proportion of non-missing data between the control and treatment groups differs significantly since we never observe outcomes for those who leave the district. These students are exclusively contained in the control group since nobody can be in a magnet program yet outside of the district. The other main difference between our application and Lee's application is sample size. Lee reports over 3000 observations in the treatment group before and after trimming. These sample are much larger than the ones in our application. Trimming can, therefore, lead to small sample estimation problems in some applications.

5 Conclusions

We have considered a research design that arises when randomization is used to determine access to oversubscribed programs offered by public school systems. We have developed a new empirical method which deals with selective attrition. Our approach classifies potential participants as stayers, always-takers, leavers, and those that are at risk. We show that the last type of households causes the selective attrition problem. These "at risk" households are also most interesting from a policy perspective since the decision to remain in public schooling crucially depends on the outcome of the lottery. If selective attrition matters, point identification of local average treatment effects for compliers cannot be established. Instead we show how to construct and estimate informative bounds.

Table 9: Comparison with Lee Estimator

	Our Estimator	Lee's Estimator
ES Offenses	UB : -0.28 (0.09) LB : -0.26 (0.09) Point Estimate Range : 0.02 Simple TE CI : [-0.46 , -0.08]	UB : -0.33 (0.08) [347] LB : -0.27 (0.08) [357] Point Estimate Range : 0.06 Simple TE CI : [-0.49 , -0.11] $\hat{p} = 0.337$
ES Tardies	UB : -0.95 (0.73) LB : 0.52 (0.87) Point Estimate Range : 1.47 Simple TE CI : [-2.38 , 2.23]	UB : -3.58 (0.58) [217] LB : -0.68 (0.74) [306] Point Estimate Range : 2.90 Simple TE CI : [-4.72 , 0.77] $\hat{p} = 0.362$
HS Math	UB : 87.09 (57.62) LB : -24.22 (148.00) Point Estimate Range : 111.31 Simple TE CI : [-314.30 , 200.03]	UB : 243.69 (301.97) [33] LB : -150.83 (252.33) [33] Point Estimate Range : 394.52 Simple TE CI : [-645.40 , 835.55] $\hat{p} = 0.660$
MS Math	UB : 180.89 (124.89) LB : 91.08 (183.69) Point Estimate Range : 89.81 Simple TE CI : [-268.95 , 425.67]	UB : 382.86 (286.89) [45] LB : 65.11 (243.81) [48] Point Estimate Range : 317.75 Simple TE CI : [-412.76 , 945.16] $\hat{p} = 0.426$

We have applied our new methods to study the effectiveness of magnet programs. Our empirical results suggest that selective attrition cannot be ignored in our application. We find that magnet programs are useful tools that help the district to attract and retain students from middle class backgrounds. Finally, we have also studied the impact of magnet programs on achievement and a variety of behavioral outcomes. Our findings for achievement effects are mixed. While the point estimates of the bounds point to positive treatment effects, sample sizes are too small to provide precise estimates. For a variety of behavioral outcomes, we do not face these data limitations. Our evidence suggests that magnet programs often improve behavioral outcomes.

We believe that the techniques discussed in this paper can be extended and applied to variety of different problems. Chan and Hamilton (2006), for example, consider clinical AIDS trials and show that attrition is prevalent. Dinardo, McCrary, and Sanbonmatsu (2006) show that attrition is also a problem in the Moving To Opportunity randomized experiment. The techniques developed in this paper can be applied to study these types of questions as well.

Essay 2

The Efficacy of a Pre-Algebra Cognitive Tutor in Chile and Mexico

1 Introduction

1.1 Implementing the Math Cognitive Tutor in the Classroom

This paper examines the impact of an educational intervention on the mathematics performance of 7th grade students in public middle schools in Chile and Mexico. A math cognitive tutor (MCT) targeted at the pre-algebra level represents the intervention. Of the total time devoted to mathematics, students are expected to spend around $\frac{3}{5}$ of it in the classroom and $\frac{2}{5}$ in the computer lab using the MCT software.

There are significant pedagogical changes that are strongly suggested for teachers to use in the classroom portion of the overall curriculum. For the planning of the traditional classes, the teacher can utilize performance reports provided by the MCT system that are individualized to each student. A more collaborative "group learning" strategy, where the teacher serves more as a mentor or coach while students practice problems (instead of a pure lecturer), is also suggested for those classes. Teachers required training before beginning their use of the MCT curriculum. Previous work (Casas, Goodman, & Pelaez (2011) and Casas, Imbrogno, & Vergara (2013)) has highlighted these classroom changes and training provided to teachers. This paper does not include much further discussion of the classroom changes, but instead focuses on the computer-based MCT instruction and its impact on the 7th grade students.

1.2 The ACT-R Theory Behind the MCT

The MCT is based on Anderson's (1993a, b) cognitive theory called adaptive control of thought-rational, or ACT-R. Cognition is modeled as a system of piecemeal knowledge components. According to the theory, the link between declarative, or factual, knowledge and procedural knowledge (problem-solving skill) is strengthened as a power function of practice. Repeated attempts to solve a problem through the use of a particular skill allows students to perform that skill both more quickly and more accurately. In other words, practice improves the performance measure of time spent or errors made by reducing them

over repeated attempts. The models in Nowell & Rosenbloom (1981) and Anderson & Schooler (1991) showing this relationship mathematically can be shown in the simplest form using a performance measure P, learning rate b (less than one), and number of attempts N as:

$$P = b^N \tag{1}$$

As the number of attempts increases, the performance measure (amount of time to solve a problem, probability of making an error) decreases. A graphical representation of this relationship is often referred to as a "learning curve."

Using the MCT software, students can demonstrate proficiency in many knowledge components separately. The MCT combines student actions and a generalized power function to estimate how well the student understands each knowledge component. It uses this estimation to build individualized instruction that focuses on the specific components with which each student struggles. The MCT presents different problems to different students as they progress through the software because its choice of problems to present to each individual is determined by its interpretation of which knowledge components the student has and has not learned. The MCT tailors instruction to the demonstrated ability level of the student and selects problems designed to increase student learning in areas of weakness until a level of mastery is shown. Students gradually build up their more complex problem-solving skills by separate acquisition of a number of these smaller building blocks.

Using the ACT-R theory of individual knowledge components, the MCT is able to break down student misunderstandings at a finer grain level than even individual problems. The MCT itself tracks the knowledge components as separate skills,[30] and an example of the component skills in a given problem should suffice to demonstrate the effectiveness of the ACT-R approach in the MCT. Say a student is asked to identify the greatest common factor (GCF) of 27 and 18. A problem in the MCT proceeds in separate steps. In the first step of

[30] A knowledge component and a skill are the same thing. "Skill" is the term used by the MCT itself in its presentation to students.

the problem, he is asked to list separately the factors of 27 and 18 (skill 1). In the second step, he is asked to identify the common factors of 27 and 18 by appealing to his previous lists (skill 2). Finally, he is asked to identify the greatest common factor of 27 and 18 by referring to his last step (skill 3). See Table 1 for a visual on the problem steps. The three separate skills require a student to (1) factor numbers, (2) choose common numbers between sets, and (3) choose the greatest number from a set. If a student makes a mistake or asks for a hint in the first step, at the conclusion of the problem the MCT is unlikely to produce an immediately subsequent problem asking him to identify the GCF of two numbers again. Instead, it will track back and give the student a problem that focuses on the missing skill (skill 1) either explicitly, as in "List the factors of 36" or in another domain, such as "Please reduce the fraction $\frac{12}{48}$," where the student must begin the problem by demonstrating the same skill 1 as in the GCF problem.

Table 1: Demonstration of MCT Skills Breakdown

What is the greatest common factor of 27 and 18?

Step Description	**Skill**	**Skill Description**
List factors of 27	1	Factor a number
List factors of 18	1	Factor a number
Identify common factor of 27 and 18	2	Choose common numbers between sets
Identify greatest common factor of 27 and 18	3	Choose the greatest number from a set

1.3 Bridge to Algebra MCT

The main technological system in this study is the Bridge to Algebra MCT produced by Carnegie Learning, Inc. It covers math material commonly referred to in the U.S. as pre-algebra (such as number sense and algebraic thinking, fractions, decimals, linear functions, and number systems). Koedinger & Anderson (1993), Koedinger, Anderson, Hadley, & Mark (1997), Anderson & Schunn (2000), and Anderson (2002) provide descriptions of

the the application of ACT-R to the development of the software itself, plus early implementation and design issues. The MCT provides each student a personalized learning environment. A problem generator provides each student with a different set of problems for each skill module. When presented with a problem, the student can ask for hints during all the problem solving processes. Problems are presented in order of complexity. The system keeps track of the number of mistakes and hints used over time. When a skill is completed, the student receives feedback and moves to the next module. The principles underlying the tutor include individualized instruction, opportunities for practice, a scaffolding and hint system that focuses attention on appropriate processes for problem solving, an extensive feedback system that facilitates learning for the student, and an extensive data system that permits diagnosis of student problem solving processes. These principles are designed to enhance learning.

Students do not have to finish the entire software curriculum, and, in fact, very few of them actually do. The prepared Bridge to Algebra curriculum consists of 14 units, 57 sections, and 552 skills.[31] Students progress through the curriculum by mastering skills and sections. According to the MCT developers, these are the best measures of the student learning that has occurred via the software part of the curriculum.

Each section in the MCT contains many skills, and the student must pass all of them to pass a section. The student can advance to the next section without passing all skills once he spends sufficient time and effort on the section, as measured by the number of problems attempted. But the student would still "fail" the section even though he moves on. In order to "master" a section, the student must master all skills within that section. By design then, it is more difficult to master a section than a skill, even leaving aside the fact that there is more material involved. Mastering 9 out of 10 skills in each of four sections, plus 10 of 10 in a fifth section, would result in a skills mastered percentage of 92% (46 skills mastered out of 50), but a sections mastered percentage of just 20% (1 section mastered out of 5). Also by design, the number of "skills mastered" is much greater than the number of "sections

[31] The entire Bridge to Algebra curriculum offered by Carnegie Learning is longer. A subset was used in this study.

mastered," and the two measures are highly correlated. Because the MCT curriculum is personally adaptive for each student, the number of questions required to master a skill varies. If a student correctly answers the steps in a few problems involving a specific skill, he is adjudged to have mastered that skill. But errors and hints, again related to a specific skill, often result in more questions composed of that skill being asked of a student. Therefore, skills and sections mastered are better indicators of demonstrated understanding than simply time spent or problems faced, or even problems answered correctly. Skills mastered and sections mastered are part of the output of the MCT for the use of the teachers and students that reflect the underlying estimation of the student-specific learning curves.

1.4 Related Efficacy Work

The main rationale for this research is the following: in most countries in the world, economic development is based on an educated population, well versed in math, science, and other similar disciplines (National Research Council, 2007; Hanushek & Woessmann, 2008). An educational system that creates a supply of people well versed in mathematics and science is likely to be in a better position to improve the country's economic development. In countries where literacy, language, and mathematics understanding are low, the opportunity for economic development will be low. The goal of this study is to demonstrate the success, or lack of success, of a classroom MCT intervention. If successful, broader diffusion of this MCT could be expected.

Second, this research examines the generalizability of the MCT system. The existing assessments of this specific MCT occur in English-speaking settings. One should note that the issue here is not simply translating from English to Spanish. There are additional changes in the problem content and reprogramming to fit the local contexts. For example, word problems related to "starting a lemonade stand must be adapted for international student understanding. This study takes place in public schools in Chile and Mexico. Changes were made in the language, problems, and programming to adopt the MCT to

Spanish-speaking environments. There are no systematic studies of this MCT in a Central or South American environment, though other researchers (such as Banerjee, Cole, Duflo, & Linden, 2007) have shown positive treatment effects from similar computer-assisted math learning programs in international settings.

The third purpose is to shed more light on the processes that produce successful, or unsuccessful, MCT interventions. As suggested in Cook (2003), we seek to peer into the "black box" and describe some of the implementation quality and measurements of intervening processes that lead to effective use of the MCT. Though our measurements are different, we follow a similar thought process as Pane, McCaffrey, Steele, Ikemoto, & Slaughter (2010) in this regard. This MCT technology has been adopted in large school districts in the U.S.[32] for many middle and high school math courses. Papers on the student performance results of this technology enhanced learning system have found mixed evidence. Some studies on MCT (Koedinger & Anderson, 1993; Koedinger et al., 1997; Morgan & Ritter, 2002; Ritter, Anderson, Koedinger, & Corbett, 2007; Arroyo, Woolf, Royer, Tai, & English, 2010; Ritter, 2011) have shown positive treatment effects while others (Dynarski, 2007; Cabalo, Jaciw, & Vu, 2007; Campuzano, 2009; Pane et al., 2010) have shown insignificant or even negative treatment effects. We add to the program evaluation literature by focusing on implementation quality in addition to the more straightforward question of overall effectiveness in an effort to understand why some studies report positive treatment effects and others disagree. If the difference is a matter of proper implementation, we hope to improve upon the process of bringing the MCT into classrooms.

1.5 Country Settings

In the past two decades, several technology driven initiatives to improve education in South American countries have been introduced with varying degrees of success (de Ferranti et al., 2003; Scheurmann & Pedro, 2009; Chong, 2011). In general, the focus of these initiatives has been two-fold: providing basic technology infrastructure in the schools, plus computer

[32]Including Los Angeles, Chicago, St. Louis, Miami, Baltimore, and Pittsburgh

literacy (for teachers and students) through training. Even though there have been clear advancements in the access to information and to electronic educational materials, the promises of these types of interventions to achieve improved learning have not yet been fully accomplished. Performance in national and international tests such as Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMMS) in the region has not increased as expected, and in some cases has even deteriorated (Hanushek & Woessmann, 2009).

Chile was ranked 96th out of 133 countries regarding quality of primary education in a 2009-2010 report by the World Economic Forum (Schwab, 2009). Even though international 2009 PISA tests reported large gains in math and language education in Chile with respect to the 2006 tests, Chile is still far below the average of the Organization for Economic Co-Operation and Development (OECD) countries. From the 2009-2010 World Economic Forum report:

> The main area requiring improvement for Chile going forward remains the unsatisfactory quality of its educational system, notwithstanding increasing investment in education and rising educational attainment rates. Despite a slight improvement in both cases, primary and higher education continue to be assessed fairly poorly at 96th and 45th ranks, respectively, pointing to the need for further upgrading if Chile is to catch up with best practice countries and establish an innovation-conducive environment.

According to the Chilean Ministry of Education,[33] there is a dramatic gap in quality of education between public and private schools (K-12) in Chile. This situation has produced large inequality and critical social unrest in Chile for the past few years. With very few exceptions, the best results on the national Chilean tests (in English, The System for Measuring the Quality of Education; Spanish acronym SIMCE) are achieved exclusively in the private schools, even though they represent just 8% of the K-12 educational system. Many of the public school math teachers have little or no formal training in math. This MCT

[33]Information pulled from the Ministry's website at http://www.mineduc.cl/

implementation was done exclusively in public schools.

The situation is much the same in Mexico. Although their rank of 65th out of 133 in health and primary education in the same World Economic Forum report outstrips Chile, there is still much progress to be made, especially in math and science in the public schools. From the report:

> Last but not least, the higher education and training system (74th) does not seem to provide the economy with the necessary pool of skilled labor, notably scientists and engineers (94th), and is not creating an environment conducive to adopting new technologies (71st in the technological readiness pillar) and generating new ones (78th in the innovation pillar). Further action is needed to liberalize markets, upgrade the educational system, and improve public governance in the country.

2 Data

2.1 Test Scores and Demographics

The first international implementation for the adapted Bridge to Algebra MCT occurred in public schools of one selected district each in Santiago, Chile, and Mexico City, Mexico, in 2011. Within the selected Chilean district, all 24 schools were invited to participate. Of those, 15 expressed initial interest in participating in the study and, after further consultation about the necessary training and pedagogical changes that would be required of teachers from schools that ended up in the treatment group, 12 remained in our study sample. The final design randomized six schools into treatment and six into control.[34] In

[34]There was also an all-boys school in the Santiago district which is significantly better than the rest of the district schools in terms of socioeconomic statuses and achievement levels of students. That school participated in the study in a different way. Due to concerns that they did not have enough computers to allow appropriate access for all students, the school split its student population into treatment and control groups. Because it was so different from the other district schools, it is left out of the analysis that follows.

total, there are 310 Chilean students in six treatment schools and 358 in six control schools that are used in the results for this paper. All of the students are in 7th grade. About half the schools had a single classroom for all students, and all schools with multiple classrooms had only one teacher responsible for covering them all.

In Mexico, the randomization took place at the classrooml level in four public schools of a district near Mexico City. Each school had the same teacher teaching four different sections of math classes. Within each school, one of the four sections was randomly assigned to use the MCT while the other three served as control groups. The choice of one treatment and three control classes per school stemmed from computer access limitations. In total, there are 156 Mexican students in treatment classrooms and 478 in control classrooms. Again, the students were 7th graders.

The student achievement outcome measure used in this study comes from two comprehensive, grade-level pre-algebra exams given to all students. In both countries, the exams used were outside copies of the national standardized exams from Chile (SIMCE, as referenced earlier). One exam was given in May, near the beginning of the school year and before the MCT implementation. The other was given six months later. Both exams consisted of 44 multiple choice questions. The tests were reviewed and approved by both the Chilean education authorities and the MCT developers prior to their use in the study. The developers agreed that the material in the exams was both grade-level appropriate and covered by the software. The math material focuses on pre-algebra concepts, as does the software used by the treated students.

Unfortunately, there is little student characteristic data from both countries. The only demographic variable is gender, and student pretest scores are the only measure of prior achievement. Gender is binary, and we assign males = 1 and females = 0. Tables 2 and 3 show the average and (standard deviation) of the pretest score and percent male by classroom in both Chile and Mexico. Classroom 4 in school 4 in Mexico consisted exclusively of female students. Table 4 aggregates the same data by type within each country, and

The thrust of the empirical findings does not change when the school is included. If anything, the treatment effect is shown to be stronger when the high quality school is included.

displays it similarly.

Achievement gaps are an important consideration for many policymakers and school administrators. We already briefly discussed the differences in Chile and Mexico between public and private schools. New programs and curricula that prove useful only to select students, especially those at the top of the achievement scale, are not as likely to experience widespread adoption because of concerns over widening the achievement gap and leaving slower students behind. In order to investigate the differential treatment effects of the MCT curriculum across the ability distribution, students were separated, by school, into three separate tertiles[35] depending on their pretest scores: high, middle, or low. Next, the students in each tertile were aggregated across schools. We set up the tertile designations separately by school because there are clear differences in initial achievement levels across schools. We are interested in separating out the differential treatment effects on more versus less skilled students. By design, this must consider the tertiles within the same school or else school differences in pretest scores would also be picked up at the stage of tertile assignment; schools with a high average pretest score would have more representation in the high tertile sample, for example. Due to the discrete nature of the exam scores and the fact that border scores were assigned to the lower tertile, the total observations did not end up equal across the tertiles. In Chile (Mexico), the lowest tertile has 254 (236) total observations across treatment and control students, the middle tertile has 220 (202), and the high tertile has just 194 (196) observations.

2.2 MCT Software

The MCT software also logs each student's usage of and progress through the software-based part of the curriculum. The data logs store information regarding the total software usage time, average hints and errors per problem, as well as the total units, sections, skills, and problems encountered. The software also stores the percentage of the skills and sections

[35]Other splits were examined. The results are largely the same for quartiles and quintiles. The tertile divisions were ultimately chosen to present here because they adequately capture the different ability levels while maintaining sufficient sample sizes across each school-tertile designation.

Table 2: Student Observations in Chile

School Num	Type	Classroom Num	Obs	Pretest	Gender
1	Treatment	1	30	18.8 (3.5)	0.37 (0.49)
		2	31	21.3 (4.9)	0.55 (0.51)
		3	32	18.7 (4.3)	0.59 (0.50)
2	Treatment	1	25	21.8 (5.0)	0.64 (0.49)
		2	19	17.5 (4.93)	0.42 (0.51)
3	Treatment	1	21	19.3 (5.5)	0.67 (0.48)
4	Treatment	1	34	22.8 (6.3)	0.50 (0.51)
		2	31	24.0 (6.3)	0.52 (0.51)
5	Treatment	1	21	15.6 (3.9)	0.29 (0.46)
		2	21	17.6 (3.9)	0.81 (0.40)
		3	21	16.6 (4.6)	0.52 (0.51)
6	Treatment	1	24	21.3 (4.6)	0.63 (0.49)
7	Control	1	28	17.2 (4.3)	0.46 (0.51)
		2	29	18.0 (4.2)	0.52 (0.51)
		3	23	17.1 (5.1)	0.30 (0.47)
8	Control	1	25	20.9 (5.6)	0.64 (0.49)
		2	23	17.7 (4.3)	0.48 (0.51)
		3	25	19.0 (6.3)	0.72 (0.46)
9	Control	1	32	24.3 (5.5)	0.53 (0.51)
		2	31	24.6 (5.7)	0.45 (0.51)
		3	29	23.3 (5.8)	0.59 (0.50)
10	Control	1	18	18.0 (5.5)	0.61 (0.50)
		2	21	20.8 (6.2)	0.67 (0.48)
11	Control	1	23	15.5 (5.2)	0.57 (0.51)
		2	24	18.4 (3.5)	0.42 (0.50)
12	Control	1	27	18.6 (4.3)	0.37 (0.49)

Table 3: Student Observations in Mexico

School Num	**Type**	**Classroom Num**	**Obs**	**Pretest**	**Gender**
1	Treatment	1	40	23.3 (5.9)	0.45 (0.50)
	Control	2	41	21.3 (6.0)	0.59 (0.49)
	Control	3	37	21.7 (5.6)	0.65 (0.48)
	Control	4	38	20.8 (5.8)	0.68 (0.47)
2	Treatment	1	41	25.0 (6.0)	0.41 (0.50)
	Control	2	44	25.3 (5.4)	0.36 (0.49)
	Control	3	39	26.4 (6.9)	0.46 (0.51)
	Control	4	40	24.4 (6.1)	0.35 (0.48)
3	Treatment	1	36	23.6 (5.1)	0.50 (0.51)
	Control	2	41	22.0 (5.2)	0.42 (0.50)
	Control	3	39	24.9 (6.0)	0.46 (0.51)
	Control	4	39	21.5 (5.1)	0.46 (0.51)
4	Treatment	1	39	18.1 (6.9)	0.74 (0.44)
	Control	2	44	20.5 (6.0)	0.82 (0.39)
	Control	3	37	20.7 (6.4)	0.78 (0.42)
	Control	4	39	17.8 (5.6)	0 (N/A)

Table 4: Characteristics by Type - Aggregated

Country		**Treatment**	**Control**
Chile	*Observations*	310	358
	Pretest	19.9 (5.5)	19.8 (5.8)
	Gender	0.54 (0.50)	0.52 (0.50)
Mexico	*Observations*	156	478
	Pretest	22.5 (6.5)	22.3 (6.3)
	Gender	0.53 (0.50)	0.50 (0.50)

seen by the student that were mastered. Obviously this data only exists for the treated students since the control group did not use the software. Table 5 shows the average and (standard deviation) of the units completed, total usage hours, skills mastered, and sections mastered for each treatment classroom.

Table 5: MCT Process Data

Country	**School, Class**	**Usage Hrs**	**Units**	**Sections Mas**	**Skills Mas**
Chile	1, 1	17.8 (2.8)	11.1 (2.9)	31.4 (12.0)	360.5 (141.1)
	1, 2	17.9 (3.6)	12.5 (2.1)	36.8 (10.1)	428.8 (104.3)
	1, 3	17.8 (3.0)	9.2 (2.8)	25.0 (10.9)	275.4 (126.3)
	2, 1	8.5 (2.4)	5.9 (3.5)	18.8 (11.6)	174.5 (130.0)
	2, 2	9.4 (2.8)	5.3 (3.6)	16.8 (11.2)	154.1 (125.6)
	3, 1	10.3 (1.8)	8.5 (3.5)	26.1 (13.4)	259.2 (153.0)
	4, 1	8.7 (2.4)	6.6 (2.8)	22.1 (10.8)	196.0 (114.4)
	4, 2	8.7 (2.6)	6.8 (3.9)	22.8 (13.7)	207.3 (152.7)
	5, 1	13.2 (3.5)	7.2 (2.9)	17.5 (7.5)	195.5 (99.8)
	5, 2	11.5 (2.2)	6.6 (2.8)	19.6 (9.3)	189.2 (106.9)
	5, 3	7.9 (2.2)	4.9 (2.9)	15.0 (10.0)	133.8 (102.2)
	6, 1	18.6 (2.8)	11.1 (2.9)	31.3 (12.5)	357.6 (143.5)
	All students	12.8 (5.0)	8.2 (3.9)	24.3 (12.8)	252.3 (154.2)
Mexico	1, 1	7.8 (1.3)	9.0 (2.4)	27.5 (10.1)	249.5 (115.0)
	2, 1	5.8 (1.9)	6.3 (2.3)	18.2 (7.5)	136.1 (70.1)
	3, 1	14.3 (3.4)	11.3 (1.6)	34.0 (8.4)	351.8 (103.5)
	4, 1	1.3 (0.77)	2.4 (1.0)	6.2 (3.8)	37.6 (24.5)
	All students	7.2 (5.0)	7.1 (3.8)	21.2 (12.9)	190.3 (144.1)

Whether measured by units completed or total hours, the students in Chilean schools 1 and 6 and Mexican school 4 spent the most time, on average, using the software. There are obviously differences across the classrooms and schools in the amount of time their students

spent using the software. The MCT curriculum calls for 40% of class time to be devoted to the software. In practice, this would constitute two days of a typical school week, which over a six month time period would mean roughly 25 hours spent in the computer lab in lieu of traditional math lectures. Note that this standard for total MCT software usage hours by students was not met, on average, by any classroom in any school. Students were also encouraged to use the software during their free time and even after-school when the lab was available. Conversations with the school administrators, which will be more fully explained and evaluated later, confirmed that the twice per week standard for computer use was rarely met. Students often only visited the lab once each week during class time, and very few of them accessed the software in their own time. Some of the schools did not have enough computers for every student, further hindering their personalized time with the software when they were forced to share the machines. Units, sections mastered, and skills mastered are highly correlated measures[36] across students (less so for usage hours).

2.3 Schools

The differences between schools are another component of the analysis in this paper. Our research team determined values for the treatment schools in four specific areas of interest (basic inputs, infrastructure, implementation, and learning environment) in an attempt to quantify the differences between them. It is valuable to examine how these school-level variables might affect student software usage. The basic inputs variable is an attempt to quantify all of the following: quality of students and teachers, budget issues, socio-economic status of students, absenteeism rates, and discipline rates in the schools. The infrastructure measure was probably the most important one in this context, and likewise the one where with the lowest average score. Infrastructure captures internet connection quality, presence of tech support, and the computer to student ratio. Each treatment school was expected to have at least one computer per student in the computer labs, but that was in fact not the case in some of the schools. Schools which did not provide enough computers were obviously at a severe disadvantage concerning student access to the software part

[36] Smallest pairwise correlation between any combination is 0.93.

of the MCT curriculum. Students were forced to rotate and take turns in the computer lab, restricting the hours of access and amount of material covered in the math program. Connectivity to the internet was also necessary to run the software. Implementation is a measure of teacher and principal enthusiasm, motivation, and commitment toward the new curriculum. Training for instructors is a necessary part of the MCT implementation, and training session attendance and participation by teachers was incorporated into the rating. Finally, learning environment is a measure of the frequency of MCT use, peer cooperation among students in both the classroom and lab, and student motivation and behavior in the lab.

Each rating was made by three different, affiliated members of the research team based on conversations, notes, and feedback from the Chilean project manager and the teacherss and principals in the schools. Raters were blind to the school itself in the write-ups. Each of the four criteria was evaluated on a 1-5 scale, with 5 being the best rating. There was a high convergence among raters.[37] Table 6 shows the ratings given to each school for each variable. It also includes a total rating which is simply the sum of the four component pieces. School 2 in Chile and school 4 in Mexico rate particularly poorly on these measures. School 3 in Mexico is the only school with a "perfect" rating.

2.4 Surveys

We administered surveys for nearly all students in the treatment schools.[38] The students were asked questions on a range of topics related to mathematics, computers, and the MCT software. Their survey responses were anonymous and could only be linked to students at the school level. The surveys requested that students indicate their perceptions in the

[37] All three researchers rated all ten treatment schools (six in Chile, four in Mexico) on the four characteristics using the 5-point scale. On 31 of the 40 school-characteristic ratings, all three researchers assigned the exact same score. The remaining 9 school-characteristic ratings which showed differences amongst the raters never differed by more than one point by any pairwise set of raters and were resolved through discussion and consensus in order to assign one value for them. Fleiss' κ of 0.79 reveals substantial agreement among raters.

[38] Mexican school 4 has no available survey data.

Table 6: School Characteristics

Country	**School**	**Basic Inputs**	**Infra**	**Imple**	**Learn Env't**	**Total**
Chile	1	3	3	5	4	15
	2	2	1	3	2	8
	3	3	4	4	3	14
	4	5	3	5	4	17
	5	3	3	5	5	16
	6	5	4	5	5	19
Mexico	1	5	5	4	5	19
	2	4	1	5	5	15
	3	5	5	5	5	20
	4	1	1	3	3	8

following areas: (1) Ease of use and clarity of the MCT software, (2) Teacher help with the MCT, (3) Computer lab infrastructure, and (4) Effectiveness of the MCT. The anonymous aspect of the surveys reduces concerns about biased reporting by students. The surveys are complementary to the school characteristics discussed in the previous section. The school variables are derived measures from conversations with the researchers and principals involved in the project at the school level, plus overall input and perspective from the Chilean project manager. It is a top down view of the differing school characteristics. In contrast, the survey responses can be thought of as a bottom up view of the schools. They are the beliefs, feelings, and perceptions of the students themselves. The students likely have no idea how their particular school or classroom matches up against other schools in the survey topics, but their input is quite relevant since the curriculum change applies directly to them.

The survey questions associated with each topic area are shown below. All questions use a scale ranging from "very low" to "very high," with the responses agreeing with the survey

question statement (i.e. for the first component of Ease of Use, the "very low" answer said "very unsatisfied").

Surveys were also administered to the teachers in the six Chilean schools which adopted the MCT. The survey questions and responses from teachers are shown in the Appendix.

1. Ease of Use and Clarity of MCT Software
   - Indicate how satisfied you are with the ease of use of the Cognitive Tutor.
   - Indicate how satisfied you are with the clarity of the instructions that the Cognitive Tutor offers.
2. Teacher Help with the MCT
   - Indicate how satisfied you are with the help the teacher gave you to use the Cognitive Tutor.
   - Indicate to what extent your teacher helped you use the Cognitive Tutor in the lab.
   - Indicate to what extent your teacher handled the Cognitive Tutor adequately in the math classes (i.e. he knew the contents, steps, hints, etc)
3. Infrastructure
   - Indicate your level of agreement with the following statement about computers: The computers worked adequately in the lab.
   - Indicate your level of agreement with the following statement about computers: The internet connection worked in the lab.
4. Effectiveness of the MCT
   - Indicate to what extent your math learning improved with the Cognitive Tutor.
   - Indicate your level of agreement with the following statement about the Cognitive Tutor: The Cognitive Tutor is a useful resource to learn math.

- Indicate your level of agreement with the following statement about the Cognitive Tutor: I would like to keep using the Cognitive Tutor in the math classes.
- Indicate your degree of comfort with this new way of learning math.

3 Research Questions

The main objective of this study is to estimate the causal impact of the Bridge to Algebra Cognitive Tutor curriculum on Chilean and Mexican students' pre-algebra achievement. The design includes treatment and control groups with before and after measures of math ability. We also investigate the process variables of the MCT itself to determine whether the amount of use is reflected in the achievement measure. Because the MCT curriculum requires significant changes in the classroom structure, school technological infrastructure, and student and teacher behavior, the study incorporates measures of school characteristics that include computer access, student background, discipline issues, teacher fidelity to training, and classroom fidelity to suggested usage of the MCT. Last, students are surveyed in the schools using the MCT to better understand its perceived ease of use and effectiveness.

Using this framework as a guide, we address the following questions:

1. Does the MCT significantly affect math performance compared with students in a controlled condition?
2. Do the MCT process indicators at the student level affect changes in math performance?
3. Do school characteristics affect the process indicators in the MCT?
4. What are the student attitudes about the MCT experience?

4 Research Model

4.1 Initial Balance Check

It is necessary to check whether the randomization process that determined which schools would implement the MCT curriculum and which would not created balanced treatment and control groups before evaluating results. If the groups are not equal on observable characteristics and pretest scores, it would not only call into the randomization processes employed but also the relevance of the results. Since the randomization was done differently in each country, we evaluate each country separately. The only observable characteristic we have is gender. Balance checks are run on both the proportion of male students and the average pretest score by treatment assignment using two-sided t-tests of equivalence with unequal variance. The comparisons are also done separately by tertile.

4.2 Treatment Effects

We follow the hierarchical linear modeling (HLM) approach advanced by Raudenbush & Bryk (2002) and used in a similar design in Pane, et al (2010) to estimate the causal impact of the Bridge to Algebra MCT on student math performance. We have students nested within classrooms within schools. The treatment group is represented by the students who used the MCT, and the control group by the students who did not. The estimation and results for the two countries are kept separate because the unit of randomization was different in the two settings. In Chile, entire schools were randomized into treatment and control groups; in Mexico, randomization occurred within each school such that one classroom was chosen for treatment and the rest were kept as controls.

Let Y_{ijk} be the posttest score for student i in classroom j in school k. Similarly, let y_{ijk} represent the pretest score and X_{ijk} student covariates.[39] At the student level (level 1), I model the posttest score as a function of the pretest score, covariates, classroom level (level

[39] Commonly, these could include race, gender, free/reduced lunch status, etc. In this study, all I have access to is information on gender.

2) variables (μ_{jk} for now), and an error term $\epsilon_{ijk} \sim N(0, \tau_1^2)$.

$$Y_{ijk} = \mu_{jk} + \beta_1 y_{ijk} + \beta_2 X_{ijk} + \epsilon_{ijk} \tag{2}$$

The classroom level equation incorporates the treatment assignment. The model will follow the experimental design from Mexico. Let type be denoted as T_{jk}. The treatment classes have $T = 1$ and the control classes have $T = 0$. Let $\overline{y_{jk}}$ represent the class average of the pretest score. Though we are agnostic about the nature of the effect of the average class pretest score on individual student performance on the posttest in this paper, common arguments for its inclusion revolve around peer effects in the classroom. For $\eta_{jk} \sim N(0, \tau_2^2)$, μ_{jk} is modeled as:

$$\mu_{jk} = \gamma_0 + \gamma_1 T_{jk} + \gamma_2 \overline{y_{jk}} + \eta_{jk} \tag{3}$$

Combining Equations (2) and (3) yields:

$$Y_{ijk} = \gamma_0 + \gamma_1 T_{jk} + \gamma_2 \overline{y_{jk}} + \beta_1 y_{ijk} + \beta_2 X_{ijk} + \epsilon_{ijk} + \eta_{jk} \tag{4}$$

Looking at Equation (4), the paramater of interest is γ_1. The estimate of γ_1 is the treatment effect, in terms of standardized test scores, for the 7th grade students using the MCT.

In Chile, the randomization occurred at the school level. A similar line of reasoning as shown above helps derive the following empirical model for Chile, where all that changes in the equation itself is the elimination of the j subscripts on $\overline{y}$, T, and η. This obviously also changes the estimation slightly.

$$Y_{ijk} = \gamma_0 + \gamma_1 T_k + \gamma_2 \overline{y_k} + \beta_1 y_{ijk} + \beta_2 X_{ijk} + \epsilon_{ijk} + \eta_k \tag{5}$$

Equations (4) and (5) are estimated across the entire sample within each country, and then separately according to each tertile. We used the *xtmixed* command in Stata to estimate the models. Specifications are run both including and not including the classroom pretest average ($\overline{y_{jk}}$).

4.3 MCT Process Data

The HLM models proposed in the last section are also used, with slight modification, to address the second research question. Here we investigate whether or not accomplishing more of the software itself is predictive of better math test performance. The MCT process variables can be considered "mediator variables" in the model of Baron & Kenny (1986). The control students do not use the MCT software at all, so they are not included in this analysis. Rather, looking at just the treated students, we explore the relationship between software usage (in terms of units completed, total usage hours, sections mastered, and skills mastered) and performance (as measured by posttest scores) by controlling for pretest score, gender, and school/classroom. The approach is to estimate Equation (6) below for Mexico, which is very similar to Equation (4) shown previously.[40] The switch to λ from γ is purely for notational convenience to further distinguish the equations. Here, the type T_{ijk} is dropped and MCT_{ijk} represents the specific MCT data considered. Separate regressions are run for each of the four MCT process variables. The country's are evaluated separately. The paramater of interest is λ_1. The estimate of λ_1 is the amount by which we expect a student's posttest score to increase when he increases the value of the MCT variable under consideration by one.

$$Y_{ijk} = \lambda_0 + \lambda_1 MCT_{ijk} + \beta_1 y_{ijk} + \beta_2 X_{ijk} + \epsilon_{ijk} + \eta_{jk} \tag{6}$$

[40]The change to the estimation in Chile is obvious, and mirrors the change between Equations (4) and (5).

4.4 School Characteristics

We are interested in explaining the relationship between the school characteristics, MCT process variables, and posttest scores. However, we only have ten schools containing students using the MCT. Regression techniques with such a small sample size are not useful. To investigate the relationship between the school characteristics and other data, our approach is quite simplistic. We average the MCT variables and posttest scores for each school, and then check the pairwise correlations between them and each of the four school characteristics (basic inputs, infrastructure, implementation, and learning environment). Carnegie Learning stresses that school administrators and teachers must be on board with this curriculum change for the MCT to be effective. In addition, there must be sufficient computer access available for students. As discussed earlier, our conversations with the on-location research team and individual school officials revealed varying degrees of proper implementation across schools. We want to know whether implementation and infrastructure differences possibly help explain the large differences we see across students and across schools in terms of completion and mastery of the Bridge to Algebra MCT and test scores.

4.5 Surveys

The survey responses are aggregated across classrooms within each school. We set up a combination of questions that closely matches specific areas of interest in the evaluation of the MCT, as explained earlier. Since each interest area (ease of use of tutor, teacher help with tutor, infrastructure, and effectiveness) is composed of multiple responses, the results we report indicate the percentage of people across all questions in an area who answer "very high" or "high." These survey responses are helpful since they capture student perceptions of school infrastructure and, more importantly, the benefits of the MCT software. Surveys were not conducted in the control schools since the questions are focused on the subject of the MCT itself, so there is unfortunately not a comparison group for the results on student attitudes.

5 Results

5.1 Balanced Sample

The results on the balance check are shown in Table 7. The table displays the t-statistic resulting from the two-sided test, with the 95% confidence interval shown in parentheses below it. A negative t-statistic denotes that treatment students, on average, had a higher pretest score or percentage male. Only one of the t-statistics are significantly different from zero. There is a higher percentage of male students in treatment than in control for low tertile group in Mexico. Overall, on both pretest scores and percentage male, the treatment students match the control students. The randomization design effectively split our sample into reliable experimental groups. The treated and control groups only differ by treatment assignment.

5.2 Treatment Effects

The first outcome measure we address is whether or not the MCT is an effectice curriculum for increasing student test scores. Tables 8 and 9 shows the average pre- and posttest scores, plus the difference scores, by school and type in both countries. The difference score is the posttest score minus the pretest score. If a difference score is positive, a student scored better on the posttest than he did on the pretest. As shown by comparing both tables, the average treatment students have greater (positive) difference scores than the average control students. In Chile, every individual treatment school had a positive average difference score but one and every control school had a negative average difference score. These tables support the descriptions of Chilean public primary schools in the World Economic Forum's report - namely, that students are lagging behind and underperforming in math. The average student not using the Cognitive Tutor actually scores *lower* on a similar test after 6 months of pre-algebra instruction. In Mexico, where the randomization was done within

Table 7: Balance Check

Country	**Students**	**Pretest**	**Percent Male**
	All	-0.38	-0.49
		(-1.02, 0.69)	(-0.10, 0.06)
Chile	High Tertile	0.10	1.38
		(-1.22, 1.35)	(-0.04, 0.25)
	Middle Tertile	0.23	-1.91
		(-0.68, 0.86)	(-0.25, 0.00)
	Low Tertile	-1.00	-0.04
		(-1.05, 0.34)	(-0.13, 0.12)
	All	-0.36	-0.51
		(-1.39, 0.96)	(-0.11, 0.07)
Mexico	High Tertile	-0.94	1.01
		(-1.72, 0.61)	(-0.08, 0.25)
	Middle Tertile	-0.99	1.70
		(-1.26, 0.42)	(-0.02, 0.29)
	Low Tertile	0.01	-3.61
		(-1.13, 1.14)	(-0.40, -0.12)

t-statistic shown above

95% confidence interval in parentheses below

each school by classroom, the average difference-in-difference score[41] by school was positive for every school.

Table 8: Test Scores for Treatment Students

Country	**School**	**Pretest**	**Posttest**	**Difference Score**
Chile	1	19.57 (4.40)	19.58 (5.84)	0.01 (5.09)
	2	19.91 (5.37)	21.14 (6.19)	1.23 (5.45)
	3	19.33 (5.46)	22.62 (7.24)	3.29 (5.52)
	4	23.37 (6.29)	24.22 (7.62)	0.85 (5.27)
	5	16.60 (4.16)	16.97 (4.50)	0.37 (4.54)
	6	21.33 (4.56)	21.08 (4.22)	-0.25 (5.08)
	All students	19.93 (5.47)	20.56 (6.51)	0.63 (5.13)
Mexico	1	23.28 (5.93)	28.33 (7.04)	5.05 (4.38)
	2	25.05 (5.97)	27.95 (6.57)	2.90 (4.89)
	3	23.64 (5.13)	22.56 (6.39)	-1.08 (4.76)
	4	18.08 (6.88)	19.10 (7.56)	1.03 (4.36)
	All students	22.53 (6.53)	24.60 (7.87)	2.06 (5.08)

Average (standard deviation)

In order to quantify this difference in performance, we turn to the results of the estimations of Equations (4) and (5), shown in Table 10. Both pre- and posttest scores are transformed into z-scores (separately by country and by pre- and posttest) with mean 0 and standard deviation 1 to make the empirical results more easily interpretable. The coefficient value on *Type* represents the increase in standard deviations of the posttest score that treated students are predicted to gain over control students. In both Chile and Mexico, there a positive and statistically significant treatment effect. Using the top line specification in both countries, where all students are included in the estimation, students who use the

[41] In other words, the average difference score of the treated students less the average difference score of the control students.

Table 9: Test Scores for Control Students

Country	**School**	**Pretest**	**Posttest**	**Difference Score**
Chile	7	17.46 (4.45)	17.28 (4.07)	-0.19 (4.94)
	8	19.23 (5.57)	18.45 (6.63)	-0.78 (5.46)
	9	24.08 (5.63)	23.71 (7.11)	-0.37 (5.11)
	10	19.51 (6.00)	17.87 (6.36)	-1.64 (6.66)
	11	17.00 (4.60)	15.89 (4.13)	-1.11 (4.90)
	12	18.56 (4.30)	18.11 (5.57)	-0.44 (5.59)
	All students	19.77 (5.81)	19.11 (6.49)	-0.65 (5.32)
Mexico	1	21.28 (5.75)	23.69 (7.44)	2.41 (6.07)
	2	25.37 (6.15)	27.26 (6.74)	1.89 (4.14)
	3	22.82 (5.59)	22.15 (6.16)	-0.66 (4.46)
	4	19.68 (6.10)	18.48 (6.74)	-1.20 (5.46)
	All students	22.31 (6.26)	22.92 (7.47)	0.60 (5.30)

Average above, standard deviation in parentheses below.

MCT score nearly $\frac{1}{5}$ of a standard deviation of the posttest score higher on the posttest than their control group peers. This translates to nearly 1.2 extra problems on the 44 question exam at the end of the school year that treatment students answer correctly, even though both groups began the year with equivalent scores on the pretest.

Table 10: Treatment Effects

Country	**Students**	**Class Pretest**	**Student Pretest**	**Gender**	**Type**
	All		0.58 (0.03) ***	0.06 (0.05)	0.18 (0.09) **
Chile	All	0.29 (0.11) ***	0.56 (0.03) ***	0.06 (0.05)	0.17 (0.08) **
	High Tertile		0.88 (0.09) ***	0.05 (0.10)	0.26 (0.17)
	High Tertile	0.35 (0.22)	0.82 (0.11) ***	0.03 (0.10)	0.23 (0.13) *
	Middle Tertile		0.66 (0.10) ***	0.09 (0.07)	0.17 (0.10)
	Middle Tertile	0.24 (0.23)	0.49 (0.18) ***	0.09 (0.07)	0.17 (0.10) *
	Low Tertile		0.55 (0.09) ***	0.09 (0.07)	0.14 (0.10)
	Low Tertile	0.20 (0.15)	0.47 (0.11) ***	0.08 (0.07)	0.14 (0.10)
	All		0.67 (0.03) ***	-0.02 (0.05)	0.19 (0.09) **
Mexico	All	0.25 (0.16)	0.67 (0.03) ***	-0.03 (0.05)	0.18 (0.10) *
	High Tertile		0.72 (0.09) ***	0.10 (0.09)	0.19 (0.14)
	High Tertile	0.32 (0.27)	0.67 (0.10) ***	0.10 (0.09)	0.18 (0.14)
	Middle Tertile		0.84 (0.13) ***	-0.15 (0.09) *	0.34 (0.18) *
	Middle Tertile	-0.01 (0.31)	0.84 (0.17) ***	-0.15 (0.09)	0.34 (0.18) *
	Low Tertile		0.48 (0.10) ***	-0.06 (0.09)	0.09 (0.15)
	Low Tertile	0.62 (0.23) ***	0.40 (0.10) ***	-0.07 (0.09)	0.08 (0.13)

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

Note that the coefficient estimates on $StudentPretest$ are all positive and less than one. This makes sense in our context. Scores between the pretest and posttest should be highly

correlated (smarter students score high on both, weaker students score low on both), and the coefficient being less than one represents regression to the mean on the posttest. An additional answer correct on the pretest would be expected to raise a student's posttest score, but by less than a full correct answer since the extra correct answer on the pretest could be a random guess unrelated to the underlying student ability that the tests are meant to ascertain.

The results by tertile in Table 10 are also illuminating. The number of observations decreases to the point of making the treatment effect in most specifiations statistically insignificant.[42] However, the point estimate on *Type* is positive for every specification. The personized nature of the MCT allows students of all ability levels to improve their math test scores.

5.3 MCT Process Data and Test Scores

As Pane et al. (2010) explain in their paper, it is difficult to disentangle the unobserved effects of individual ability or motivation from the instructional effect of the software through a specification like Equation (6), and hence a significant λ_1 simply helps confirm that the skills required for progression through the software are strongly related to the skills measured by the standardized tests. We generally agree with this sentiment but would stress that a positive finding here lends credence to the argument that the software itself is integral to the new curriculum. At first it would seem obvious that mastering more of the software part of the curriculum would be essential for achieving higher test scores. Unfortunately, upon reflection this relationship is not actually that obvious. Being in the treatment group meant an entire change in math teaching and learning for that school year. The incorporation of computer technology in the math courses is just one aspect of the overall shift in teaching strategy seen in the treatment classrooms, albeit the most obvious one. Students completed more group projects and teachers were encouraged and trained

[42] We reach the 10% level of significance in Chile for the high and middle tertiles (with classroom pretest included) and in Mexico for the middle tertile (both specifications).

to act more as facilitators and coaches than as lecturers in the classroom. The software aspect of the curriculum shift, where students are sent to the computer lab for two out of the five class days each week, is the only part of the MCT treatment program for which we have data. But it is quite conceivable that other aspects of the pedagogical shift caused the treatment effect finding above, and the process of the software is not particularly integral for the treatment students' math achievement. In addition, the exams are constructed by the Chilean education ministry, not the developers of the MCT software. If, for example, the math software heavily frontloads much of the tested material but then covers non-essential topics throughout the remaining units, then students who accomplish more of the software program would likely not see higher exam scores than their peers who got through only the early units. Demonstrating that MCT mastery is aligned with standardized test mastery is a relevant finding for the efficacy of the software on its own.

The z-scores are again used for the test data in these estimations. Table 11 shows the results of Equation (6) and its equivalent specification for Chile for each separate MCT variable. For convenience, only λ_1 is shown in the table, though the rest of the estimates can be obtained from the authors. Each column of Table 11 comes from a different estimation, where only the listed MCT variable (units, total hours, sections mastered, skills mastered) is included as MCT in Equation (6). Since we do not include control students in these regressions, we do not separate students by tertile due to concerns over sample size. The results for both countries are basically the same. The total hours of usage is not significant in predicting posttest scores in Mexico (and marginally significant in the *negative* direction in Chile), but the coefficient estimates on number of units completed, skills mastered, and sections mastered are all positive and significant. This aligns with the contention by Carnegie Learning that the "mastered" variables are predictive of performance. Mastered sections and skills encompass both student effort and student ability, while units completed is reflective of effort put forth on the software portion of the curriculum.

The coefficients on sections and skills mastered in the regressions relate the marginal effect on posttest scores of increasing sections or skills mastered by one unit while controlling for school. In terms of magnitude for the software effects, consider the following. Across the

Table 11: MCT and Test Scores

Country	**MCT Data**			
	Usage Hrs	**Units**	**Skills Mas**	**Sections Mas**
Chile	-0.016 (0.009) *	0.064 (0.012) ***	0.002 (0.0003) ***	0.022 (0.003) ***
Mexico	-0.027 (0.020)	0.080 (0.029) ***	0.002 (0.0006) ***	0.031 (0.007) ***

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

310 students in Chile, the sections mastered (skills mastered, units) variable has an average of 24.3 (252.3, 8.2) and standard deviation of 12.8 (154.2, 3.9). The regression results show that an increase in usage of the MCT software by one standard deviation of sections mastered (skills mastered, units) improves posttest scores by 0.28 (0.31, 0.25) standard deviations. That is a consequential increase in achievement on the posttest. In Mexico, the sections mastered (skills mastered, units) variable has an average of 21.2 (190.3, 7.1) and standard deviation of 12.9 (144.1, 3.8). Our results there show that an increase in usage of the MCT software by one standard deviation of sections mastered (skills mastered, units) improves posttest scores by 0.40 (0.28, 0.30) standard deviations. The *Bridge to Algebra* MCT curriculum, taken as a whole, is effective at improving math scores for students of all abilities. Furthermore, the notion that students can expect to do better on the math exams when they have accomplished more units and mastered more of the sections and skills taught in the software is supported. The processes of the software piece of the MCT curriculum are effective at improving math score outcomes.

5.4 School Characteristics and Software Usage

We are interested in explaining the relationship between the school characteristics and the MCT process variables. Carnegie Learning stresses that school administrators and teachers must be on board with this curriculum change for the MCT to be effective. In addition,

there must be sufficient computer access available for all students. There are large differences across schools in terms of average completion and mastery of the Bridge to Algebra MCT.

Table 12 shows the correlations between each of the four school characteristics (basic inputs, infrastructure, implementation, and learning environment) and the average value for the MCT data (usage hours, units completed, sections mastered, and skills mastered) by school. It is clear from the table that infrastructure and implementation school ratings are highly correlated with student MCT usage. Though this result is based on just ten observations and is therefore not the strongest in this paper, we believe it has practical importance for policy considerations. Every school was expected to have one computer per student and reliable connectivity to the internet, but in reality this was not seen. Those schools which experienced this ideal infrastructure possessed an environment which allowed their students to excel, while those who adopted the MCT curriculum but did not have the ability to properly use it saw their students lag behind. Committed teachers, principals, and administrators (implementation) are central to realizing the effectiveness of the MCT. This is evidenced most strongly by teacher fidelity to pre-rollout training. Students will also master more skills using the MCT when the frequency of lab use and the conditions in the lab (learning environment) are high. School and governments considering the adoption of the MCT curriculum need to fully commit the time, energy, and resources to the endeavor. Simply sending students to an inadequate computer lab to use the software every now and then will be decidedly less effective than consistently utilizing proper facilities. The schools which most closely followed the recommended implementation activities saw their students complete and master more of the MCT software curriculum. In that light, the results presented in this section speak to the possibility that the treatment effects shown earlier are lower bounds of the true treatment effect. If all schools were able to properly take advantage of the MCT curriculum, we have reason to believe that the treatment students as a whole would have experienced even greater posttest scores.

Table 12: Correlations of School Characteristics and Software Usage

	Basic Inputs	**Infrastructure**	**Implementation**	**Learning Env't**
Usage Hrs	0.22	0.55	0.60	0.33
Units	0.54	0.78	0.61	0.49
Skills Mas	0.45	0.76	0.57	0.39
Sections Mas	0.61	0.80	0.58	0.44

5.5 Student Attitudes about MCT

Table 13 presents student responses on the four dimensions measured by multiple questions per dimension. To simplify presentation, we indicate the percentage of people across questions in a dimension who feel very positive or positive. For example, 83% of the students in Chilean school 1 asked questions about the "Ease of Use of Tutor" said they were very satisfied or satisfied on both questions within this dimension.

Our fourth and final research question addressed student attitudes toward the MCT. In general, students report high levels of satisfaction with the new curriculum and its implementation. Most of the students rate the ease of use of the Tutor, teacher help with the Tutor, and effectiveness of the Tutor highly or very highly. The process of the MCT itself seems to lead to positive dispositions toward the technology. Infrastructure gets the lowest ratings, which is not surprising since the schools are primarily in poorer areas. The lower rating of infrastructure by students when compared to other survey response areas also matches the lower values we saw for infrastructure when compared to the other school characteristics.

The attitude results are a complementary outcome to the positive treatment effects. Students feel the MCT helps them learn mathematics, they enjoy using it, and they find the teacher to be supportive. As shown in the Appendix, teachers also had a positive disposition toward the MCT curriculum.

Table 13: Student Surveys

Country	School	Ease of Use	Teach Help	Infrastructure	Effectiveness
Chile	1	83%	92%	22%	90%
	2	71%	83%	66%	90%
	3	87%	86%	82%	88%
	4	83%	80%	55%	87%
	5	89%	94%	60%	89%
	6	90%	95%	84%	93%
Mexico	1	86%	91%	37%	88%
	2	48%	34%	48%	52%
	3	64%	85%	13%	85%

Percentage answering high/very high in aggregate

6 Conclusions and Discussion

The results of this paper add to the growing body of work that investigates technology-based math curricula. It is the first to our knowledge that looks at the MCT curriculum in Central and South American schools. All previous MCT studies have focused on U.S. schools. The results presented in this paper are largely supportive of the MCT *Bridge to Algebra* curriculum for Chilean and Mexican middle school students. Schools which expressed interest in adopting the curriculum were randomly assigned to treatment or control groups.

Though they scored the same on the pretest, treatment students outperformed their control group peers on the standardized exam posttest in both Chile and Mexico. The treatment effect of the MCT curriculum is significant and positive. The overall treatment effect shows that treated students answer an additional 1.2 questions correct on the 44 question final exam than their control group peers.

The finding on treatment effects is not driven by students in just one or two treatment schools performing well, or conversely those in a few control schools severely dragging

down the overall control group. In addition, the positive treatment effect exists across the student ability distribution. Though there was generally insufficient power due to limited observations, every regression specification that separated students by initial score into three groups (high scorers, middle, and low) resulted in a positive treatment effect. The interactive and personalized structure of the software part of the curriculum and the emphasis on group-based collaborative work on math projects seems to help all students, regardless of initial math ability.

Within treatment schools, we took advantage of the wealth of data provided by the MCT software. This paper is among the first to employ this data set from the MCT in evaluating student outcomes. Even the IES paper for Congress (Campuzano et al., 2009) published for the U.S. Department of Education's What Works Clearinghouse, widely considered the most comprehensive report on the effect of technology use in U.S. classrooms, only incorporated the actual time logged in (the equivalent to our "usage hours" MCT variable) from the various softwares. This is especially relevant since the IES report generally showed a lack of technology usage time mattering for achievement purposes, as measured by outside exams. We would concur with that finding. But in this paper, *accomplishing* more of the program is positively related to higher posttest scores even after controlling for school effects and pretest scores. Simply spending more time logged in to the software (usage hours) is not significantly related to posttest scores, but actually completing more of the 14 units or mastering more of the 57 sections or 552 skills was.

We developed a process for evaluating the degree of efficacy for proper implementation in schools, and a way to rate four different characteristics of schools that matter in the implementation of the MCT. Those schools which were more prepared to handle the demands of this new curriculum saw their students accomplish more of it. Though this finding seems completely intuitive, we believe it is worth particular emphasis. Schools with sufficient computers, reliable internet connections, and committed principals and teachers saw students accomplish more of the software program. In turn, those students could be expected to achieve higher scores on the exams. The adoption of the MCT curriculum requires large investments in time and money. School administrators need to prepare their teaching staff

and supply enough technological infrastructure to realize the educational benefits of the investment. With the proper inputs, the processes of the MCT curriculum, including both the software component and the in-class changes, lead to student improvements in math abilities and enjoyment of the coursework. The take away message for schools considering the MCT is best summed up by the admonition that if you are going to do something, do it right.

There are some alternative explanations to the results in this paper, though we have tried to mitigate them to the best of our ability. It is possible there is a large selection effect occurring here, and that these findings are not broadly reflective of the results a random school would find if they adopted the MCT. In other words, the findings are constrained to those schools willing to participate. In large part, we do not disagree with that statement. This is a substantial change from a traditional textbook- and lecture-based curriculum. School administrators who are unwilling to make the effort to ensure the best possible learning environment and computer facilities very well may not see a positive treatment effect. This is supported by the results on school characteristics even when considering schools who wanted to adopt the new curriculum and ended up doing just that. In addition, this concern is often encountered in education policy papers. Due to the randomization within the group of schools willing to participate in treatment, we believe that our results show high internal validity. The required tradeoff in a clean experimental design often necessitates concerns over external validity. The inclusion of both random and fixed effects through HLM (such as in Equation (4)) is the most that can be done in this setting to show broader applicability.

The amount of missing data was not terribly disconcerting in this study. Mexican school 4 did not return our surveys. Otherwise, we matched enrollment figures to test scores to MCT data for a very high percentage of students. For example, in Chilean treatment schools, our original enrollment-based target was 313 students. We ended up with two test scores (pre and post) and MCT data for 310 of those same students. The overall rate of missing information is similar to other education studies that actually report on the issue, if not better. We were able to match almost every treatment student with an initial and final

test score to his software data. We observed no students transferring from the treatment condition to the control condition, or vice versa.

The results presented here are robust to other specifications, including just school- or classroom-random effects models (instead of incorporating HLM, which includes fixed effects). We also estimated the treatment effect model using difference scores as the dependent variable and excluding the independent pretest scores on the right hand side. The results are basically the same, which is not surprising considering that we had initially balanced samples based on pretest scores.

It is possible that these results are driven largely by teacher effects, rather than any effect from the curriculum itself. We unfortunately have no information on the teachers other than their names, so it is impossible to check balance between treatment and control groups on teacher observables (such as years teaching or highest degree attained) or previous teacher-student output (such as last year's student standardized test scores by teacher). However, we have no reason to believe there are large discrepancies here. In fact, in all the Mexican schools and some of the Chilean ones, the same teacher taught all the classes within the same school. If there are some pedagogical learning gains from incorporating the MCT curriculum, this would result in mitigated treatment effects since the control groups would receive some benefit, too.

The attitude data provides an additional way to think about the effectiveness of the MCT. The reported ease of use seems consistent with the actual improvements in math scores. The positive responses in Table 13 suggest that the students would be open to future classroom uses of the software. Reporting positive outcomes for both test scores and attitudes represents a more comprehensive picture of effectiveness.

There were specific research questions posed earlier in this paper. We have shown the following:

1. The MCT improves math performance for treated students over control students.

2. All process indicators except for usage hours are significantly and positively related

to math performance indicators. Accomplishing more of the MCT curriculum is predictive of larger posttest scores for students in treatment.

3. Better school characteristics, especially infrastructure and implementation, lead to increases in MCT completion and mastery.

4. Students are able to understand and use the MCT on their own and receive help from the teacher when it is needed. The students also believe that the MCT is an effective tool for learning math at the pre-algebra level.

Our initial strategy was to see how the MCT impacted learning of mathematics. Our road map for future research is to examine more closely the effect of the MCT conditional on school, teacher, and student characteristics and to identify how they contribute to improving math performance. We would also like to incorporate more details from the software, such as hint-seeking behavior of the students, in a fashion similar to Equation (6). Finally, we need to learn how performance in a tutor-based class impacts on math performance in subsequent years. The IES reports led by Campuzano (2009) and Dynarski (2007) have shown that many math achievement gains using technology-infused curricula fade after the initial year unless the students continue using similar curricula through future grades.

Essay 3

Student Hint-Seeking Behavior Using the Math Cognitive Tutor

1 Introduction

The Math Cognitive Tutor series has many possible mechanisms that combine to produce the learning gains demonstrated by the positive effects on test score in my last chapter. One possibility is that students using the MCT receive more problems to practice, and this more extensive, iterated process of practice and feedback leads to larger learning gains. Another possibility is the nature of the differentiation for which the software is purposefully designed. Many scholars (e.g. Hattie, 2009) believe this personalization is essential to improving student performance, and traditional classrooms with one teacher lecturing at the exact same pace to 25 students are particularly ill-equipped to differentiate pace on a topic-by-topic basis amongst learners. In the MCT, students receive more practice on the skills in which they do not demonstrate proficiency while more quickly skipping through those which they are deemed to understand. A third potential avenue for the effectiveness of the MCT lies in its hint feature. Intelligent tutoring systems such as the MCT offer students the opportunity to ask for a series of hints on each step of a problem. The hints are gradually more helpful until the final hint essentially gives away the answer. This feature of the MCT may help students quickly overcome misunderstandings on certain skills, reduce unproductive time, and learn more efficiently (Anderson et al, 1989; McKendree, 1990). The value of the hint feature is explored in this essay. Ultimately, the issue at hand is whether or not the help provided to students by the *Bridge to Algebra* MCT in the form of hints is actually helpful.

An ideal evaluation of the MCT hint system would follow a similar experimental design to that seen in essay 2. All students would use the MCT. Half of them would randomly be selected for the treatment group, and they would have access to the hint feature throughout the evaluation. The other half in the control group would not be able to request hints. All students would take an outside pre- and post-test, and a specification similar to Equation (4) in essay 2 would reveal whether or not having access to the hint feature improves student performance. For equity and other considerations, this was not possible in the Chilean and Mexican research projects. The evaluation of the hint system can only occur in this context

by looking retrospectively at hint usage of the students who used the MCT.

There are difficulties in evaluating the effectiveness of hints or any other help-seeking behavior by students in this context. The first lies in defining a relevant outcome measure. What exactly does it mean in the context of the MCT to "help" a student? Does it mean to assist them in performing correctly on the next problem, or does it entail a lasting learning effect that can be picked up in future attempts or on an outside exam? In essay 2, it was obvious that the outside exam scores were the relevant outcome measure, especially since the exams themselves were forms of the national standardized exams and we were evaluating the effectiveness of this new software-based math curriculum. But in the context of evaluating the micro features of the hint system, it is difficult to argue that hints requested on certain steps of problems should readily translate to performance on the post-test. Rather, I will follow the example (while changing the methodology) of much of the learning science literature (see, for example, Beck et al (2008), Lee et al (2008), and Goldin, Koedinger, and Aleven (2012)) and focus on student performance on the next few problems following a hint request.

Second, the absence of a student-specific counterfactual to a hint request and subsequent performance can easily lead to a problem of selection bias. Students who request more hints are more likely to lack the knowledge and skills necessary to solve the problems correctly. Simply comparing the outcomes of "hint-takers" to peers who did not request hints would support the notion that hints produce lower outcomes.

Other researchers have explored the effectiveness of hints in the MCT in other contexts. The most common methodologies for evaluating the effect of hints on performance in the learning science literature are referred to as learning decomposition and Bayesian knowledge tracing. Learning decomposition (Beck (2006), Zhang, Mostow, and Beck (2007), and Beck (2007)) extends the classic learning curve (described briefly in essay 2) by taking into account the heterogeneity of different learning opportunities for a single skill.

$$Performance = A * e^{-b(t_1+\beta*t_2)} \tag{1}$$

In the equation above, A measures student performance on the first trial and b the learning rate for trial opportunities of types t_1 and t_2. The different types of learning opportunities (for example, factoring a number as a step in a problem involving greatest common factors might be different from factoring a number in order to decompose a quadratic) are allowed to have differential effects on the performance measure, but ultimately the goal of learning decomposition is to estimate a learning curve.

Bayesian knowledge tracing (BKT) is another form of evaluating student knowledge using MCT data. BKT assumes that student knowledge is represented by a set of binary variables denoting whether each of many skills has been mastered or not (Yudelson, Koedinger, and Gordon (2013)) based on a binary outcome on student-problem steps of either right or wrong. Student knowledge can increase (learn) or decrease (forget) over time, and it is determined by performance on specific problems (with built in "guess" and "slip" parameters denoting a correct response when a student does not know the skill and an incorrect one when he does, respectively). Tutor interventions, including hint requests, can teach students and hence increase actual knowledge, or simply "scaffold" short-term performance with no requisite increase in knowledge (Beck et al (2008)). Given sufficient student-problem observations for each skill, Bayes nets can be constructed that estimate all of the specific parameters of the model. The most important parameter estimate for determining the value of hints in these models denotes the probability of learning a skill following a hint request. Beck et al (2008) estimate the effect of help being an 8% relative improvement over no help in a reading tutor context.

In the learning science literature, other papers have explored or advanced these methodologies to estimate the effect of the hint system. While some papers (Aleven et al (2004), Aleven et al (2003)) have shown that students occasionally "abuse" the hint system (usually by clicking through hints without absorbing any of their content), most work shows evidence that on-demand hints can positively impact learning. Aleven and Koedinger (2000) showed that asking for help after one or two errors on a step was associated with fewer errors on the given step and a reduction in the time needed to complete the step (as compared to another hintless attempt at solving the problem). They also showed that lower-performing

students have low metacognition regarding hint requests; in other words, they struggle with knowing what they don't know and hence request a sub-optimal number of hints. Goldin, Koedinger, and Aleven (2012) showed that hints are more likely to be requested by less proficient students and on steps that are more difficult. Further, success after first hints is less likely than after second or third hints, and the more proficient the student, the less likely it is that the student will benefit from a hint. In a setting using math tutors that explicitly provided hints at various points, Arroyo et al (2000) showed that boys seemed to do better with the lower hint version, their self-confidence was harmed by increased hints, and they spent less time looking at hints than female counterparts.

In the economics of education literature, the education production function generally regards the output of a student (test score) as being determined by implications of school spending levels (teacher salaries, classroom size) and student demographics (race, gender, parental influence, economic status). Hanushek's (2003) meta-analysis showed that these traditional inputs do not matter much for the output. In that light, more recent work has looked at students as active decision makers whose effort choices matter for education output. The personalized learning environments offered by MCTs are increasingly prevalent in educational settings, and they are beginning to be evaluated by economists. Haelermans and Ghysels (2014) showed that MCTs lead to a substantial and significant increase in math performance growth. The authors also focused on the returns to practice time using the tutor, revealing that increased effort on the students' parts leads to improved performance (this contradicts my findings in essay 2). Fryer (2010) showed that paying students to improve test scores outright was ineffective for doing so, but paying them to read more books was effective at raising scores. His conclusion was that input incentives can raise test scores by increasing student effort, but that output incentives on test scores do not increase that same effort because students are unaware of their own education production function. In other words, they are often unable to make the link on their own that reading more books could help improve reading test scores.

There are three different approaches in the economics of education lierature for dealing with student effort. The first methodology, advanced by McKenzie and Staaf (1974),

explained that effort is tranformed into new knowledge according to some learning rate, determined primarily by the student's aptitude. Basically, they posited a linear knowledge accumulation function as written below.

$$FinalKnowledge = InitialKnowledge + Aptitude * Effort \tag{2}$$

Equation (2) is in the same spirit as the learning curve equations explored earlier, albeit with a different functional form. Here, effort was unobservable but it was backed out by solving the equation above using IQ scores for *Aptitude* and separate exam scores at the beginning and end of an experiment for the difference in *Knowledge*. This model is in the same vein but much simpler than the Bayesian knowledge tracing models used in the learning science literature since the exam scores are considered perfect proxies for knowledge levels (i.e. student performance is exactly equal to student knowledge, with no attendant guess or slip parameters). Cooley (2010a, b) introduced the concept of students making decisions regarding how much effort to put forth in the classroom, where costs depend on peer and teacher influences and effort is unobservable to the econometrician but can be proxied for using another measure, such as ability. Dickey and Houston (2013) and Babcock and Betts (2009) surveyed students and teachers, respectively, to measure student effort, and then included both ability (from SAT scores) and effort as regressors in predicting learning. Both papers showed that effort is more important than ability in the education production function.

However, with the explosion in the education data available from tools similar to the MCT, effort need not be considered unobservable nor merely derivable from surveys. The MCT has a number of variables that could be considered student effort: usage hours, number of problems attempted, units covered, or, as this essay explores, the number of hint requests.

1.1 Hint Request Levels

Students are able to ask for hints from the MCT system at any point while solving problems. All hint requests have to originate from the student though; the system will not "ping" a

student if he struggles with errors or takes too long to attempt an answer to a step in a problem. There are three successive levels of hints offered in the MCT, and most steps of problems offer one hint at each level. The first level is feature or interface based. It offers no help on the underlying math skill involved, but instead directs the student toward the area of the screen where an answer should be input. The second level is the "meat" of the hint stage, and for some problems there can be multiple hints available at this level. These hints are definition or example based. They state the problem-solving principle that is applicable for the problem. As an example, if the problem step requires the student to list the factors of 27, second level hints might include statements such as "The factors of 27 include all numbers that can multiply to 27" or "Remember that 1 is always a factor of every number." Finally, the third level of hint is referred to as a "bottom out hint," and it supplies the answer for the student.

2 Data

In Chile, 310 students used the MCT; in Mexico, 156 students used it. The MCT log files exist at the level of the student-problem. In total, the Chilean students completed 126,004 problems (406 per student) and their Mexican counterparts completed 37,910 (243 per student). Each student-problem contains the following information from the tutor log files: student ID, unique problem identifier, unit of the problem, section of the problem, time to complete the problem, total number of hint requests on the problem, and total number of errors made on the problem. The problems are ordered in the same manner they were completed by the student, so each problem is also assigned a corresponding number within each unit and section denoting the order in which the student completed it. The student level data used in the previous work is also linked to the MCT log data. This includes pre- and post-test scores from an outside examination, aggregate MCT completion metrics (total hours spent using the tutor, units completed, skills mastered, and sections mastered), student gender, and school and classroom assignment.

It is important to point out some limitations of the existing data set so as to make it

clear what can and cannot be investigated. There are no "time stamps" in the log files. It is impossible to tell, for example, if a student completed the first two problems in a given section on a Monday and then did not log in to the system again until Friday when he attempted the next (third) problem. The log data also views each problem as a complete whole. As discussed in essay 2, one of the cognitive psychology foundations of the MCT is the breakdown of problem steps into specific skills (or knowledge components) that allow the software to ensure that students master skills, not problems, before moving forward. This is also the foundation of the learning decomposition and BKT models discussed in the introduction. There is no identification of the specific skills involved in each problem in this data. The abbreviated version of the Bridge to Algebra MCT evaluated in this study contained 552 skills across 57 sections, or about 10 skills per section. A given problem contains multiple skills. But there is no guarantee that consecutive problems, even within the same section, contain the same set, or even a similar subset, of skills. The hints and errors in a given problem are also not separated by step. Since there is no finer grained detail than the level of the problem for hints and errors, it is impossible to tell if a student, in a problem where he requested one hint and made one error, asked for that hint immediately following the error, immediately preceding the error, or on a step of the problem unrelated to the one on which he made an error. Finally, the hint requests are a count of the number of times a student clicked on the hint button. The majority of problems have three levels of hints that gradually reveal more information to the student. In the MCT data, there is no way of knowing which levels or how many of the levels were accessed by the student, even by looking at the hint counter. A student who used three hints on a given problem did not necessarily ask for all three levels of hints; instead, he could have clicked the "first hint" button three times. Despite these limitations, the data is useful for exploring the helpfulness of the hint feature.

Both the hints and errors per problem are heavily skewed in both countries since hints and errors are constrained at the lower end by zero but have no upper limit. As shown in Tables 1 and 2, the median values for both variables in both countries are very reasonable: 0 hint requests and 1 error. Tables 3 and 4 break down the lower levels of hints and

errors, respectively, in more detail. Both tables show the number (percentage) of problems completed in each country with the corresponding hints and errors. It is obvious that errors are more prevalent than hints in both the Chilean and Mexican student populations. In addition, Tables 5 and 6 show that students rarely ask for a hint on a problem without committing at least one error. The ">0 Hints, 0 Errors" cell in both Tables Tables 5 and 6 reveal that students rarely request help via the hint feature without making an error on the same problem (1.8% of student-problems in Chile and 0.9% in Mexico). While completing problems, students are either requesting a hint prior to their attempt and the hint is not providing enough learning support to allow the student to correctly answer the problem, or they are asking for hints after making mistakes and realizing they do not understand the problem. Students may not be asking for hints at productive times, as argued by Aleven and Koedinger (2000), but this essay will evaluate the effect of the hint requests they do make on improving their outcomes, as defined by a lack of errors on subsequent problems. I am forced to use performance on subsequent problems, instead of next attempts within a problem as some of the other hint effectiveness work does, because I do not have finer grained data than the student-problem level and therefore I cannot discern the hint/error sequence within a problem.

Table 1: Hints Per Problem

	Chile	Mexico
Mean	0.76	0.54
Std Dev	1.55	1.42
Median	0	0
75th %-ile	1	1
90th %-ile	2	2
95th %-ile	4	3

Table 2: Errors Per Problem

	Chile	Mexico
Mean	3.15	2.67
Std Dev	5.30	5.51
Median	1	1
75th %-ile	4	3
90th %-ile	8	7
95th %-ile	12	10

Table 3: Hints by Problem

	Chile	Mexico
0	81,311 (64.5%)	28,187 (74.4%)
1	22,813 (18.1%)	5,434 (14.3%)
2	10,253 (8.1%)	2,063 (5.4%)
≥ 3	11,627 (9.2%)	2,226 (5.9%)

Table 4: Errors by Problem

	Chile	Mexico
0	37,913 (30.1%)	13,313 (35.1%)
1	25,214 (20.0%)	8,082 (21.3%)
2	15,660 (12.4%)	4,458 (11.8%)
3	11,960 (9.5%)	3,347 (8.8%)
4	8,372 (6.6%)	2,165 (5.7%)
5	5,739 (4.6%)	1,495 (3.9%)
≥ 6	21,146 (16.8%)	5,050 (13.3%)

Table 5: Hints vs Errors in Chile

		Errors		
		0	>0	Total
	0	35,587 (28.2%)	45,724 (36.3%)	81,311 (64.5%)
Hints	>0	2,326 (1.8%)	42,367 (33.6%)	44,693 (35.5%)
	Total	37,913 (30.1%)	88,091 (69.9%)	

Table 6: Hints vs Errors in Mexico

		Errors		
		0	>0	Total
	0	12,961 (34.2%)	15,226 (40.2%)	28,187 (74.4%)
Hints	>0	352 (0.9%)	9,371 (24.7%)	9,723 (25.6%)
	Total	13,313 (35.1%)	24,597 (64.9%)	

2.1 Within Student Comparison

Since I do not know when in the sequence of the problem a student makes an error or requests a hint, I have to look across problems to evaluate hint effectiveness. My within-student comparison shows the improvement a student makes over a three-problem sequence when he requests at least one hint on the second problem. Think of a three-problem sequence within a section for a given student that occurs at any point in the overall problem sequence of that section. On the first and third problems, he does not request a hint. On the second problem, he requests (at least) one hint. I am interested in the difference in performance on the 1st and 3rd problems of the sequence. Notice that this sequence will necessarily leave out many student-problem observations, and many sequences in which hints could be evaluated. For example, a four-problem sequence featuring hint requests on the middle problems 2 and 3 but no request on 1 and 4 would *not* be included here. If a student asks for a hint on every question in a section, he is not included. If he never asks for a hint, he is not included. If he asks for a hint on consecutive questions, those questions cannot be included. The only thing I am including is the "problem 1: no hint - problem 2: hint - problem 3: no hint" sequence on three consecutive questions.

Tables 7 and 8 show the results on three relevant outcome measures *Incorrect* is the probability that the student made at least one error on the given problem, *Errors* is a count of the number of errors on that problem, and *Time* measures the total time, in seconds, required for the student to complete the problem. For all three measures of student performance, smaller values are better. The data in the columns labeled "1st Problem" and "3rd Problem" is listed as mean (standard error) and the t-value in the subsequent column comes from the two-sample t-test with unequal variance and has the usual significance designations (* 10%, ** 5%, *** 1%). Tables 7 and 8 support the notion that the hints provided by the MCT are beneficial for short-term improvements in student performance. Students were significantly less likely to make an error on the third problem in the sequence than on the first. They also made fewer total errors and finished in less time on the third problem.

Table 7: Improvement from 1st to 3rd Problem in Chile

N = 5,607	1st	3rd	t-value
Incorrect	0.64 (0.006)	0.62 (0.006)	2.70 ***
Errors	1.58 (0.037)	1.36 (0.028)	4.81 ***
Time	85.9 (1.9)	74.8 (1.2)	4.98 ***

Table 8: Improvement from 1st to 3rd Problem in Mexico

N = 1,641	1st	3rd	t-value
Incorrect	0.68 (0.012)	0.62 (0.012)	3.38 ***
Errors	1.70 (0.056)	1.36 (0.043)	4.77 ***
Time	92.2 (2.7)	84.4 (2.6)	2.06 **

However, there are a few drawbacks to this approach. First, this does not provide a robust view of hint-seeking since a specific pattern was needed to include the student problems as an observation. Second, as the "power law of learning" says (briefly explained in essay 2), students should gradually improve their performance as the number of practice problems increases. Therefore, we would expect that students improve as they move from the first to third problem of *any* sequence, regardless of the presence of a hint request. Although some of the improvements seen above can certainly be explained by the additional practice, the median "three-problem sequence" in Tables 7 and 8 in both countries occurs during the 6th-8th problems in a section. Across all students, the average errors per problem and time to complete a problem are actually *higher* in the 8th problem than the 6th (for students completing both), though this is most likely due to a sample selection problem itself. Because more proficient students are able to advance in the MCT once they demonstrate mastery of the attendant skills in a section, some of the better students do not attempt the 6th or higher problems in a given section. Tables 7 and 8, while informative, are certainly not a complete case for the effectiveness of the hint mechanism in the *Bridge to Algebra*

MCT.

3 Model

As suggested in the learning science literature, student learning curves show quick, steep drops in many outcomes (time to complete a problem, probability of making an error, probability of asking for a hint) as the student progresses through the first few practice problems related to a given skill. Looking across all students, this inverse relationship is supported in the data in both countries. Figures 1 and 2 show the rapid early decline within sections in hint requests and errors in Chile.[43] After the fourth problem in the section, the rates of both errors and hints flatten out for a long stretch before increasing again around the 11-12th problem due to more proficient students moving on to the next section.

Since the bulk of the action in reducing error and hint request rates occurs within the first few problems of a section, I focus my attention on those problems exclusively. Consider students i, section of the MCT j, and schools k. Define Y as the difference in errors between problems 3 and 1 within a given section and $Hints$ as the total number of hint requests on problems 1 and 2. If students improve their performance and make fewer errors as they progress, then Y, as defined by "problem 3 errors - problem 1 errors," will be negative. Using the first and third problems as the comparison in error reduction follows the same pattern as the within-student comparison shown previously.

$$Y_{ijk} = \beta_0 + \beta_1 Hints_{ij} + \beta_2 Gender_i + \beta_3 Pretest_i + \beta_4 Gen_i * Hint_{ij} + \mu_j + \mu_k + \epsilon_{ijk} \quad (3)$$

School fixed effects and section fixed effects are also included in the main specification. The section fixed effects are meant to pick up differences in difficulty across sections since harder sections both elicit more hints from students and cause more errors per problem.[44]

[43] The Mexico figures look largely the same but are omitted in this document. They are available from the author upon request.

[44] The hint and error probabilities (per problem) within a given section are highly correlated - 0.75 in

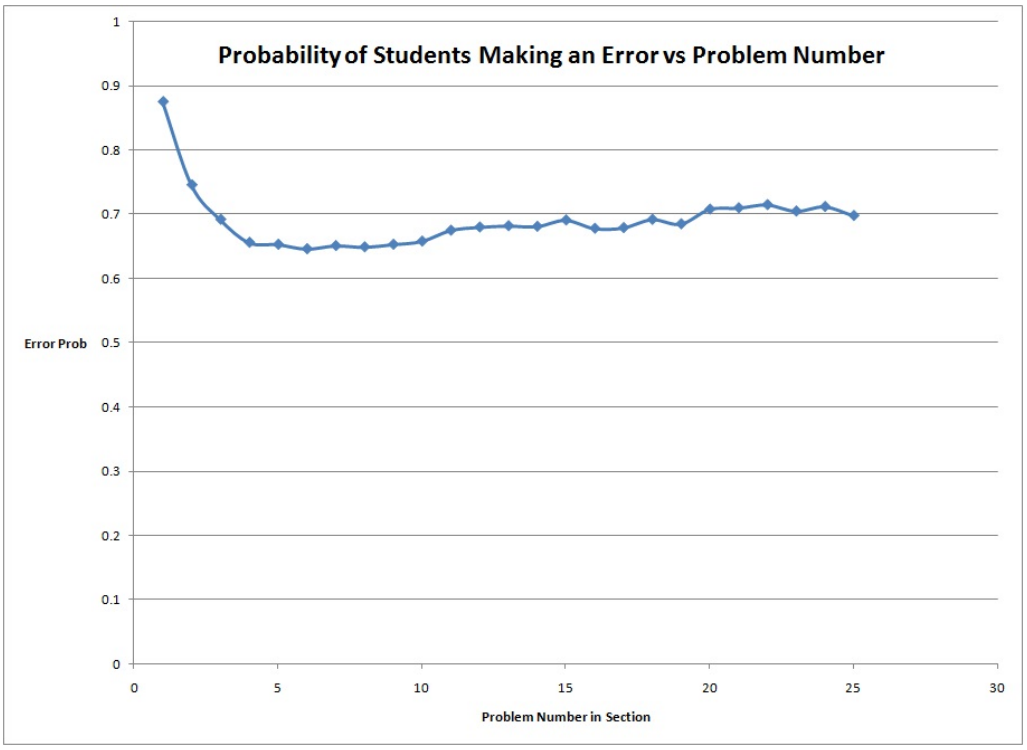

Figure 1: Error Rates on Problems (Chile)

The main parameter of interest in β_1. It represents the marginal improvement in error reduction from problem 1 to 3 of asking for an additional hint on problems 1 and 2 in a section. A negative point estimate of β_1 is supportive of the ability of the hint feature to improve student performance since it means that problem 3 errors will decrease by *more*.

Endogeneity of the hint count could be a problem if, as suggested by other work, less apt students request more hints. In order to control for the differences in difficulty, and hence hint requests, across sections, I instrument for $Hints_{ij}$ using $Hints_{-ij}$.[45]

Two sets of student-section observations are considered in the analysis. The first includes all of the student-sections observed in both countries. In order to control for the fact that more capable students are able to advance through more sections (in the same given amount of time), I want to ensure that the results of the first set of regressions are not

Chile and 0.71 in Mexico.

[45]This also entails instrumenting for $Gen_i{*}Hint_{ij}$ with $Gen_i{*}Hint_{-ij}$.

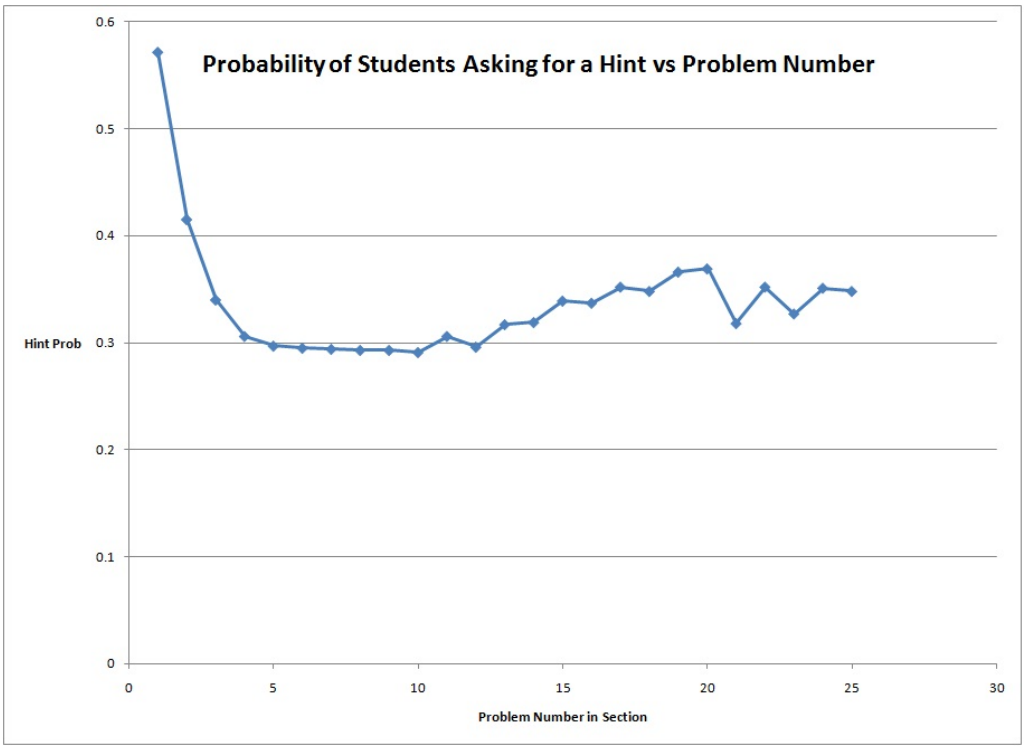

Figure 2: Hint Rates on Problems (Chile)

driven by better students completing more sections and hence recording more student-section observations. The second set of regressions includes only those sections that were completed by at least half the students in each country, respectively.

4 Results

The results shown below are separated by country, regression type (OLV vs IV), and observation set. For the IV regressions including all students, the first stage results are also included in Tables 11 and 12 in order to demonstrate the strength of the included instruments.[46] Tables 9 and 10 show the OLS results for Chile and Mexico, respectively, using the full set of student-section observations. The top rows of both Tables show the coef-

[46]The first stage results from the IV regressions using only the smaller set of student-sections are available from the author, but are omitted here for brevity.

Table 9: Performance Improvement and Hint Requests in Chile

Hints	-0.346***	-0.354***	-0.356***	-0.397***	-0.357***
	(0.022)	(0.023)	(0.023)	(0.026)	(0.035)
Gender		-0.218***	-0.213***	-0.190***	-0.064
		(0.073)	(0.073)	(0.070)	(0.084)
Pretest		-0.013**	-0.016**	-0.032***	-0.032***
		(0.006)	(0.007)	(0.007)	(0.007)
GenHint					-0.083*
					(0.043)
School fixed?	No	No	Yes	Yes	Yes
Section fixed?	No	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 6,960

ficient estimates on *Hints*. In all specifications, the estimate is negative and significant, indicating that every hint that a student requests on problems 1 and 2 in a section will further *reduce* the errors made on problem 3 respective to problem 1. Based on Table 2, these effects are quite substantial. In Chile, students averaged 3.15 errors per problem, across all student-problems completed. In the full OLS specification of Equation (3), each additional hint request on problems 1 and 2 reduces the net problem 3 less problem 1 errors by 0.36 errors. The OLS results also suggest that male students[47] benefit more from hints than their female classmates. As expected, the parameter estimate on *Pretest* is negative, indicating that more proficient students with higher pretest scores improve their error rate from problems 1 to 3 more than students with lower scores.

[47]Those who have *Gender* = 1.

Table 10: Performance Improvement and Hint Requests in Mexico

Hints	-0.652***	-0.660***	-0.677***	-0.632***	-0.548***
	(0.041)	(0.041)	(0.041)	(0.043)	(0.053)
Gender		-0.151	-0.158	-0.142	0.063
		(0.103)	(0.103)	(0.099)	(0.107)
Pretest		-0.010	-0.011	-0.007	-0.009
		(0.009)	(0.010)	(0.010)	(0.010)
GenHint					-0.211***
					(0.078)
School fixed?	No	No	Yes	Yes	Yes
Section fixed?	No	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 2,955

Tables 11 and 13 show the first stage IV results using $Hints_{-ij}$ to instrument for $Hints_{ij}$ when all student-sections are included. Informally, the strength of the instruments is demonstrated by the significant F-statistics and significant point estimates (shown in bold) of the instruments in the first stage regressions. Further, in all specifications, the Cragg-Donald Wald statistic rejects the null hypothesis that the instruments are weak. The second stage results are shown in Tables 12 and 14. The results largely reflect those shown by the OLS regressions - namely, there is a negative and significant point estimate for $Hints$, so requesting a hint will improve student performance by reducing errors more as students move from problem 1 to problem 3 within a section. In both regression types, the Mexican students benefit from each additional hint they request in problems 1 and 2 more than their Chilean counterparts. The differences by gender that were shown in the OLS results are somewhat mitigated by the IV estimation.

The smaller subsets of observations used in the results below only include the sections that were completed by at least half the students. The OLS results look much the same as above when comparing Tables 9 and 15 for Chile and Tables 10 and 16 for Mexico. The IV results remain substantially the same in Mexico (Tables 14 and 18) between the two samples, although the magnitude of the hint effect is a bit mitigated with the smaller sample. However, in Chile, the results are significantly different. The coefficient on $Hints$ switches and is significant and positive. The coefficient on $Pretest$ is also positive in all IV specifications for Chile with the smaller set of student-sections, and it is significant when school fixed effects are not included. This suggests that better students, as proxied by their pretest score, would not expect to decrease their errors as much as less apt peers.

5 Discussion and Conclusion

The Chilean and Mexican 7th grade students who had the opportunity to study pre-algebra with the MCT incorporated hints into their usage of the software. In the early problems of each section, as all students rapidly make learning gains as measured by reduced errors, hints, and time required to solve problems, requesting a hint(s) increases those gains. The

Table 11: Performance Improvement and Hint Requests in Chile - IV Specification, First Stage Results, Endogenous Variable Across Top

	Hints	Hints	Hints	Hints	GenHints
Hints Instrument	**0.967***** (0.023)	**0.971***** (0.023)	**0.980***** (0.023)	**1.04***** (0.033)	0.009*** (0.023)
Gender		-0.056 (0.042)	-0.061 (0.042)	0.122 (0.081)	0.110* (0.057)
Pretest		-0.069*** (0.004)	-0.054*** (0.004)	-0.055*** (0.004)	-0.027*** (0.003)
GenHint Instrument				-0.120*** (0.046)	**0.908***** (0.032)
School fixed?	No	No	Yes	Yes	Yes
F-stat	1692	711	284	253	494

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 6,960

Table 12: Performance Improvement and Hint Requests in Chile - IV Specification, Second Stage

Hints	-0.222***	-0.223***	-0.225***	-0.170***
	(0.044)	(0.044)	(0.043)	(0.060)
Gender		-0.207***	-0.201***	-0.033
		(0.073)	(0.073)	(0.137)
Pretest		-0.005	-0.008	-0.009
		(0.007)	(0.008)	(0.008)
GenHint				-0.110
				(0.088)
School fixed?	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 6,960

Table 13: Performance Improvement and Hint Requests in Mexico - IV Specification, First Stage Results, Endogenous Variable Across Top

	Hints	Hints	Hints	Hints	GenHints
Hints Instrument	**0.880***** (0.048)	**0.907***** (0.047)	**0.921***** (0.046)	**0.946***** (0.062)	0.013 (0.040)
Gender		-0.152*** (0.042)	-0.157*** (0.052)	-0.100 (0.107)	0.050 (0.069)
Pretest		-0.048*** (0.005)	-0.049*** (0.005)	-0.049*** (0.005)	-0.030*** (0.003)
GenHint Instrument				-0.057 (0.093)	**0.861***** (0.060)
School fixed?	No	No	Yes	Yes	Yes
F-stat	338	152	101	86.4	184

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 2,955

Table 14: Performance Improvement and Hint Requests in Mexico - IV Specification, Second Stage

Hints	-0.976***	-0.963***	-0.979***	-1.01***
	(0.104)	(0.101)	(0.099)	(0.133)
Gender		-0.190*	-0.197*	-0.262
		(0.106)	(0.106)	(0.203)
Pretest		-0.024**	-0.025**	-0.025**
		(0.010)	(0.011)	(0.011)
GenHint				0.067
				(0.206)
School fixed?	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 2,955

Table 15: Performance Improvement and Hint Requests in Chile

Hints	-0.284***	-0.298***	-0.302***	-0.413***	-0.411***
	(0.030)	(0.030)	(0.031)	(0.076)	(0.060)
Gender		-0.097	-0.095***	-0.086***	-0.077
		(0.097)	(0.097)	(0.061)	(0.068)
Pretest		-0.022**	-0.027***	-0.035***	-0.035***
		(0.009)	(0.010)	(0.011)	(0.011)
GenHint					-0.006
					(0.049)
School fixed?	No	No	Yes	Yes	Yes
Section fixed?	No	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 3,827

Table 16: Performance Improvement and Hint Requests in Mexico

Hints	-0.530***	-0.520***	-0.524***	-0.536***	-0.434***
	(0.064)	(0.064)	(0.064)	(0.108)	(0.123)
Gender		-0.084	-0.068	-0.073	0.107
		(0.130)	(0.131)	(0.160)	(0.177)
Pretest		0.014	0.011	0.010	0.006
		(0.012)	(0.012)	(0.014)	(0.015)
GenHint					-0.232***
					(0.077)
School fixed?	No	No	Yes	Yes	Yes
Section fixed?	No	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 1,447

Table 17: Performance Improvement and Hint Requests in Chile - IV Specification, Second Stage

Hints	0.352***	0.348***	0.343***	0.315***
	(0.075)	(0.074)	(0.074)	(0.106)
Gender		-0.077	-0.069	-0.148
		(0.104)	(0.104)	(0.229)
Pretest		0.022**	0.009	0.009
		(0.011)	(0.011)	(0.011)
GenHint				0.054
				(0.150)
School fixed?	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 3,827

Table 18: Performance Improvement and Hint Requests in Mexico - IV Specification, Second Stage

Hints	-0.433***	-0.449***	-0.467***	-0.595***
	(0.139)	(0.135)	(0.132)	(0.165)
Gender		-0.075	-0.063	-0.298
		(0.133)	(0.133)	(0.218)
Pretest		0.017	0.014	0.019
		(0.013)	(0.014)	(0.015)
GenHint				0.304
				(0.288)
School fixed?	No	No	Yes	Yes

Estimated std errors are reported in parentheses.

Significance denoted as ***1%, **5%, *10%

N = 1,447

hint feature of the software is effective at helping students reduce their errors on the early problems.

In much of the learning curve analyses, hints and errors are considered equals - both show a lack of mastery on the student's part, and both suggest more testing of that specific skill is warranted. This work used a different approach to evaluating the hint feature of the MCT from much of the learning science literature. Instead of estimating student-specific parameters related to their mastery of skills, and then how much the hints may have helped that learning process, I take a much more "macro" view of the hint requests. Overall, students reduce their errors over the first few problems when encountering the new math material in new sections. The results in this essay show that the students who make use of the hints can expect to reduce their errors *more*, regardless of the specific skills being tested in each problem. Asking for a hint in the MCT can certainly be viewed as "effort" on the part of the student since an active decision separate from attempting to solve the problem steps must be made. In that case, students who put forth more early effort in sections by asking for more hints can expect to reduce their errors more quickly early on.

In the longer term, the MCT might be better served by updating its hint system to include an automatic hint feature. This and other work has demonstrated the overall usefulness of the hints feature using a variety of methodologies. Aleven and Koedinger (2000) showed that some students struggle to understand when to ask for hints. In light of the evidence of hint effectiveness and low metacognition on help-seeking, the MCT could improve its pedagogy by targeting hints at specific students at specific times. Based on the results in this essay, adding an early problem in each section that students solve with hint guidance at each step might speed up the learning process.

References

Abadie, A. & Imbens, G. (2006). "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica*, *74:1*, 235-67.

Abdulkadirgolu, A., Angrist, J., Dynarski, S., Payne, T., & Pathak, P. (2009). "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." Working Paper no. 15549, National Bureau of Economic Research, Cambridge, MA.

Ahn, H., & Powell, J. (1993). "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics*, *58*, 3-29.

Aleven, V., & Koedinger, K. (2000). "Limitations of Student Control: Do Students Know When They Need Help?" In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*. Berlin, Germany: Springer-Verlag. 292-303.

Aleven, V., McLaren, B., Roll, I., and Koedinger, K. (2004). "Toward Tutoring Help Seeking: Applying Cognitive Modeling to Meta-Cognitive Skills." In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*. Berlin, Germany: Springer-Verlag. 227-39.

Angrist, J. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, *80*, 313-36.

Angrist, J., Imbens, G., & Rubin, D. (1996). "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, *91*, 444-56.

Anderson, J., Conrad, F., & Corbett, A. (1989). "Skill Acquisition and the LISP Tutor." *Cognitive Science*, *13*, 467-505.

Anderson, J. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. (1993). "Problem Solving and Learning." *American Psychologist*, *48*, 35-44.

Anderson, J. (2002). "Spanning Seven Orders of Magnitude: A Challenge for Cognitive Modeling." *Cognitive Science*, 85-112.

Anderson, J., & Schooler, L. (1991). "Reflection of the Environment in Memory." *Psychological Science*, *2*, 396-408.

Anderson, J. & Schunn, C. (2000). "Implications of the ACT-R Learning Theory: No Magic Bullets." In R. Glaser, *Advances in Instructional Psychology (Vol. 5)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Arroyo, I., Woolf, B., Royer, J., Tai, M., & English, S. (2010). "Improving Math Learning Through Intelligent Tutoring and Basic Skills Training." *Intelligent Tutoring Systems Part II: 10th International Conference, ITS 2010*, 423-432. Pittsburgh, PA: Springer Berlin.

Arroyo, I., Beck, J., Beal, C., Wing, R., & Woolf, B. (2001). "Analyzing Students Response to Help Provision in an Elementary Mathematics Intelligent Tutoring System. In *Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments*, 34-46.

Babcock, P., & Betts, J. (2009). "Reduced-Class Distinctions: Effort, Abililty, and the Education Production Function." Working Paper no. 14777, National Bureau of Economic Research, Cambridge, MA.

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 1235-64.

Barnard, J., Frangakis, C., Hill, J., & Rubin, D. (2003). "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association*, *98*, 299-311.

Baron, R. & Kenny, D. (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology*, *51:6*, 1173-82.

Beck, J., Chang, K., Mostow, J., and Corbett, A. "Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology." Intelligent Tutoring Systems, ed. Beverly Woolf. 383-394.

Bjorklund, A., & Moffitt, R. (1987). "The Estimation of Wage and Welfare Gains in Self-Selection Models." *Review of Economics and Statistics*, *69*, 42-49.

Blundell, R., Gosling, A., Ichimura, H., & Meghir, C. (2007). "Changes in the Distribution of Male and Female Wages Accounting for Parameter Sets in Econometric Models." *Econometrica*, *75:2*, 323-63.

Cabalo, J., Jaciw, A., & Vu, M. (2007). *Comparative Effectiveness of Carnegie Learning's Congitive Tutor Algebra I Curriculum: A Report of a Randomized Experiment in Maui School District*. Palo Alto, CA: Empirical Education.

Campuzano, L., Dynarski, M., Agodini, R., Rall, K., & Pendleton, A. (2009). *Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts.* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Casas, I., Goodman, P., & Pelaez, E. (2011). "On the Design and Use of a Cognitive Tutoring System in the Math Classroom." *Technology for Education (T4E), 2011 IEEE International Conference*, 9-17. Chennai, India: Curran.

Casas, I., Imbrogno, J., & Vergara, A. (2013). "A Cognitive Tutoring Strategy for Math Teaching and Learning in Latin America." Manuscript submitted for publication.

Chernozhukov, V., Hong, H., & Tamer, E. (2007). "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica*, *75:5*, 1243-84.

Chong, A. (2011). *Development Connections: Unveiling the Impact of New Information Technologies.* New York, NY: Palgrave MacMillan.

Cook, T. (2003). "Why Have Education Evaluators Chosen Not to Do Randomized Experiments?" *The ANNALS of the American Academy of Political and Social Sciences*, 114-149.

Cooley, J. (2010). "Can Achievement Peer Effect Estimates Inform Policy? A View from Inside the Black Box." Working Paper, under revise and resubmit at *Review of Economics and Statistics.*

Cooley, J. (2010). "Classroom Peer Effects." *The New Palgrave Dictionary of Economics*, Eds. Steven N. Durlauf and Lawrence E. Blume, Palgrave Macmillan, 2010.

Cullen, J., Jacob, B., & Levitt, S. (2006). "The Effect of School Choice on Student Outcomes: Evidence from Randomized Lotteries." *Econometrica*, *74:5*, 1191-1230.

de Ferranti, D., Perry, G., Gill, I., Guasch, L., Malloney, W., Sanchez, C., & Schady, N. (2003). *Closing the Gap in Education and Technology.* Washington, DC: The World Bank.

Das, M., Newey, W., & Vella, F. (2003). "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies*, *70*, 33-58.

Dickey, S. & Houston, R. (2013). "Student Choice of Effort in Principles of Macroeconomics." *Journal of Economics and Economic Education Research*, *14:2*.

Dobbie, W., & Fryer, R. (2009). "Are High Quality Schools Enough to Close the Achievement Gap? Evidence from a Social Experiment in Harlem." Working Paper no. 15473, National Bureau of Economic Research, Cambridge, MA.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., & Campuzano, L. (2007). *Effectiveness of Reading and Mathematics Software Products: Findings From the First Student Cohort*. National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Fisher, R. (1935). *Design of Experiments*. New York: Hafner.

Frangakis, C., & Rubin, D. (2002). "Principal Stratification in Causal Inference." *Biometrics*, *58*, 21-29.

Frolich, M., & Lechner, M. (2010). "Combining Matching and Nonparametric IV Estimation: Theory and an Application to the Evaluation of Active Labor Market Policies." Working Paper, http://ideas.repec.org/p/usg/dp2010/2010-21.html.

Goldin, I., Koedinger, K., & Aleven, V. (2012). "Learner Differences in Hint Processing." In *Proceedings of the 5th International Conference on Educational Data Mining*, Chania, Greece.

Haelermans, C., & Ghysels, J. (2014). "The Effect of an Individualized Online Practice Tool on Math Performance: Evidence from a Randomized Field Experiment." Working Paper, Presented at AEA 2014, Philadelphia, PA.

Hansen, L.P. (1982). "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, *50:4*, 1029-53.

Hanushek, E., & Woessman, L. (2008). "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature*, 607-688.

Hanushek, E., & Woessman, L. (2009). *Schooling, Cognitive Skills, and the Latin American Growth Puzzle*. Cambridge, MA: National Bureau of Economic Research.

Hastings, J., Kane, T., & Staiger, D. (2010). "Heterogeneous Preferences and the Efficacy of Public School Choice." Working Paper, http://justinehastings.com/research/jh-papers/item/108-hastings-kane-staiger-rev2010.

Hattie, J. (2009). *Visible Learning - A Synthesis of Over 800 Meta-Analysis Relating to Achievement*. New York, NY: Routledge.

Heckman, J. (1974). "Shadow Prices, Market Wages, and Labor Supply." *Econometrica*, *42:4*, 679-94.

Heckman, J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica*, *46*, 931-60.

Heckman, J. (1979). "Sample Selection Bias as a Specfication Error." *Econometrica*, *47:1*, 153-61.

Heckman, J. (1990). "Varieties of Selection Bias." *American Economic Review Papers and Proceedings*, *80*, 313-18.

Heckman, J., Ichimura, H., & Todd, P. (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies*, *64:4*, 605-64.

Heckman, J., & Robb, R. (1985). *Alternative Methods for Evaluating the Impact of Interventions in Longitudinal Analysis of Labor Market Data.* Cambridge: Cambridge University Press.

Heckman, J., Urzua, S., & Vytlacil, E. (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics*, *88:3*, 389-432.

Heckman, J., & Vytlacil, E. (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, *73:3*, 669-738.

Heckman, J., & Vytlacil, E. (2007). "Econometric Evaluations of Social Programs." Pts 1 and 2. In *Handbook of Econometrics*, vol 6B, ed. James Heckman and Edward Leamer. Amsterdam: Elsevier/North-Holland.

Horowitz, J., & Manski, C. (2000). "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association*, *95*, 77-84.

Hoxby, C., & Murarka, S. (2009). "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement." Working Paper no. 14852, National Bureau of Economic Research, Cambridge, MA.

Hoxby, C., & Rcokoff, J. (2005). "The Impact of Charter Schools on Student Achievement." Working Paper, http://www.rand.org/content/dam/rand/www/external/labor/seminars/adp/pdfs/2005hoxby.pdf.

Imbens, G., & Manski, C. (2004). "Confidence Intervals for Partially Identified Parameters." *Econometrica*, *72*, 1845-57.

Imberman, S. (2011). "Achievement and Behavior in Charter Schools: Drawing a More Complete Picture." *Review of Economics and Statistics*, *93:2*, 416-35.

Koedinger, K., & Anderson, J. (1993). "Effective Use of Intelligent Software in High School Math Classrooms." *Proceedings of the AIED 1993 World Conference on Artificial Intelligence in Education*, 241-248. Charlottesville, VA: Association for the Advancement of Computing in Education.

Koedinger, K., Anderson, J., Hadley, W., & Mark, M. (1997). "Intelligent Tutoring Goes to School in the Big City." *International Journal of Artifical Intelligence in Education*, 30-43.

Lee, D. (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, *76:3*, 1071-1102.

Lee, L.F. (1979). "Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables." *Econometrica*, *47*, 977-96.

Manski, C. (1990). "Nonparametric Bounds on Treatment Effects." *American Economics Review Papers and Proceedings*, *80*, 319-23.

McKendree, J. (1990). "Effective Feedback Control for Tutoring Complex Skills." *Human Computer Interaction*, *5*, 381-413.

McKenzie, R. & Staaf, R. (1974). "An Economic Theory of Learning." Blacksburg, VA: University Press.

Moffitt, R. (2008). "Estimating Marginal Treatment Effects in Heterogeneous Populations." *Annals of Economics and Statistics*, *91/92*, 239-61.

National Research Council. (2007). *Rising Above the Gathering Storm.* Washington, DC: National Academy Press.

Newell, A., & Rosenbloom, P. (1981)." Mechanisms of Skills Acquisition and the Law of Practice." In J.R. Anderson (Ed.), *Cognitive Skills and Their Acquisition*, 1-55. Hillsdale, NJ: Lawrence Erlbaum Associates.

Neyman, J. (1923/1990). "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles." Section 9 (trans. from Polish). *Statistical Science*, *5*, 465-80.

Pane, J., McCaffrey, D., Steele, J., Ikemoto, G., & Slaughter, M. (2010). "An Experiment to Evaluate the Efficacy of Cognitive Tutor Geometry." *Journal of Research in Educational Effectiveness*, 254-281.

Quandt, R. (1972). "A New Approach to Estimating Switching Regression Models." *Journal of the American Statistical Association*, *67*, 306-10.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical Linear Models.* Thousand Oaks, CA: Sage Publications.

Ritter, S. (2011). *The Research Behind the Carnegie Learning Math Series.* Pittsburgh, PA: Carnegie Learning.

Ritter, S., & Morgan, P. (2002). *An Experimental Study of the Effects of Cognitive Tutor Algebra I on Student Knowledge and Attitude.* Pittsburgh, PA: Carnegie Learning.

Ritter, S., Anderson, J., Koedinger, K., & Corbett, A. (2007). "The Cognitive Tutor: Applied Research in Mathematics Education." *Psychonomics Bulletin and Review*, 249-255.

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

Rouse, C. (1998). "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics*, *113*, 553-602.

Rubin, D. (1974). "Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, *66*, 688-701.

Rubin, D. (1976). "Inference and Missing Data." *Biometrika*, *63*, 581-92.

Rubin, D. (1978). "Bayesian Inference for Causal Effects." *Annals of Statistics*, *6*, 34-58.

Scheurman, F., & Pedro, F. (2009). *Assessing the Effects of ICT in Education: Indicators, Criteria, and Benchmarks for International Comparisons.* Paris, France: European Union.

Schwab, K. (2010). *The Global Competitiveness Report: 2009-2010.* Bern, Switzerland: World Economic Forum.

Yudelson, M., Koedinger, K., & Gordon, G. (2013). "Individualized Bayesian Knowledge Tracing Models." In *Proceedings of 16th International Conference on Artificial Intelligence in Education (AIED 2013)*, Memphis, TN. 171-180.

Zhang, J., & Rubin, D. (2003). "Estimation of Causal Effects Via Principal Stratification when some Outcomes are Truncated by Death." *Journal of Educational and Behavioral Statistics*, *28*, 353-68.

Zhang, J., Rubin, D., & Mealli, F. (2009). "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification." *Journal of the American Statistical Association*, *104*, 166-76.