ESSAYS ON THE ECONOMICS OF EDUCATION

By

Andrew Jacob Bibler

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics - Doctor of Philosophy

2016

ABSTRACT

ESSAYS ON THE ECONOMICS OF EDUCATION

By

Andrew Jacob Bibler

Chapter 1 estimates the impact of dual language and immersion education on student achievement. Dual language classrooms provide English Language Learners (ELLs) an opportunity to receive instruction in their native language as they transition to English fluency. This might allow ELLs to build a stronger foundation in core subjects and lead to better academic outcomes. Dual language and immersion education have also grown substantially in popularity among English speaking families across the U.S., as they present an option to learn content in, and presumably become fluent in, a second language. Despite the spike in practice, there is little causal evidence on what effect attending a dual language school has on student achievement. I examine dual language and immersion education, and student achievement using school choice lotteries from Charlotte-Mecklenburg School District, finding local average treatment effects on math and reading exam scores of more than 0.06 standard deviations per year for participants who were eligible for English second language (ESL) services or designated limited English proficient (LEP). There is also some evidence that attending a dual language school led to a lower probability of having limited English proficient status starting in third grade. For applicants who were not eligible for ESL services or designated as LEP, attending a dual language school has resulted in higher end of grade exam scores of about 0.09 and 0.05 standard deviations per year in math and reading, respectively.

Chapter 2 builds upon recent research discussing apparent gender differences in returns to parental investments. Parental time investments are one potential mechanism, if correlated with household structure differentially by gender, that could help explain documented gender differences in non-cognitive skills, as well as the sensitivity of outcomes and behavior to family structure for boys. This paper investigates gender differences in parental time investments around changes in household composition. I find that, although both boys and girls experience decreases in parental time investments following a change in family structure from a two-parent to a single-parent house-

hold, the loss for boys is relatively large. The difference is strongest in paternal weekday investments, for which boys lose an additional 24 minutes per day, equivalent to roughly 35% of average paternal weekday investments. There is no significant evidence that mothers compensate for the loss by increasing their investments to boys relative to girls.

Chapter 3 discusses the construction of confidence intervals in teacher value-added (VA). The use of teacher value-added models to measure teacher effectiveness is expanding rapidly, with teacher value-added estimates being incorporated into teacher evaluation systems and potentially high-stakes decisions. However, we still know little about the precision of value-added estimates, or the performance of the resulting confidence intervals, which could play an important role in the decision-making of districts and policymakers. Our study aims to fill a gap in the literature by providing comparisons of confidence interval performance for the OLS-Lag (DOLS) value-added estimator. We use simulated student achievement data to study the behavior of standard errors and confidence intervals for teacher value-added estimates under several student grouping and classroom assignment scenarios. We propose a simple method for calculating confidence intervals, which includes a critical value adjustment, and compare it's performance to that of the confidence intervals you get from using a typical variance estimator with standard normal critical values. We find that this method generally leads to 95% confidence interval coverage rates near 95%, which we consider a desirable feature. In particular, the method has advantages over standard confidence interval calculations when value-added estimates are based on a small number of classrooms per teacher and students are grouped on unobservable heterogeneity. We then use student-level administrative data to compare standard errors and confidence intervals across the different methods. Our proposed method assigns 18.5% of teachers in the top decile of the value-added distribution confidence intervals with a lower bound above the 75th percentile of the distribution, a much lower percentage than that from ignoring correlation in unobservables, 60%, or using cohort-school clustering, 75.8%. These differences indicate that the policy conclusions drawn from using the value-added estimates may depend on the choice of confidence interval. Hence, knowledge of the reliability of value-added models could be a critical part of the decision making process of administrators and policy makers.

ACKNOWLEDGEMENTS

I would like to thank Todd Elder for his guidance and help throughout. I am also grateful to Scott Imberman, Stacy Dickert-Conlin, and Madeline Mavrogordato for serving on my dissertation committee and always offering helpful advice, as well as Kelly Vosters, and Michelle Maxfield for providing feedback on the different chapters. I have also benefited from useful discussions and comments from many other classmates, faculty, and seminar participants throughout this process. I would like to note that the third chapter is co-authored with Cassandra Guarino, Kelly Vosters, and Jeffrey Wooldridge, and thank them for working with me.

I am grateful to the North Carolina Education Research Data Center (NCERDC), and Charlotte-Mecklenburg School District for providing the data used in the first chapter, and in particular to Kara Bonneau, Susan Frieje, and Lindsay Messinger for their assistance in that process. This research was supported by a Pre-Doctoral Training Grant from the IES, U.S. Department of Education (Award R305B090011) to Michigan State University, and by IES Statistical Research and Methodology grant R305D10028.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Dual Language Education and Student Achievement

1.1 Introduction

Dual language classrooms use a non-English language of instruction for a significant amount of the curriculum. They are primarily used to provide instruction to English language learners (ELLs) in their first language and to promote bilingualism and biculturalism among native English speakers. There are two types of dual language classrooms.[1] *Two-way* classrooms typically enroll students from two different language backgrounds and teach curriculum in both languages. There were only about ten such programs in the U.S. in 1980, but that number was almost 250 by 2000 Howard and Sugarman [2001]. In contrast, most students in a *one-way* (*immersion*) classroom share a similar language background, but receive instruction in a second language. The number of one-way classrooms registered with the Center for Applied Lingustics increased from fewer than 50 to almost 450 over the last few decades (Center for Applied Linguistics, 2011). Recent expansions in several states have driven these numbers even higher.[2] Despite the growth, there is little causal evidence on the effect of dual language education on student achievement. In this paper, I use school choice lotteries from Charlotte-Mecklenburg School District (CMS) to estimate the effect of attending a dual language school on student achievement.

Districts target dual language education to two types of students: English language learners (ELLs) who might benefit from receiving instruction in their home language, and other students who want to have instruction in a second language. For ELLs, dual language education might

[1] Although there are two different types of classrooms studied here, two-way dual language and language immersion, I will refer to both as dual language for simplicity.

[2] Utah passed legislation for funding of dual language programs in 2008, and since then has implemented programs in 118 schools in 22 districts. New York City added or expanded 40 programs in 2015.

allow for an easier transition to full English instruction, providing a potential route for improving outcomes of the growing and struggling ELL population. The alternative is often placement in an English-only classroom coupled with English second language (ESL) services, which could mean missing important classroom instruction time and disruption to the student and his or her peers, and ultimately making students more likely to fall behind. On the other hand, placement in an English-only classroom might expedite the development of English skills, leading to faster re-classification out of ELL status and higher scores on standardized exams that are written in English.

Districts also target dual language education to English speaking students who desire to learn in a second language, and the influx of dual language programs seems to be driven in large part by demand from English speaking families Watanabe [2011].[3] The primary goal for districts in offering dual language education to English speakers is to provide an option that allows them to become bilingual, biliterate, and bicultural. In addition, districts can use dual language schools, as well as other specialized programs, to offer a more diverse set of schooling options and compete with charter and private schools to retain students residing in the district boundaries. Dual language programs are often promoted using high test performance of participants as evidence of increased cognitive development Maxwell [2012, 2014]. However, lack of formal training in English could slow progress as measured by scores on standardized exams. It is unclear whether dual language programs would increase or decrease test scores for this group of students.

Whether dual language education has any effect on student achievement and the direction of those effects are empirical questions, but there is very little causal evidence due to endogeneity from self selection into the programs. In this paper, I estimate the causal effect of attending a dual language school on achievement on standardized math and reading exams by exploiting quasi-random assignment from oversubscribed admissions lotteries. I focus on students who applied through the Charlotte-Mecklenburg school choice lottery for their kindergarten year, and specified a dual language school as their first choice. I use quasi-random assignment to a dual language school through the lottery as an instrument for dual language school attendance to identify the local average treatment effect of dual language schooling on achievement. The treatment differs by whether or not the

[3] About 70 percent of the estimation sample in this study are students who were never identified as English language learners or limited English proficienct in the data.

student uses English as a home language. For a native English speaker, the treatment is to receive instruction in a second language and the alternative to attending a dual language school is receiving instruction in their home language. For ELLs, the typical alternative to attending a dual language school is to receive instruction in English (not their home language) accompanied by other ESL services. Because of this divide in treatment, I estimate effects separately for two subgroups using a proxy for whether or not the student was proficient in English when they entered school.

The first group is made up of students who were eligible for English second language (ESL) services or were designated as limited English proficient (LEP)[4] at any point in the data. I will refer to this group of students as the "ESL/LEP" sample. In the ESL/LEP sample, I find that attending a dual language school leads to increased scores on math and reading exams of more than 0.06 standard deviations per year of participation. Furthermore, I find some evidence that attending a dual language school has led to a lower probability of being designated as limited English proficient in grades three through six for students in the ESL/LEP sample. The second subgroup is made up of students who were never eligible for ESL services or designated LEP. I will refer to this group as the "English" or the "non-ESL/LEP" sample. Among this group, I estimate that attending a dual language school leads to 0.09 standard deviations higher achievement in math per year of participation, and 0.05 standard deviations higher achievement in reading per year. The estimates are statistically significant and generally robust to a number of alternate specifications.

1.1.1 Literature

Bilingual education broadly refers to educational programs that are targeted toward ELLs and include some amount of home language instruction. Drawing conclusions from previous literature is complicated by the fact that bilingual education can take several forms, and the degree to which home language instruction is used varies within and across program types. Two-way dual language

[4]North Carolina uses the term limited English proficient (LEP) to refer to students who do not use English as a primary language in their home, and score below a specified cutoff on an English skills test. There is some variation among researchers and school districts in how they refer to this group of students. The term English language learner (ELL) is often used in place of, or interchangeably with LEP. I use the term LEP when referring to the designation given to students in North Carolina because that is the term the state uses, but use the terms ELL and LEP interchangeably when referring to this group of students generally.

classrooms tend to use the non-English language for a large proportion of instruction (50% or more) throughout elementary school. Instruction is not generally based on current English ability of ELL students, as two-way classrooms enroll dominant speakers of both languages and provide instruction in both languages. Other forms of bilingual education, such as transitional bilingual education and structured English immersion, are more focused on expediting English fluency and do not necessarily group ELLs and non-ELLs in the same classroom.[5]

Some prior research focuses on the achievement gap for ELL students in dual language (DL) programs, showing that ELLs participating in DL programs have higher test scores than ELLs in non-dual language classrooms. Two detailed reports on six districts in North Carolina, including CMS, find that ELLs in two-way programs score higher than students in English-only classrooms on end-of-grade exams Thomas and Collier [2009], Thomas et al. [2010]. Collier and Thomas [2004] summarize 18 years of results on one- and two-way programs from 23 different school districts. Students in both program types close at least 70% of the ELL test score gap by the end of fifth grade, but this could be driven by self selection.

One technique sometimes employed in an attempt to overcome the self selection issue is matching on pretest scores or other observable characterisitcs. Cazabon, Lambert, and Hall [1999] studied a two-way program in Cambridge, MA. They used a pretest to match each DL student with a control, and show that ELLs assigned to a dual language classroom outperformed the control group on English based math and reading exams. Cobb, Vega, and Kronauge [2009] also match students on observable characteristics to consider the impact of DL education on achievement and find positive effects in writing and math for native Spanish speakers, with the effects being more pronounced one year after completion of the program.

Several attempts have been made by researchers to summarize estimates from the most methodologically rigorous studies. In a meta-analysis on transitional bilingual education programs, Rossell and Baker [1996] deemed only 25 percent of the studies they considered to be methodologically acceptable. A transitional bilingual education model teaches reading in the native language in early grades, but moves to complete English instruction as early as second grade. They determined that

[5]See Valentino and Reardon [2015] for a good description of some of the differences between dual language classrooms and other forms of bilingual education.

a relatively small percentage of the most rigorous studies estimate positive effects of transitional bilingual programs. While other meta-analyses agree that much of the prior research is flawed, they suggest that among acceptable studies there are positive effects of bilingual programs across subjects and in different types of programs Greene [1998], Slavin and Cheung [2005], Willig [1985]. Most of the literature they examined does not account for selection bias, but there are a handful of studies that used random assignment in an attempt to estimate causal effects. However, these studies were generally based on small samples (e.g. less than 175 students) and from nearly 30 years ago Greene [1998].

More recently, Valentino and Reardon [2015] use data on student preferences from a large urban district to compare the test scores of students in bilingual education and English-only programs. They study student performance on exams across program types conditional on the type of program that the student preferred. The assignment is quasi-random, but they do not use knowledge of the assignment mechanism to completely exploit the randomness in the assignment. They find that dual language students progress faster in math and English language arts performance after second grade, leading to better long-run performance than any of the other three programs (including English immersion) Valentino and Reardon [2015]. Similar to this study, Steele et al. [2016] exploit random assignment from oversubscribed admissions lotteries into dual language programs in Portland, Oregon. They report mostly positive, insignificant effects for dual language students on reading and math exam scores. However, there are two important differences in this study. First, they pool two subgroups of students - ELLs and non-ELLs - together, despite the fact that the treatment effects for these two groups could be of different signs and magnitudes, which would have important policy implications.[6] Another important difference between the Steele et al. study and this paper is that the dual language programs in Portland Public Schools (PPS) are strand programs, meaning that they only make up a portion of the school. All of the dual language programs in CMS are housed in three schools, where every classroom in the school is a dual language classroom.

Other recent studies have examined causal effects of bilingual education programs, but the classrooms in these studies are not necessarily two-way dual language programs. Slavin et al. [2011]

[6] About nine percent of students are ELLs at the time of application and fifteen percent have a non-English home language Steele et al. [2016].

use random assignment in kindergarten to either an English immersion or transitional bilingual classroom to study differences in English and Spanish reading scores for several years following the assignment. They found that students assigned to a transitional classroom scored lower on English reading exams in early grades, but there were no statistically significant differences by fourth grade. Guo and Koretz [2013] use a difference-in-differences framework to study the effect of a Massachusetts policy that shifted the early elementary education for ELLs from a several year transitional bilingual model to a one-year sheltered (or structured) English immersion model. Structured models target instruction to the current English ability of the students while expediting English fluency relative to transitional models, so this represents a clear shift away from instruction in the home language of the students. Similar to the findings of Slavin et al. [2011], Guo and Koretz find that the policy had no effect (or a small positive effect) on fourth grade English reading scores. Chin et al. [2013] use district level variation in the number of LEP students in Texas to study whether having a bilingual education option improves achievement for LEPs and their non-LEP peers. They identify the treatment effects using the discontinuity generated by a Texas rule that districts with at least twenty LEP students who share a common language in a specific grade must offer a bilingual education option to those students. They do not find significant increases in the test scores of LEP students from districts that offer bilingual education, but do find an increase in the scores of non-LEPs in districts that offer bilingual programs. Their findings suggest that offering bilingual education resulted in positive spillover effects to non-LEP students. Estimating peer effects directly has rarely been done in this setting and with mixed evidence Cho [2012], Geay et al. [2013].

While test scores are one outcome of interest, districts may also care about the duration of LEP classification of language minority students. When a student enrolls in a district in North Carolina, their parent takes a survey that asks about the languages the student uses at home. The district uses that survey, and possibly interviews with the parents and/or student, to determine the home language of the student. If the home language of the student is not English, then the student must take a test that determines LEP status and eligibility for ESL services. When a student is identified as LEP based on the score of the placement test, they are required to continue testing annually until they are re-classified out of LEP status. LEP classification is important for several reasons.

It is another measure of student progress that differs from the math and reading exams. Second, students with LEP status may be eligible for testing accommodations. Lastly, offering ESL services is costly, so districts benefit from programs that expedite reclassification, all else equal. Umansky and Reardon [2014] show that dual language participants in a large urban district are reclassified out of LEP status at a slower rate in early grades, but have higher total reclassification and English proficiency than students from English immersion classrooms by the end of high school. Similarly, Steele et al. [2016] report slower reclassification out of ELL status for dual language participants throughout elementary and middle school. When estimating treatment effects though, they find that attending a dual language classroom led to a higher probability of exiting ELL status starting in fifth grade. In addition to estimating the effect of dual language schooling on achievement, I estimate the effect of dual language schooling on LEP classification among students who were ever identified as LEP or eligible for ESL services.

For native English speakers, there is concern that attending dual language schools may promote bilingualism at the expense of achievement as measured by standardized tests, which are written in English. The question for many parents is whether their child can attend a dual language school and become bilingual without falling behind in other subjects. Learning in a second language could create confusion or frustration that would negatively impact achievement, especially in the short-run. On the other hand, the mental juggling involved with thinking in two languages might promote cognitive development. This theoretical connection has previously been made in research related to working memory, which is used to store and process information and execute related tasks Baddeley and Hitch [1974], Baddeley [2003], Alloway [2010]. Working memory can be considered a measure of ability to learn. It is strongly correlated with academic outcomes and much of the growth in working memory capacity takes place before adolescence Alloway [2010]. Working memory is closely associated with second language acquisition Baddeley [2003], native language vocabulary Dufva et al. [2001], and listening and reading comprehension Chrysochoou et al. [2011], Dufva et al. [2001], but empirical evidence directly supporting a causal link between second language acquisition and cognitive development through working memory is sparse. In this specific setting, students apply for entry into a DL school for their kindergarten year but don't take their first high stakes

exam until third grade, so students and teachers have some time to overcome any initial difficulties in adjusting to the new language. The gap in time between school assignment and testing allows teachers and administrators to commit to teaching in the second language in the first few years of school when there are no high stakes exams looming. If working in two languages can boost cognitive development, then one might expect it to show up in this environment.

Some prior literature has pointed out the positive achievement gap for English dominant students in dual language programs. English speaking participants of two-way programs in North Carolina score higher than their peers on end-of-grade exams and have better attendance Thomas and Collier [2009], Thomas et al. [2010]. Other research uses matching on pretests and observables, finding that learning a portion of curriculum in a second language does not hinder progress and might be associated with positive effects on achievement in reading Cobb et al. [2009], Cazabon et al. [1999]. Again, Steele et al. [2016] find that dual language instruction led to positive, although often insignificant, effects on math and reading scores. Their effects are not directly comparable because they pool all students, but their sample is comprised of mostly native English speakers.

The most important contribution of this study is to provide causal estimates of the effect of attending a dual language school on achievement using school choice lotteries in Charlotte-Mecklenburg School District and, in particular, estimating separate treatment effects for two groups regarded very differently for policy purposes. This is a valuable addition to the existing research on bilingual education, because most prior literature does not credibly identify any treatment effect and does not disentangle effects for English language learners compared to native English speakers. The next section provides details on the lottery used in CMS. Section 3 discusses the data and some descriptives. In section 4 the empirical strategy used for the main results is discussed. Section 5 presents the empirical results and section 6 concludes.

1.2 Lottery

Every student enrolled in Charlotte-Mecklenburg School District is assigned to a neighborhood school based on geographic zones. The district uses a school choice lottery to allocate seats for students

who wish to opt out of their neighborhood school. The empirical strategy used in this paper makes use of exogenous variation created from oversubscribed lotteries, so it is useful to describe how the lottery operates and why it facilitates the identification of treatment effects. This section provides details on the lottery.

1.2.1 Magnet Programs and Priority Groups

All CMS students can submit up to three programs in order of preference through a centralized lottery. All students with an older sibling in a school are guaranteed a seat in that school by making it their first choice.[7] Then non-guaranteed seats are assigned in three rounds. In the first round, only first choices are considered. If there are fewer applicants than seats available to a given program, then all of the applicants to that program will be assigned to their first choice. Identification comes from comparing winners and losers from the same lottery, so estimates are driven by oversubscribed lotteries. When the number of applicants is greater than the number of available seats (the choice is oversubscribed), seats are awarded quasi-randomly. Seat assignment is not completely random, because the probability of winning for a particular student depends on the priority group that the student is assigned to. Priority groups refer to sets of students that meet (or do not meet) some pre-specified criteria. In CMS, over the sample period they are based on geographic location and whether the student's neighborhood school is a Title I choice school.

With that in mind, the district gives priority with a couple of apparent goals. First, they care about transportation costs and allowing students to attend schools that are close to home. Students who live within close proximity to a full magnet school are given priority. In addition, the district is split into four geographic zones. Magnet schools offer transportation to at least one, and up to four of the zones, leading the zones to be referred to as transportation zones. Students who live in a zone served by a magnet are given priority for admission to that school over students who live in a zone that is not served by that school. Students outside of the zone can still apply, but living outside of

[7]Students who meet admission criteria and have a twin or older sibling assigned to a magnet program receive guaranteed admission to that program. The applying student must specifiy it as their first choice in order to be guaranteed admission through sibling preference. The sibling guarantee requires that the students have the same residence and at least one common parent or guardian.

the school's transportation zone means they have a lower probability of winning, all else equal. They are also required to provide their own transportation. The district also cares about equity. They show this by offering priority to students who are assigned to Title I *choice* schools. Title I schools are those with a high percentage of students eligible for free and reduced price lunch (FRPL). A Title I school becomes a Title I *choice* school if they fail to meet adequate yearly progress in the same subject for two consecutive years. No Child Left Behind (NCLB) requires that the district allow students assigned to Title I choice schools the opportunity to attend a non-Title I choice school, but it does not require the district to allow students to choose the school they are offered. In fact, they could be offered a school that they did not apply for in the lottery.

Assigning students to priority groups alters the probabilities of winning, and means that assignment is not unconditionally random. I use lottery (program of application by year by priority group) fixed effects to exploit the fact that winners should be randomly chosen within these groups. In addition to priority groups, all applicants are ordered based on randomly assigned numbers. When a choice is oversubscribed, the combination of priority groups and randomly assigned numbers determine who wins the lottery. The next subsection discusses the priority groups, and gives more detail on how lottery winners are determined during and after the first round.

1.2.1.1 Priority Groups

Seats are allocated based on priority group and lottery number. The top priority for applicants to full magnet schools in CMS is given to students who live within one-third mile of the school, but only twenty percent of seats can be assigned through that priority.[8] For example, if there are ten seats available to a specific full magnet school and more than two applicants live within one-third mile of the school, then the students with the first two numbers win under that priority. Then they move to the second priority group, students with Title I choice neighborhood schools.[9] Only ten percent

[8]Students with this priority are still subject to the lottery if demand from this priority exceeds 20% of seats available. The area can be extended beyond 1/3 of a mile by the superintendent if the number of students enrolled meeting this criteria for a specific grade is less than 15.

[9]The first seats awarded are for students who qualify for FRPL and those below their grade level in reading. For kindergarten students, below grade level in reading is defined as having a personalized education plan.

of available seats can be assigned through this priority. Continuing with the example, the student with the first number who meets the second priority is assigned a seat, but the rest of the students assigned to Title I choice schools remain unassigned. Finally, they move to the third priority, all students who live in transportation zones served by the magnet school.[10] There is no limit on the number of seats assigned through this priority, so in this example, students with the next seven numbers who live in the transportation zone are admitted. The last priority is for students from transportation zones not served by the magnet school.[11] In this example, if more than two students meet priority one, then that priority group is oversubscribed. The identification strategy relies on comparing students who met a specific priority and won with students who met that priority and did not win. Similarly, if more than one student meets the second priority, then that lottery is oversubscribed as well. Finally, if more than seven students meet priority three, then that lottery is oversubscribed. In such a case, students from all three of those priority groups contribute to the estimates. In contrast, consider what happens to the students in the last priority group, those from outside of the transportation zone. Since no students in the last priority group won a seat, those students do not directly contribute to the estimates.

After going through all first choices, second choices are considered. If a student's second choice is already full from the first round of assignments, then they remain unassigned in the second round. Then third choices are considered. All students are assigned to a default neighborhood school based on pre-determined geographic zones if not otherwise assigned in the lottery. Since the lottery considers student choices in order, students are most likely to win a choice by picking it first, and more seats are awarded in the first round than in the second or third. In the following analysis I restrict to students who made a dual language school their first choice. The treatment assignment variable is a dummy variable for winning their first choice, which should be random within lottery.

1.2.1.2 Creating Lottery Fixed Effects

[10] This priority first limits the number of seats from any particular neighborhood school assignment zone to be proportional to the potential number of applicants to the school. Then priority opens up to all students in the transportation zone. This restriction does not seem to be practically important.

[11] Applicants from outside of the transportation zone must provide their own transportation.

Although lottery fixed effects are not explicitly given in the data, I use available information to construct fixed effects. The data contain up to three choices for every student in order of preference, as well as sibling placement, Title I choice placement, FRPL status, and transportation zone.[12] I start with the sample of all applicants without a guaranteed seat and proceed in the following way to generate lottery fixed effects.

1. Proxy Title I choice school using whether or not any student from their neighborhood school was placed under the Title I choice option that year.
2. Generate priority groups using FRPL, transportation zone, and Title I choice proxy.
3. Lottery fixed effects are priority-year-program of application combinations.

Since the lottery fixed effects are generated, they are a proxy to the true lottery fixed effects. The assignment, conditional on lottery, provides the exogenous variation used to estimate causal effects.

1.3 Data

There are three dual language schools in CMS. All three are full magnet schools, meaning that admission requires a lottery application and every student in the school participates in the dual language program. Collinswood Language Academy and Oaklawn Language Academy offer two-way English-Spanish classrooms, and Waddell Language Academy (formerly Smith Language Academy) offers full immersion strands in Mandarin, French, German and Japanese. Collinswood started in 1997 and now houses grades K-8. In kindergarten, 90% of instructional time is in Spanish Thomas and Collier [2009]. In grades one through five, half of the content is taught in each language. Oaklawn is a newer program, started in 2004, but follows a similar model to that of Collinswood. The curriculum is taught 90% in Spanish in kindergarten, 75% in first grade, and 50% in grades two

[12] CMS stopped reporting FRPL after 2010. For 2011 I proxy for FRPL at the time of application using FRPL from the NCERDC data.

through five.[13] Spanish is by far the most common non-English language among students in CMS, and the two-way programs are targeted toward native speakers of both languages. The German and French one-way immersion classes offered at Waddell have complete foreign language instruction in grades K-2, whereas the Mandarin and Japanese classes teach one hour in English per day in grades K-2 Thomas and Collier [2009]. All four programs at Waddell target 90% of instructional time in the non-English language in grades 3 – 5. The one-way programs primarily target English speaking students, but they admit ELLs who speak the partner language or another language altogether.

CMS and the North Carolina Education Research Data Center (NCERDC) provided the data for this study. CMS provided eight years (2006 – 2013) of lottery results with assignment into the three dual language schools. NCERDC linked the lottery data from CMS with statewide data. The following analysis will focus on end-of-grade exam scores in math and reading, which begin in third grade. Linking the lottery data with statewide data provides information on end-of-grade exam scores, and allows for the tracking of students who leave the district but stay in the North Carolina public school system. Since lottery results could impact school attendance decisions, this helps to mitigate attrition issues Rouse [1998], Steele et al. [2016].

My analysis sample includes students entering kindergarten from the 2006-2007 through 2010-2011 school years who submitted an application for a dual language school in the CMS school choice lottery, and were linked from the CMS data to the NCERDC data.[14] I start with the 2006-2007 school year because of changes implemented that year to the lottery system, including how the priority groups were determined. End-of-grade exams start in third grade, so the last year of entry used when estimating effects on exam scores is the 2010-2011 school year.[15]

Since estimation relies on applicants with non-guaranteed seats in oversubscribed lotteries, there are a couple of things worth noting. From the first row of Table B.1, between 20 and 30 percent of the seats in each school were awarded to students with sibling guarantee. Those students are dropped from the estimation sample. The second row of Table B.1 shows the percentage of applicants to

[13]Both schools use team teaching, divided by language of instruction.

[14]NCERDC was able to link between 93% and 97% of all observations from the CMS data in each year. Among observations of rising kindergarteners in the CMS data who chose a dual language school first in the lottery over the sample period, 93.5% were matched with the NCERDC data.

[15]The latest exam scores are from the 2013 - 2014 school year. This will be updated as NCERDC releases new exam scores each year.

each school that won their first choice. Only 56 percent of applicants who listed Collinswood as their first choice won their first choice, and 78 percent of first choice applicants to Waddell won their first choice in the lottery.[16] The CMS assignment mechanism only considers first choices in the first round of seat allocation, so if a school fills up in the first round, then second and third choice applicants to that school will not win a seat. Table B.2 shows application numbers for students who chose one of the dual language schools as their second or third choice, but not as their first choice. From column 1, 49 percent of students who made a dual language school their second choice, won their first choice to a non-dual language school. About 14 percent of students who chose a dual language school with their second choice and not their first, won that choice, but only 10 percent attended a dual language school.

The CMS data also contain the neighborhood school that each student is assigned to, which helps to describe the outside options that students are foregoing to enter a dual language school. Characteristics of the neighborhood schools of the applicants are informative for thinking about the counterfactual. Language of instruction is not the only thing that changes for the student when they opt out of their neighborhood school and into a dual language school. Specifically, there could be changes in peer quality and composition of the student body. Mean characteristics of the schools that applicants are opting out of are displayed in Table B.1. Applicants to Oaklawn come from schools that have a relatively high proportion of minorities (12 percent white), 76 percent of students on free and reduced price lunch, and score 0.3 standard deviations below the state average on end-of-grade math and reading exams. They come from neighborhood schools that score worse than the average for all applicants. On the other hand, applicants to Waddell and Collinswood come from neighborhood schools with a smaller percentage of FRPL students (57 percent and 65 percent, respectively), but still score below the state average on end-of-grade math and reading exams.

The three dual language schools are generally higher performing than the other schools in their respective neighborhoods. Table B.3 shows that over 75 percent of students at Oaklawn are at grade level in reading, but only 50 percent of the students at schools in the area near Oaklawn are

[16]These percentages include students with guaranteed seats, so they are overestimates of the percentage of winners among those with non-guaranteed seats. About 43 percent of non-guaranteed applicants to Collinswood won and 69 percent of non-guaranteed applicants to Waddell won.

at grade level in reading.[17] Self selection and peer effects could play a significant role in the high performance of DL schools, but there are other features that might hurt their performance relative to neighborhood schools. Specifically, DL schools experience higher teacher turnover and begin with larger classes in kindergarten. Dual language classrooms need teachers who are fluent in the language of instruction, so the schools in CMS often recruit teachers from abroad. The teachers are permitted to work in the U.S. for a limited amount of time, leading to higher turnover. This is particularly true in Collinswood and Oaklawn. Table B.3 shows that over 50 percent of the teachers in each of those schools has zero to 3 years of experience, compared to about 30 percent in the neighboring schools and other magnets. Not all teachers and staff members in dual language schools come from abroad, nor are they necessarily fluent in a second language. Since they often implement team teaching, in most grades there is at least one English speaking teacher. Table B.3 shows that the dual language schools do have highly experienced teachers, although they have a smaller proportion than the neighboring schools and other magnets. From column seven, 37 percent of teachers at other magnets have 11 or more years of experience, but that number is only 25 percent at Collinswood (column 4) and 27 percent at Waddell (column 2). Since students can not enroll in a dual language school after kindergarten (or first grade) without meeting a minimum language requirement, the schools start with larger class sizes, anticipating some attrition throughout elementary school. The average kindergarten class has 21.4 students at Collinswood and 22.3 at Waddell, as seen in columns 2 and 4 of Table B.3. That is 3 more students than the other schools in their respective areas. From column 7 of Table B.3, other magnet schools have 18.7 students in a kindergarten class on average. Although there are several differences between dual language schools and the typical neighborhood school, greater teacher turnover and larger early elementary school class size are two characterstics of DL schools that could lead to lower achievement.

Figures A.1-A.3 provide descriptive comparisons of average standardized math and reading scores by LEP and DL status. Figures A.1-A.2 compare average standardized math and reading scores for

[17] I refer to schools in their area as the neighborhood school zone that the dual language school is in as well as all of the school zones contiguous to that zone. Since all of the dual language schools are full magnet schools, no students are automatically assigned to them. Instead each student has a neighborhood school assignment that they attend unless they opt out through the lottery, change address, or enroll in a charter or private school.

dual language and non-dual language students.[18] These are descriptive comparisions of DL and non-DL students similar to what has generally been examined in prior studies. They represent a good starting point, but ignore useful information on lottery fixed effects and sibling placement. Figure A.1 graphs the comparison for non-LEP students. Non-LEP, dual language students score well above the state average in reading in every grade, and there is a divergence between dual language students and the rest of the district from grades three through eight. In seventh and eighth grade the dual language students score more than 0.3 standard deviations above the state average in reading. Non-LEP, dual language students also score well above the state average in math, but the gap between DL students and the rest of the district displays a slight downward trend with grade. The dual language, non-LEP students score about 0.3 standard deviations above the mean in grades 4 and 5, but about 0.2 standard deviations above the mean in eighth grade. The counterfactuals in Figure A.1, lines for the non-DL students, include all non-DL students in the district, most of whom had no interest in attending a dual language school. Since there are likely systematic differences between DL applicants and non-applicants, estimates generated from this sort of analysis should not be considered causal. Figure A.2 displays the analogous comparison for students who were identified as LEP in at least one grade, three through eight. Non-dual language, LEP students score below the state average in math and reading in every grade. On the other hand, LEP students who attend dual language schools score about 0.2 standard deviations above the state average in math in third grade, and more than 0.2 above the average in every grade after that. They also score at the state average in reading in third grade, and above it in every grade after third. Once again, Figure A.2 provides evidence that LEP, DL students score above their non-DL peers in math and reading on average, but the differences should not be interpreted as causal effects.

While these graphs provide useful descriptions of the gaps in test scores for DL students, they do not provide causal evidence on the differences in scores. For causal evidence, I turn to the randomization created by the oversubscribed lotteries. Only lottery applicants with non-guaranteed

[18]Figures A.1-A.2 graph average standardized (by year and grade across the state) residualized scores. They are residuals from linear regressions of standardized exam scores on grade dummies, year dummies, sex, FRPL status, and exceptionality. For the purposes of Figures A.1-A.2, dual language students are all students who attended a dual language school in any grade, 3 - 8. Figure A.1 uses all students who were not identified as limited English proficient in any grade, 3 - 8, and Figure A.2 uses all students who were identified as limited English proficient in at least one of those grades.

seats are used to estimate causal effects because the estimation strategy relies on comparing winners and losers of the same lottery. Tables B.4 and B.5 describe the lottery winners and losers. Columns 1 - 3 describe the application sample, which includes all applicants regardless of whether they have valid testscores in the data. Columns 4 - 6 describe the estimation samples, which are restricted to applicants who remained in the sample long enough to have test scores available. Average math and reading scores on their first exam are displayed for lottery winners in column 4 and those who lost the lottery in column 5. The differences in these scores give the raw test score gaps after restricting to the estimation sample. From Table B.4, among the non-ESL/LEP students, lottery winners scored about 0.24 standard deviations (0.52 - 0.28) higher than lottery losers on their first end-of-grade math exam.[19] These differences still do not warrant a causal interpretation, because they ignore lottery fixed effects. The analogous differences between winners and losers in the ESL/LEP sample are shown in Table B.5. Lottery winners scored 0.08 standard deviations above the state mean on their first math exam, and lottery losers scored about 0.2 standard deviations below the state mean. That is a difference of about 0.28 standard deviations in favor of lottery winners on their first end-of-grade math exam. Similarly, lottery winners in the ESL/LEP sample scored about 0.24 standard deviations higher than lottery losers on their first end-of-grade reading exam, although both groups scored below the state average.

If students perfectly complied with the lottery assignment, then assignment would be synonomous with attendance and the causal effect could be estimated using OLS regressions of achievement on assignment/attendance. However, students do not perfectly comply with initial assignment from the lottery. From Table B.4, 89.9 percent of first choice lottery winners and 37.5 percent of first choice lottery losers from the non-ESL/LEP estimation subsample attend a dual language school, meaning that there is non-compliance among winners and losers. The first row of Table B.5 gives the analogous figures for the ESL/LEP subsample. Ninety-three percent of lottery winners from the ESL/LEP estimation sample attend a dual language school and 29 percent of lottery losers attend a dual language school.

Winners are not bound to attend the school they won the lottery for, and lottery losers can

[19]The scores included are from the first exam score available for each student, which is typically the third grade score.

end up in a dual language school despite losing the initial lottery. There are several ways this can happen. First, they could win a seat to a different dual language program with their second or third choice in the lottery. This is somewhat unlikely since they are typically filled up by students making them their first choice, but it does happen. From Table B.4, about 31 (21) percent of lottery losers in the non-ESL/LEP estimation sample chose a dual language school with their second (third) choice. More than 11 percent of the lottery losers in that sample won a seat in a dual language program with their second or third choice. Table B.5 shows that lottery losers from the ESL/LEP sample were less likely to choose a dual language school for their second (third) choice, as only 23 (12) percent did, and only 2 percent of them won a seat in a dual language school. Second, students who do not win their first choice are placed on a waiting list for that school, which is accessed if seats become available. If a lottery winner chooses not to take the seat offered to them, the seat is offered to the next student on the waiting list. This is likely a major source of non-compliance from lottery losers. From the non-ESL/LEP estimation sample in Table B.4, 10.1 percent of winners do not end up attending a dual language school. From Table B.5, 7.1 percent of lottery winners in the ESL/LEP estimation sample do not attend a dual language school. Even if a winning student enrolls in the dual language school and attends that school, but eventually exits, that seat can be offered to another student. The waiting list can be accessed all the way through the first academic quarter of the school year. Lastly, students can reapply in the school choice lottery for the subsequent year and win a seat.[20] As discussed in further detail in the next section, non-compliance does not invalidate the empirical strategy used in this paper.

For causal inference, assignment must be a significant predictor of attendance and must be exogenous conditional on lottery fixed effects. I first examine whether assignment is a significant predictor of attending a DL program. I test for differences in DL attendance between winners and losers, conditional on lottery fixed effects, by regressing the dummy variable for attending a DL school on a dummy for winning and lottery fixed effects. The first row of Tables B.4 and B.5 displays the estimated coefficients on the dummy for winning the lottery, which indicate whether winning the lottery actually predicts DL attendance. Rejecting the null hypothesis of no effect

[20]There is also a second lottery mainly for students who enrolled in CMS after the deadline for the first lottery, but second lottery applicants are placed at the end of the waitlist for oversubscribed programs.

indicates that winning is correlated with attendance. The test in column 6 of Table B.4 suggests that lottery winners in the non-ESL/LEP estimation sample are about 52 percent more likely to attend a dual language school than lottery losers. Column 6 of Table B.5 shows that in the ESL/LEP estimation sample winners are almost 67 percent more likely to attend a DL school. Both estimates are statistically significant, suggesting that winning the lottery is a good predictor for attending a DL school, which is necessary to implement the identification strategy used in this paper.

The remaining tests, found in columns three and six of Tables B.4 and B.5, give some indication whether the lottery results are truly random. Assignment is random within lottery groups, so the generated lottery fixed effects are included in each test. Since lottery groups depend on geographic location and free and reduced price lunch status, we shouldn't expect fixed characteristics of applicants to be unrelated to winning the lottery unconditionally. A rejection of the null hypothesis suggests that winning the lottery might be related to that characteristic in some non-random way and generally gives cause for concern about the identification strategy proposed below. Tests in column 3 of Tables B.4 and B.5 are for the application sample, which is the sample that the randomization actually took place in. None of the tests in the non-ESL/LEP application sample reject the null hypothesis. After restricting to the estimation sample, the only rejection in column 6 of Table B.4 is on the coefficient in the regression of a dummy variable for black on winning the lottery. From Table B.5, there is a rejection of the null hypothesis for the dummy variable for Hispanic in both the application and estimation samples. There are at least two reasons why a test might reject even if the initial assignment is random. The first could be from non-random attirition from the sample. Since I am estimating effects on math and reading scores, students who do not remain in the sample long enough to observe test outcomes must be dropped for estimation. Even though assignment is random at the time of application, it is not necessarily random when restricting to the applicants that remain in the sample through third grade. Staying in the district could be related to winning the lottery and the resulting attrition would lead to selection bias Rouse [1998], Steele et al. [2016].[21] While this would not explain the rejection in the ESL/LEP application sample, it could

[21] Attrition is likely higher because of the lag between application and testing and the focus on students entering kindergarten. I have at least one set of exam scores for about eighty-five percent of the applicant sample (see the rows labeled "Non-missing Test Scores" in Tables B.4 and B.5).

explain the rejection in the non-ESL/LEP estimation sample. I include initial tests in Tables B.4 and B.5 for non-random attrition, which are from OLS regressions of having at least one available set of test scores on winning the lottery. Both tests fail to reject the null hypothesis, suggesting that non-random attrition is not an issue.[22] Another possible explanation is that assigment is actually random, and the rejection of the null hypothesis is an artifact of measurement error in the constructed proxies used for lottery fixed effects. Since priority groups depend on free and reduced price lunch status and characteristics of the student's neighborhood school, I control for this flexibly by including free and reduced lunch by cohort dummy variables and neighborhood school fixed effects, in addition to the lottery fixed effects. To further alleviate concerns of endogeneity and non-random attrition, I include a number of robustness checks including using weights based on the probability of remaining in the sample.

1.4 Empirical Strategy

The effect of attending a dual language school on achievement can be estimated directly using OLS to estimate equation 1.1.

$$Y_{i,j,g,s} = \gamma \cdot 1[DualLanguage]_{i,j} + \beta \cdot X_{i,j,g,s} + \Omega_j + N_s + \varepsilon_{i,j,g,s} \tag{1.1}$$

Where $Y_{i,j,g,s}$ represents an end-of-grade math or reading exam score of student i in grade g who applied to lottery j from neighborhood school s. The key variable, $1[DualLanguage]_{i,j}$, is a dummy variable that is equal to one if the student attended a dual language school.[23] Lottery fixed effects, Ω_j, are included because winning the lottery is not unconditionally random, but students are drawn

[22] I include a more formal discussion of non-random attrition below, as well as a discussion of estimates weighted for non-random attrition.

[23] Using enrollment in the year of the exam is one way to measure participation. That leaves a lot of time between application and when enrollment is measured. One might worry that this could bias estimates since students have time to apply to other schools or simply withdraw from the dual language program, both of which are likely non-random. For this reason, I prefer using enrollment in kindergarten as the participation measure. In one year of the data (2007), the school of attendance in kindergarten is missing for a non-trivial portion of applicants, many of whom show up in a dual language school in first grade. For this reason, I actually measure attendance as showing up in a dual language school in either kindergarten or first grade.

randomly within lottery. Fixed effects for the neighborhood school[24] that the student was assigned to at the time of application, N_s, and student level covariates, $X_{i,j,g,s}$, are also included. Grades are pooled for estimation, so grade of exam dummy variables are included in $X_{i,j,g,s}$. One concern with this approach is that, although the assignment is random conditional on lottery fixed effects, compliance with initial assignment may not be random, leading to a biased and inconsistent estimator for the average treatment effect. Compliance might be non-random for a couple of reasons. In particular, over 30 percent of lottery losers end up attending a dual language school. Students who attend a dual language school, despite losing the lottery for their first choice, might be systematically different from the students who lost and did not end up attending a dual language school. For example, students who chose a dual language program with their second and/or third choice are more likely to attend a dual language school relative to those who did not specify a dual language school with their second and/or third choice. Non-compliance could represent strength of preferences for dual language schooling or for their neighborhood school, or the ability of parents to maneuver their way into their first choice school. Since OLS estimators are biased and inconsistent if attending a dual language school is non-random, I focus on estimating the intention-to-treat and local average treatment effects which are consistent when assignment is random conditional on lottery fixed effects.

I follow a standard approach for estimating treatment effects using applicants for oversubscribed lotteries Deming et al. [2014], Rouse [1998]. The intention-to-treat effect is estimated by regressing end-of-grade math and reading scores on a dummy for winning the lottery and a set of covariates in the sample of lottery applicants, as shown in equation 1.2.

$$Y_{i,j,g,s} = \gamma^{ITT} \cdot 1[LotteryWinner]_{i,j} + \beta^{ITT} \cdot X_{i,j,g,s} + \Omega_j^{ITT} + N_s^{ITT} + \varepsilon_{i,j,g,s}^{ITT} \tag{1.2}$$

Where $1[LotteryWinner]_{i,j}$ indicates whether student i was a winner of lottery j. The coefficient of interest, $\hat{\gamma}^{ITT}$, is an estimate of the intention-to-treat Imbens and Angrist [1994]. The difference between equations 1.1 and 1.2 is that equation 1.2 replaces the variable of interest, $1[DualLanguage]_{i,j}$, with the assignment variable, $1[LotteryWinner]_{i,j}$. The estimators from equations 1.1 and 1.2 are

[24]Neighborhood school refers to the school that the student was assigned to at the time of the lottery. This is the school that the student would be assigned to attend in kindergarten unless the student either opts out during the lottery, enrolls in a charter or private school, or changes address.

not estimating the same parameter, but $\hat{\gamma}^{ITT}$ is consistent under the assumption that assignment is random. Whereas, consistency of $\hat{\gamma}$ requires the less plausible assumption that attending a dual language school is random. Both the intention-to-treat and local average treatment effect estimators share this advantage over the OLS estimator from equation 1.1.

Equations 1.3 and 1.4 describe a two-stage estimation strategy using the dummy for winning the lottery as an instrument for attending a dual language school. Now $\hat{\gamma}^{LATE}$ is an estimate of the local average treatment effect, the effect for those who are induced to participate by winning the lottery Imbens and Angrist [1994]. In the main specification, the effects are estimated by pooling grades and interacting the treatment dummy with years of treatment (grade of exam plus one). Dummy variables are included for grade of exam, leading to a per-year of participation intepretation.

$$1[DualLanguage]_{i,j} = \gamma^{DL} \cdot 1[LotteryWinner]_{i,j} + \beta^{DL} \cdot X_{i,j,g,s} + \Omega_j^{DL} + N_s^{DL} + \varepsilon_{i,j,g,s}^{DL} \quad (1.3)$$

$$Y_{i,j,g,s} = \gamma^{LATE} \cdot \hat{1}[DualLanguage]_{i,j} + \beta^{LATE} \cdot X_{i,j,g,s} + \Omega_j^{LATE} + N_s^{LATE} + \varepsilon_{i,j,g,s}^{LATE} \quad (1.4)$$

I perform specification checks to alleviate concerns about exogeneity of the treatment or non-random attrition, including using weights based on the estimated probability of remaining in the sample.[25] Weighting the regressions adjusts for non-random attrition related to observable characteristics. Including neighborhood school fixed effects in all of the main estimates further restricts the comparisons to help with concerns about misspecified lottery fixed effects. Neighborhood school is defined as the school that the student would have been assigned to if they did not win any seat in the lottery, change address, or enroll in a charter or private school. Having the same neighborhood school means that the students live in the same geographic area and have the same outside schooling option. For comparison, I also show estimates from an alternative specification that does not include neighborhood school fixed effects.

In addition to estimating the effect of attendance on acheivement, I estimate the effect of attending a dual language school on limited English proficiency status. I interact the treatment and

[25] Remaining in the sample means that the student has valid end-of-grade exam scores for at least one grade. Weights are based on estimated probabilities from logit regressions of an indicator for staying in the sample on race, gender, FRPL, and a dummy for winning the lottery.

attendance variables with each grade (three through six), and estimate the effect on having limited English proficient status in each grade on the ESL/LEP sample. Prior research suggests that dual language participants re-classify at a slower rate in early grades, but eventually surpass their non-dual-language-schooled peers Umansky and Reardon [2014]. This is a good point of reference, although we should not necessarily expect these results to be the same. These are two different contexts, and I focus on estimating effects in a select subsample, unlike the district wide analysis by Umansky and Reardon [2014].

1.5 Results

I begin by providing first stage ($\hat{\gamma}^{DL}$ from equation 1.3) and treatment effect ($\hat{\gamma}^{ITT}$ from equation 1.2 and $\hat{\gamma}^{LATE}$ from equation 1.4) estimates from the main specification in Table B.6. Panel A of Table B.6 shows estimates for the non-ESL/LEP sample of applicants, students who were never identified as eligible for English second language services or as limited English proficient. The estimated effect on math scores in column 5 suggests that among compliers, attending a dual language school led to an increase in math scores of 0.089 standard deviations. This can be interpreted as a per-year gain in achievement. Column 7 shows an effect on reading scores for this sample of 0.053 standard deviations per year. Both estimates are statistically significant at the ten percent level. These estimates are promising for the growing practice of dual language education for native English speakers. At least in CMS, the dual language schools have been successful in delivering instruction in a second language, and increasing math and reading exam scores for English proficient students. Although these estimates do not separate out the mechanisms through which achievement gains are operating, they show that it is possible to successfully promote bilingualism and increase academic achievement.

Panel B of Table B.6 shows estimated treatment effects for the ESL/LEP sample. The estimated effect of attending a dual language school on math scores in column 5 is 0.078. From column 7, the estimated effect on reading scores in this sample is 0.064. Both estimates are statistically significant at the five percent level. While these estimates are large, they are in line with the fact that treatment

is muti-year and begins at a young age. The estimates suggest that dual language education can be an effective teaching method for ELLs and help to reduce achievement gaps in math and reading. Consider the achievement gaps between LEP and non-LEP students, which are displayed in Figure A.3. The district average math and reading scores for LEP students are below the state averages in every grade, and below the district non-LEP averages in every grade. The largest gap in math scores is about 0.2 standard deviations, so the estimate of 0.078 standard deviations per year is large enough to more than close that gap by third grade. The largest disparity in reading scores is a little more than 0.4 standard deviations. The estimated effect on reading scores of 0.064 standard deviations is enough to close the gap in test scores by the end of elementary school.

As shown in Table B.7, the estimates are not sensitive to the omission of neighborhood school fixed effects. The estimate on reading scores for the non-ESL/LEP sample without neighborhood school fixed effects, shown in column 7 of Table B.7, is 0.057 and statistically significant at the ten percent level. The estimated impact on math scores in that sample, reported in column 5, is 0.086, and is statistically significant at the five percent level. Estimated effects in the ESL/LEP sample without neighborhood school fixed effects are reported in Panel B of Table B.7. The estimated effect on math scores is reported in column 5. It is a little smaller than in the main specification, now 0.063, but still statistically significant at the five percent level. The estimated effect on reading scores, reported in column 7, increases from 0.064 to 0.069, and is still statistically significant at the five percent level. The exclusion of neighborhood school fixed effects makes very little difference. All four of the estimates remain positive, in the same range as the initial estimates, and statistically significant.

Since non-random attrition would lead to inconsistent estimators and test scores are missing for a non-trivial portion of applicants, I include estimates that are weighted by the inverse of the estimated probability of having test scores in the data. I estimate the probability of remaining in the sample long enough to have valid test scores using logit regressions on dummy variables for race/ethnicity, gender, FRPL, and winning the lottery, then use the inverse of the estimated probabilities as weights in the estimation. The estimated probabilities are summarized in Panel A of Table B.8. From column 3 in Table B.8, the estimated probability of having a set of test

scores in the data among those who won in the non-ESL/LEP subsample is almost 85 percent, and the estimated probability for those who lost the lottery is only slightly lower at 82 percent. The estimated average partial effect of winning on remaining the sample is 0.015, and not statistically different from zero. Similarly, in the ESL/LEP subsample, the estimated average partial effect of winning on remaining in the sample is 0.019 with a standard error of 0.042. The estimates suggest that winning the lottery is not a strong predictor of remaining in the sample, alleviating concerns about non-random attrition. Panel B shows linear tests for non-random attrition, including tests that condition on lottery fixed effects and neighborhood school fixed effects. From column 2 in Panel B of Table B.8, winning the lottery does not appear to be strongly correlated with remaining in the sample of English proficient students. The estimated coefficient on winning is -0.002 with a standard errro of 0.038. Among the sample of ESL/LEP students, the estimated coefficient on winning is positive, 0.052, but statistically insignificant with a standard error of 0.049. These linear tests provide further evidence that non-random attrition is not an issue, because despite attrition rates of around fifteen percent on average, attrition is not strongly correlated with winning the lottery.[26] Despite the apparent lack of correlation between winning the lottery and remaining in the sample, weighted estimates are reported in Table B.9 to show that the estimates are not sensitive to weighting. Panel A shows the inverse probability weighted estimates for the non-ESL/LEP sample. From colmun 5 of Panel A, the estimated treatment effect for math scores is now 0.089, the same as the initial estimate, and significant at the ten percent level. The weighted estimate for reading in that sample, from column 7, is 0.052, almost identical to the intial estimate. The weighted estimates on math and reading scores in the non-ESL/LEP sample are the same as the estimates from the initial specification, suggesting that non-random attrition is not likely to be a significant factor. Estimates for the ESL/LEP sample are shown in Panel B of Table B.9. The estimated average treatment effects on math and reading scores are the same as the initial estimates, and both are still statistically significant at the five percent level. Weighting has almost no impact on the point estimates, which suggests that non-random attrition is probably not inflating the estimates much,

[26]The estimates in Table B.8 are based on a single observation for each individual applicant, and a dummy variable indicating whether the student has valid test scores in any grade. Another way to investigate non-random attrition would be to expand that data to include an observation for each individual for each grade that they could have tested in. The results are not sensitive to this alternative method. See Table G.5 for more evidence.

if at all.

I report estimates of heterogeneous treatment effects in Table B.10; these allow effects to differ by gender (columns 1-2), program type (columns 3-4), or race/ethnicity (columns 5-7). Heterogeneous treatment effects are estimated by interacting dummy variables indicating mutually exclusive sets of students with the attendance variable, and using the same dummy variables interacted with the assignment variable as instruments. Estimates for the non-ESL/LEP sample are reported in Panel A. Columns 1 and 2 of Panel A show that effects on math scores for females, 0.106, are stronger than for males, 0.068. Similary, the estimated effect on reading for females is 0.069 and statistically significant at the five percent level, and the effect for males is a statistically insignificant 0.030. On the other hand, columns 1 and 2 in Panel B suggest that the effect is stronger for males in the ESL/LEP subsample. The difference in heterogeneous effects by gender between samples is somewhat striking, and may reflect the difference in treatments. A big part of the treatment for students in the ESL/LEP subsample is likely that they receive some instruction in their home language as opposed to English immersion coupled with ESL services. On the other hand, treatment in the non-ESL/LEP sample is typically receiving instruction in a second language as opposed to English immersion. The differences in heterogeneity could result from differing treatments and potentially different mechanisms facilitating the effects.

Effects for one-way and two-way programs are reported in columns 3 and 4. The difference comes down to which school the student applied to since Waddell contains all of the one-way programs and the other two schools, Collinswood and Oaklawn, house two-way programs only. The size of the estimated effects are similar by program type for the non-ESL/LEP sample, but the estimates on effects for one-way programs have much larger standard errors. For example, the estimated effect on math scores for one-way programs in that sample is 0.081, but the standard error is 0.127. The estimate on math for two-way programs is 0.090 with a standard error of 0.046. There is also a statistically significant estimated effect for two-way applicants of 0.054 on reading scores, but the estimated effect for one-way applicants is smaller and statistically insignificant. Panel B in Table B.10 reports estimated treatment effects for students in the ESL/LEP sample for one-way and two-way programs. Similar to the non-ESL/LEP sample, estimates for one-way programs are very noisy.

The estimate on math scores for one-way programs is -0.018 in this sample, but the standard error is 0.176. The estimated effects for two-way applicants in the ESL/LEP sample are 0.079 and 0.065 on math and reading scores, respectivley. Both estimates are statistically significant at the five percent level.

Finally, I estimate heterogeneous effects by race/ethnicity in the non-ESL/LEP sample in columns 5, 6, and 7 of Panel A in Table B.10. The estimated impact on math scores is largest in the white subsample, but estimates for the black and Hispanic subsamples are also positive and the estimate for the Hispanic subsample is statistically significant at the ten percent level. The estimated treatment effect on math scores for the white subsample is 0.190, which is large relative to most other estimated effects, and is significant at the five percent level. The estimated effect on the black subsample is 0.046 but it is statistically insignificant. The estimated treatment effect on math scores in the Hispanic subsample is 0.090 and significant at the ten percent level. Estimated effects on reading scores are relatively similar across the white, and Hispanic subsamples. From Panel A of Table B.10, the estimated effect on reading scores in the black subsample, 0.034, is less than half the size of that estimate in the white subsample, 0.084, but both estimates are statistically insignificant. The only significant effect on reading scores is on the Hispanic subsample, 0.115, and it is significant at the one percent level.

I do not estimate effects for each race/ethnicity in the ESL/LEP sample, because 85% of the students in that sample are Hispanic. Any estimate for other races would be unreliable. However, restricting to the Hispanic subsample using dummy interactions shows that the main finding is robust in this subsample. Column 7 in Panel B of Table B.10 shows the estimated effects on math and reading for the Hispanic students in the ESL/LEP sample. The estimated effects on math and reading scores are 0.083 and 0.062, respectively. Both estimates are statistically significant.

Table B.11 shows estimates by grade. These are estimated by interacting the DL attendance variable and/or the indicator for winning the lottery with each exam grade. The estimated impact on math exam scores for the non-ESL/LEP sample are shown in column 5 of Panel A. The estimated effect for math scores on the third grade interaction is 0.374, and significant at the five percent level. The estimated effect on math scores for sixth grade is 0.721 and significant at the five percent level.

This estimate is only identified from two of the cohorts, leading to a relatively large standard error of 0.346. The estimates for the effect on reading scores in this sample are shown in column 6 of Panel A in Table B.11, and they also appear to exhibit an increasing pattern with grade. The estimates on the third and fourth grade interactions are 0.215 and 0.151, respectively. Neither of them are statistically significant. The largest estimate is on the fifth grade term, 0.431, and it is significant at the five percent level.

Estimated effects are also reported by grade for the ESL/LEP sample in Table B.11. The estimated effects are stronger in the ESL/LEP sample, but the estimated effects for math scores do not exhibit quite as strong of an increasing pattern with grade. The effect on math scores on the third grade interaction from column 5 in Panel B of Table B.11 is 0.393 and significant at the five percent level. The estimated coefficient on the sixth grade interaction is 0.542 and is also significant at the five percent level. The largest of all of the estimated effects on math scores for the ESL/LEP sample is on the fourth grade interaction. That estimate is 0.657 and significant at the five percent level. All of the estimates on reading scores in the ESL/LEP sample are positive as well. The largest estimate, 0.531, is on the sixth grade interaction and significant at the one percent level.

In addition to estimating treatment effects on math and reading scores, I estimate the effect of attending a dual language school on LEP classification among the sample of students ever eligible for ESL services or considered LEP. I estimate the effects by regressing a dummy variable for being considered LEP in a given year on DL attendance by grade interactions. I instrument for attendance by grade interactions using a dummy for winning the lottery interacted with each grade. OLS estimates by grade are shown in columns 1 and 2 of Table B.12. Column 1 shows estimates without neighborhood school fixed effects. Every estimate in column 1 is negative, meaning that students who attend DL schools are less likely to be considered limited English proficient in each grade. The largest in absolute value is the -0.210 estimate on the sixth grade interaction and it is significant at the one percent level. The analogous treatment effects are shown in column 3. They are all negative, but only the estimate on the sixth grade interaction, -0.168, is statistically significant. These estimates are in line with the higher English reading scores, but seem to counter some results in the prior literature Umansky and Reardon [2014] yet agree with others Steele et al. [2016]. These

results are not necessarily comparable with prior literature on re-classification since estimates are specific to a set of students who applied for dual language schools in CMS. Furthermore, all of the estimates on LEP classification are noisy and most of them are not significantly different from zero. In general though, they suggest that movements forcing English immersion on ESL/LEP students might be misguided. In this setting, students attending DL schools not only score higher on math and reading exams, but they are also less likely to be considered LEP in grades 3-6.

1.6 Conclusion

Dual language magnet schools in Charlotte-Mecklenburg offer an alternative option for students to learn curriculum in a non-English language. I find that, conditional on some baseline characterstics, dual language students score higher than their peers on end-of-grade math and reading exams. One concern with this initial descriptive analysis and previous literature is that differences may be driven by self-selection. I use random assignment from school choice lotteries to estimate causal effects of attending a dual language school on student achievement. In the main specification, I estimate local average treatment effects of more than 0.06 standard deviations per year on math, and almost 0.08 standard deviations per year in reading exam scores among students who were ever eligible for ESL services or considered LEP. The effects are robust to several alternative specifications, and large enough to close the LEP - non-LEP achievement gap in math and reading if applied to an average LEP student in CMS. I find further evidence that among students in this sample, those who attend a dual language school are less likely to be considered LEP in grades three through six, although the differences are generally statistically insignificant. The estimates on achievement and LEP classification suggest that dual language education has led to large benefits for students with limited English proficiency and appears an effective way to serve the population of ELLs in CMS.

Among English first language applicants, the estimated impact on math scores of about 0.09 standard deviations per year is robust to different specifications and represents a large increase in achievement. The effect on math scores in the non-ESL/LEP subsample is substantially stronger among females and white students. The estimated effect in reading for this sample is 0.053 standard

deviations per year in the main specification. The size of the effect on reading scores is also robust across specifications. There is some evidence that effects on reading scores might be stronger among female students, but there is less evidence of heterogeneity by school type or race. For English first language students, it appears that the dual language schools in CMS provide a good opportunity for them to become bilingual and biliterate without sacrificing achievement in other areas. Not only are they not losing ground in math or reading, they are experiencing large gains in both math and reading achievement.

Future research should aim to disentangle the mechanisms that facilite the achievement gains realized by the dual language and immersion students. Although the lottery winners in both subsamples in this study have been successful at learning in a non-English language and outperforming their peers who lost the dual language school lotteries, the current study does not specify the mechanisms through which gains were realized, and therefore can not distinguish an effect of learning in a second language itself from potential differences in peer and teacher quality, among other things. Separating out these mechanisms could point to, or rule out, specific aspects of the CMS schools that are critical to the achievement gains, and should be a primary goal of future research on the topic.

Chapter 2

Household Composition and Gender Differences in Parental Time Investments

2.1 Introduction

Recent research discusses the relatively poor non-cognitive outcomes for boys raised in single-parent homes. However, we still know very little about the mechanisms facilitating these gender differences. Learning about these mechanisms is important for proposing policies or treatments that could assist boys in closing the gap in non-cognitive skills, and potentially improving outcomes for boys on other dimensions that are correlated with growing up in a single-parent household and having poor non-cognitive skills (e.g. cognitive performance, and educational attainment).

While differences in investment levels and differential returns to investments both likely play a role in generating gender gaps in non-cognitive performance, separating the importance of returns and levels of each input is a difficult task. Because many inputs are correlated with household structure, measuring the returns to one input, can be conflated by the levels of and/or returns to omitted inputs. Parental time investments, i.e. the amount of time that parents spend with their children, are a potentially important mechanism that could help explain gender differences in non-cognitive development. More specifically, if parental time investments are important in the production of non-cognitive skills, and they depend on household composition differentially for boys and girls, then time investments could help explain gender gaps in non-cognitive outcomes. In this paper, I focus on gender gaps in the level of time investments and how they relate to household structure, i.e. whether the child lives in a two-parent or single-mother household. Because fathers tend to spend relatively more time with boys as they age Baker and Milligan [2013], and single-parent households are more often headed by the mother, growing up in a single-parent household could be

more detrimental for boys in terms of time investments. This paper contributes to the literature by analyzing differential changes by gender in the level of parental time investments around transitions in household composition.

I use within-child variation in parental time investments to estimate gender gaps in investments and investigate their importance as a potential mechanism for explaining gender differences in outcomes related to household composition. Using the Panel Study of Income Dynamics (PSID) Panel Study of Income Dynamics [2014] and the accompanying Child Development Supplement (CDS), I obtain direct measures of parental time investments and explore how investment levels relate to household composition. I show that, while investments decrease for both boys and girls after transitioning to single-mother homes, the decrease is relatively large for boys. Differences are strongest through paternal weekday investments, for which boys lose an additional 24 minutes per day, which is about 35% of the average weekday investment from fathers during the first wave of the CDS. I also estimate large additional decreases for boys through the father's weekend investments, although that estimate is generally not statistically significant and the magnitude is more sensitive to specification. Combining the weekday and weekend data, I estimate that paternal investments decrease by an additional 2.3 hours per week for boys in single-mother homes, which is over 20% of the average weekly paternal investment during the first wave of the CDS. The investment gap is larger during adolescence, during which boys in single-mother homes lose over 3.3 hours per week more than girls. Furthermore, there is no strong evidence that mothers compensate for the additional loss by increasing investments to boys relative to girls.

These findings constitute important contributions to the literature by proposing another mechanism through which differences in the generation of non-cognitive skills could operate, and showing that the changes in investments are such that their importance in explaining differences in non-cognitive outcomes is plausible. Furthermore, by focusing on children who underwent changes in household structure, the findings do not rely on comparisons of investments across individuals, but rather on comparisons of changes in investments across individuals. Lastly, a key feature of this paper is the use of a direct measure of parental time investments, calculated from twenty-four hour time diaries collected as part of the Child Development Supplement to the Panel Study of Income

Dynamics. The use of a direct measure of parental time allows for transparency and clear interpretation of the results.

A brief overview of the current literature is provided in the next subsection. Section 2 describes and summarizes the data that are used. In section 3 I discuss the estimation procedure. Section 4 outlines results, and section 5 concludes.

2.1.1 Literature

Related work documents gender differences in the response to growing up in a single-parent household, and the effect on development of non-cognitive skills Bertrand and Pan [2013]. Bertrand and Pan document a gender gap in non-cognitive behavior that widens with age. They find that behavior is more sensitive to family structure and parental inputs for boys, but find no systematic differences in the home environment or investments that could explain much of the gap. They use measures of inputs such as the HOME index[1], Warmth index[2], and whether the child was spanked last week, all of which are correlated with family structure. A closer examination of the relationship between time investments and family structure might reveal gender differences in inputs that could help explain both the apparent differences in returns to these inputs, as well as the gap in non-cognitive skills.

In other work Jacob (2002) shows that differences in non-cognitive skills could explain a significant portion of the gap in college attendance. Other research shows that more educated and higher income parents invest more time into their children Guryan et al. [2008], that only time inputs from parents with a high level of education have a positive impact on cognitive achievement Del Boca and Mancini [2013], and that these investments have a larger effect earlier in life Del Boca et al. [2012]. Heckman and Cunha (2008) find that parental time investments are better at increasing non-cognitive than cognitive skills in general, but also find that parental time investments at younger ages have a greater effect on cognitive skills than investments made at later ages. They add that

[1]Based on parent responses to six questions about the activities that the child participates in and activities the parent participates in with the child. All of the questions were asked during the child's kindergarten year. See Bertrand and Pan, 2013.

[2]Also referred to as emotional supportiveness. Based on parent responses to a series of statements about their child, e.g. "child and I often have warm, close times together" and "being a parent is harder than I thought it would be." Responses were given in the Spring of the child's kindergarten year.

non-cognitive skills are more malleable at later stages in child development (i.e. adolescence) than cognitive skills, and that non-cognitive skills appear to promote the generation of cognitive skills Heckman and Mosso [2014].

One limitation of this research is that it only considers one dimension of a complex relationship, and no single measure of parental input captures all relevant aspects of the production process. Parental time investments are not necessarily comparable with, or preferred to other input measures. However, time investments provide a direct and clearly interpretable measure of parental investment. Bertrand and Pan (2013) focus on the HOME index, Warmth index, and whether the child was spanked last week to measure parental inputs. They find that these measures are related to whether or not the child is in a two-parent home, but that behavior of boys is much more responsive to inputs. It might be the case that other measures of parental inputs actually change differentially for boys and girls, and act as a mechanism explaining differences in behavior, which could also contribute to apparent differential gender responses to inputs.

2.2 Data

To estimate the differential changes in time investments around changes in household structure I use data from the Panel Study of Income Dynamics (PSID) and the Child Development Supplement (CDS) to the PSID. The CDS is a survey that was administered to children of PSID families in three waves (1997, 2002/2003, and 2007), and includes time diaries, and surveys of the children and their parents. The most critical component of the CDS for the purpose of this paper is the collection of twenty-four hour time diaries that catalog the activities of each child for one weekday and one weekend day. The diary data are at the activity level and include information on the duration and participants for each activity. I use the time diaries to construct measures of parental investments by counting time that the child spent with each parent in each diary. Every child in the CDS was assigned one randomly selected weekend day and weekday to record their activities. The first wave of the CDS includes children under age 13, and they are eligible for the CDS until they turn 18.[3]

[3]The age limits refer to the child's age during an initial screening. There are a small number of cases for which the child's age was outside of these limits at the point that the time diary data was recorded.

About 2,900 participants completed at least one time diary for CDS-I. More than 2,500 completed at least one for CDS-II, and over 1,400 for CDS-III. These add up to a total of 6,915 child-year observations. A total of 3,330 children completed at least one time diary in any period, and 1,086 completed at least one in all three waves. More than 1,400 completed at least one for exactly two of the waves.

There are two features of the data critical for the following analysis. The first is the presence of the time diaries used to calculate parental investments. Investments are calculated by summing time spent with mother/father[4] across activity-level data for each child. This is done separately for each weekend and weekday diary. In addition, I construct weekly investment measures to help summarize total investments by summing the weekday investment multiplied by five with the weekend investment multiplied by two. Second, I use information from the CDS and PSID surveys to construct variables describing the composition of each child's household, including presence of the child's biological/adoptive mother and father. I focus on comparing time investments for children in two-parent and single-mother households.

Figures C.1 through C.4 display cross-sectional differences in investments across gender and household type from wave I of the CDS. Figure C.1 graphs local polynomials of investments by age, for boys and girls who were in two-parent households during the first wave and those who were in single-mother households. Weekday time spent with mothers decreases dramatically as age increases across both genders and household types. The average time spent across these groups are within about thirty minutes of each other at every age in Figure C.1. However, at ages where the time spent differs, it is generally true that mothers spend more time with daughters than sons, and that mothers in two-parent households spend more time with their children. Figure C.2 displays the analogous estimates for maternal weekend investments by age. The overall levels of the investments are higher on weekend days, and gender gaps are also more pronounced. For example, in Figure C.1 the patterns for weekday maternal investments to girls are almost identical for those in two-parent and single-mother households, but in Figure C.2 single-mother's weekend investments to girls are

[4]The investment measures only include time spent with biological/adoptive parents. Similarly, when referring to parental presence in the household, I am referring to biological/adoptive parents only. For example, a child who lives in the same household as their biological/adoptive mother and a step-father is considered to be living in a single-mother household for the purposes of this study.

lower at every age than their two-parent household counterparts. The gap is roughly between thirty and sixty minutes at every age, which is a significant gap relative to that in Figure C.1 where the lines are almost indistinguishable at some ages. Similarly, there is a persistent gap between household types for maternal investments to boys across all ages, with boys in two-parent households receiving larger investments. Furthermore, the gender gap in mother's weekend investments appears to widen with age, with mothers spending more time with girls than with boys. Both Figures C.1 and C.2 demonstrate the importance of the child's age when considering time investments, as investments decrease sharply with age. For example, mothers invest between six and six and one-half hours on weekend days to their infant and toddler daughters, but that number is roughly four hours for twelve to thirteen year olds.

Figures C.3 and C.4 graph the analogous estimates for paternal investments from the weekday and weekend diaries, respectively. Fathers generally spend less time with their kids than mothers do across all household types and genders, but the decrease in investments with age is less drastic, especially for boys. In fact, Figure C.3 shows that weekday paternal investments are similarly low for both boys and girls in single-mother households across all ages. A slight increase in paternal weekend investments to boys in single-mother homes, shown in Figure C.4, leads to a small gap, in favor of boys, that appears to increase with age. The increasing gender gap in weekend paternal investments is more apparent in two-parent homes. From Figure C.4, both boys and girls in two-parent homes receive more than four hours in weekend paternal investments up until about age five, but a steady decline in weekend paternal investments for girls in two-parent homes leads to that number dropping below three hours around age twelve. However, weekend paternal investments for boys in two-parent homes remain steady at around four hours for all ages represented in the graph.

One implication of these investment patterns, particularly for boys in two-parent households, is that the proportion of total investments that come from fathers is increasing with age. This is shown more directly in Figures C.5 and C.6, which graph the proportion of the total parental investments[5] that come from each parent for those in two-parent households in the first wave for boys and girls,

[5]Total investments, $Total_i$, were calculated by weighting the weekday, WD_i, and weekend, WE_i, investments to construct a weekly investment, such that $Total_i = 5 \cdot WD_i + 2 \cdot WE_i$. The weekly measure of time spent with each parent divided by the total weekly measure is the proportion of total parental investment from that parent. The two values add to one by construction.

respectively. Notice that the proportions are roughly the same for infant and toddler boys and girls. Each group received a little less than seventy percent of total investments from their mother and a little over thirty percent from their father. The proportions change quite differently with age for boys and girls. For girls, the proportion of investments from their mothers never drops below about sixty-five percent. However, by about age twelve, boys receive under fifty-five percent of their investments from their mother. Another implication of the gender differences in the investments-age relationship is that one might expect the differential effect of household composition on parental time investments to differ by age. The increasing relative importance of paternal investments for boys, apparent in Figures C.4 and C.5, suggests that the potential for investment losses, relative to girls, increases with age.

Two important points of Figures C.1 - C.6 are that investments generally decline with age and that the relationship between investments and age differs by gender. These patterns could reflect the way parents spread their time with multiple children (Price, 2008) and the apparent preference of fathers to spend relatively more time with their sons (Baker and Milligan, 2013). In most cases, across all age groups and household structures, mothers spend a little more total time with girls, and fathers spend a little more with boys, on average. The figures demonstrate how important age is when evaluating time investments and suggests that using flexible controls for age is necessary in the analysis that follows.

Table D.1 summarizes the time investment variables and covariates by gender and household type. In particular, I separate out the individuals who underwent a change in household structure, because they are critical for estimation. Columns 3 and 4 display average characteristics for boys who underwent a change at some point. Column three includes boys who lived with both parents in the first period, meaning that the change in structure for them is going from living with both parents to living with less than both parents. On the other hand, column 4 includes individuals who underwent changes, but did not have both parents in the household in the first period. Columns 7 and 8 display the averages for girls who underwent a change in household composition at some point. The first row summarizes total weekly maternal investments, which was constructed by summing weekday investments (row 2) multiplied by five with weekend investments (row 3) multiplied by two. Girls

receive larger maternal investments than boys across all household types. Girls who were always in two-parent households received about 26.6 hours in maternal investments per week, relative to 24.8 hours per week for boys. The gap in maternal investments for children in two-parent households that eventually split, comparing column 6 with column 2, is about 3 hours per week, with girls receiving more investments. Both boys and girls in households that eventually split received larger maternal investments than those who were always in a two-parent household. From column 3, boys in families that eventually split received 25.5 hours per week on average, and those in households that never split received 24.8 hours per week. Similarly, girls in two-parent households that eventually had a change in composition received 27.4 hours per week in maternal investments, but those in two-parent households that never split receieved 26.6 hours per week. These comparisons may be misleading, because the household structure categories are also correlated with age. For example, girls in two parent households that never experience a change are just under seven years old on average, but the average age of girls in two-parent houesholds that eventually split is under five years. The difference is similar for boys. This, along with Figures C.1 - C.4, demonstrates why it is important to control flexibly for age when estimating gender gaps in the relationship between investments and household structure. Not only is age correlated with investments differentially by gender, it is also correlated with household structure. With that in mind, it is similarly true that boys in two-parent households that eventually split received higher paternal investments than those who were always in two-parent households, 16.7 and 16.3 hours per week, respectively. However, the opposite is true for girls, with girls who were always in a two-parent household receiving nearly two hours more per week in paternal investments, despite being roughly two years older on average.

From column 3, boys who experience a change in household composition but are in a two-parent household during wave I are about two years younger than boys who do not experience a change and are in a two-parent household at wave I, 4.6 years old and 6.5 years old, respectively. Boys in two-parent households who eventually see a change in household composition also have a little over one sibling in the household on average, whereas those in two-parent households who experience no change have about 1.3 siblings in the household. The differences are similar for girls. Girls who experience a change in household structure, but lived with both parents in the first period were

4.9 years old on average and had 1.1 siblings in the household at wave I, and those who are in a two-parent household and don't experience a change were about 6.8 years old with almost 1.3 siblings in the household. The racial composition of boys and girls in two-parent households that eventually split are also similar. In both cases, there are roughly equal percentages of black and white individuals in the subsamples, and the percentage of Hispanic individuals is relatively small. Lastly, the percentage of children in two-parent households in which their parents are married is about eighty-six percent for girls and nearly ninety percent for boys.

2.3 Estimation

The main contribution of this paper is the estimation of the gender gaps in time investments in single-mother households. To estimate the gender gaps, I use individual fixed effects regressions, including interactions between a dummy for being in a single-mother household with male and female dummy variables. The gender gap is the difference in the coefficients on the male and female interactions.

$$T_{it} = \alpha + \beta_M \cdot M_i \cdot MO_{it} + \beta_F \cdot F_i \cdot MO_{it} + \beta_3 \cdot Other_{it} + X_{it} \cdot \Gamma + c_i + \varepsilon_{it} \tag{2.1}$$

The left hand side variable in equation (2.1), T_{it}, represents some measure of parental time investments that child i received in wave t. For most specifications the investment measures are the amount of time that child i spent with his or her mother/father from the weekday/weekend twenty-four hour time diary, measured in hours. In the main specification, I report estimates for the weekday and weekend investments, as well as a total weekly investment constructed as a weighted sum of the weekday and weekend investments. I construct the total investment by summing the weekday investment multplied by five with the weekend investment multiplied by two. For ease of reporting and because using the weekly measure better reflects the effects on total investments, I focus on reporting estimates for total investments in alternate specifications that use the hours measures. I also include estimates based on equation (2.1) that replace the hours measures with dummy variables indicating whether the child had any positive investment from his or her mother/father. In that

specification, the outcome of interest is a dummy that equals one when the time investment is greater than zero for the given time diary.[6] For the corresponding total investment measure, the outcome is a dummy variable that is equal to one if either the weekday or weekend investment is positive.

The independent variables of interest are the interaction terms, where M_i and F_i represent male and female dummy variables, and MO_{it} represents a dummy variable indicating whether child i was in a single-mother household at wave t. There are two other types of household structures to consider. Living in the same household as both parents is the omitted category, and $Other_{it}$ indicates whether child i was in some other household type during wave t. The third category, $Other_{it}$, is constructed to make the categories mutually exclusive and comprehensive.[7] I focus on estimating gender gaps in investments for those living in single-mother households. Single-mother households are of particular interest, because children are more likely to live with their mother if the family is broken up. Furthermore, if fathers invest relatively more in boys as they get older, not having their father present in the household could hinder development for boys, even if not for girls. X_{it} represents a vector of time-varying observable characteristics including child's age and age-squared interacted with gender, the number of biological siblings in the household, dummy variables indicating the CDS wave, dummy variables indicating the presence of stepparents in or out of the household, and a dummy variable for marriage of parents in the household.[8] The gender specific age terms are important, because investments are closely related to age and differentially by gender.[9] Child level, time-constant characteristics are indicated by c_i, and ε_{it} indicates a period specific error term.

I estimate equation (2.1) using individual level fixed effects, so that $\hat{\beta}_M$ is a fixed effects estimator of investments that boys receive in single-mother households relative to those in two-parent households. Similarly, $\hat{\beta}_F$ is an estimator for investments that girls receive in single-mother households.

[6] For the weekday/weekend estimates, $T_{it} = 1$ when the weekday/weekend maternal/paternal time investment is great than zero, and $T_{it} = 0$ otherwise. For the total investment regressions, $T_{it} = 1$ when either the weekday or the weekend maternal/paternal investment is positive.

[7] $Other_{it}$ is equal to 1 if the child was in a single-father houeshold, or in a household with neither parent. This makes the estimated coefficient hard to interpret but these household structures are not the focus of this study.

[8] Estimates that also condition on the day of the week that the diary references are included in Tables H.1 and H.2 of the appendix. Results are not sensitive to this specification.

[9] In an alternate specification, I include a set of age dummy variables interacted with gender. The results are robust to this specfication.

The parameter of interest is the difference between the two investment levels, $\beta_{Diff} = \beta_M - \beta_F$. When $\hat{\beta}_{Diff} < 0$, that suggests that boys receive relatively low levels of investments in single-mother households, and $\hat{\beta}_{Diff} > 0$ suggests that boys in single-mother household are relatively well off in terms of time investments.

Because I estimate β_{Diff} using fixed effects, it is necessary to view some boys and some girls in a single-mother household during one wave and in a two-parent household in another. The procedure explained above does not restrict the direction of the change in household structure. Those who transition from a single-mother to two-parent household contribute to the estimates in the same way as those who go from a two-parent to a single-mother household. However, we might expect these two groups to be different. Omitted characteristics and behavior can directly influence transitions, as well as the level of time investments. The ages at which the child is in each household structure is also related to the direction of the transition. If the size of the gender gap differs with age, it could lead to estimating different gaps depending on the direction of the transition. I include a follow-up analysis, in which I split the sample by initial household type, and report separate estimates for the sample of those who lived with both parents in the first period and for those who did not. In addition, I estimate gender gaps by age and race to investigate whether the gap changes across those variables. I also decompose the gaps into specific activities to determine what types of activities are the main contributors to the differential investment losses.

Lastly, I investigate differential changes in parental ratings of non-cognitive behavior around the changes in household composition. Learning more about how parental time investments factor into the generation of non-cognitive skills is a primary concern. I examine differential changes by gender in externalizing behavior, internalizing behavior, and positive behavior[10] around changes in household structure. To do this, I estimate equation (2.1) using individual fixed effects and replacing the time investment measures with the behavioral measures. While these data are well-suited for

[10]Each of the three ratings is based on a series of questions asked to the child's primary caregiver. Externalizing and internalizing behavior questions ask how frequently the child exhibits some externalizing (i.e. acting out) or internalizing (i.e. inward negative behavior) and have three possible answers: not true, sometimes true, or often true. The data are coded so that higher scores mean that the child exhibits more problematic behavior. The positive behavior questions ask how "like" the child certain behaviors/characterstics are (e.g. cheerful, not impulsive), and are answered on a one to five scale, where one means the behavior/characteristic is "not at all like the child" and five means it is "totally like the child." A higher score on the positive behavior rating means that the child displays less problematic behavior.

measuring parental time investments, using parent-rated behavior measures may be problematic because parental perceptions of the child's behavior could change differentially by gender around changes in household structure.

2.4 Results

Before presenting the estimated gender gaps from the fixed effects regressions, I will briefly discuss OLS estimates of equation (2.1). Table D.2 shows estimates of the male and female interaction terms, as well as the estimated difference between the male and female interaction terms, for maternal and paternal total, weekday, and weekend investments. Estimates in Table D.2 are conditional on a set of covariates, including male and female interactions with age and age-squared. The estimated coefficient on the male interaction with single-mother household in the equation for total maternal investments from column 1, suggests that boys in single-mother households receive 1.24 hours per week more from their mothers, relative to children in two-parent households. Subtracting the coefficient on the female and single-mother household interaction gives the estimated gender difference in investments. Again from column 1, the estimated gender gap is 0.1 hours per week, suggesting that boys are relatively well off in terms of maternal investments. We can decompose the total difference by looking at the weekday and weekend gaps. The estimated weekday and weekend gaps in maternal investments have opposing signs. From column 2, the estimated gap is -0.049, meaning that boys in single-mother households received roughly three fewer minutes in weekday investments than girls. On the other hand, from column 3, boys in single-mother households received roughly eleven minutes more per weekend day than girls. In this case, the weekend gap outweighs the weekday gap, and the estimated total difference is positive. The estimates for paternal investments are shown in columns 4 through 6. From column 4 we see that boys in single-mother homes have relatively large decreases in investments, compared to girls, estimated at -7.5 and -6.5 hours per week, respectively, leading to an estimated gender gap of about one hour per week. The concern with estimating the gender gap by OLS is that it relies heavily on cross-sectional variation, but both family structure and gender are likely correlated with unobservable determinants of time investments. The fixed effects estimator

is preferable because it restricts the comparison to changes in investments within individuals who underwent a change in household structure with changes in investments for children who remain in two-parent households. The remainder of the reported estimates of equation (2.1) are fixed effects estimates.

Table D.3 reports fixed effects estimates of the gender gap in time investments based on equation (2.1) for maternal/paternal total, weekday, and weekend investments. The specification reported in panel A does not include any control variables, and the specification in panel B includes the full set of controls. Column four of panel A shows estimates for total weekly paternal investments without controls. The estimated gap in paternal investments is -1.92, meaning that paternal investments drop by nearly two hours more per week for boys in single-mother homes than they do for girls. After adding controls, while the estimated coefficients on the gender/single-mother interactions change, the estimated total gap in paternal investments remains similar. For example, the estimated coefficient on the male interaction with single-mother household goes from -10.3 without controls to -7.3 after adding controls. Similarly, the estimated coefficient on the female interaction goes from -8.4 without controls to -4.9 with controls. In both cases, the estimates suggest that paternal investments drop for boys and girls after going to single-mother households, but the decrease is relatively large for boys. From panel B, the estimated gender difference in total weekly paternal investments is -2.36, which is similar to the estimate from panel A, suggesting that paternal investments drop by nearly two and one-half hours more per week for boys in single-mother homes than they do for girls. That estimate is both economically significant, as the estimated gap is more than 20% of average paternal investments across gender and household types during wave I, and stastically important with a standard error of 1.08. The gender difference is strongest through weekday investments, for which the estimated gap, from column 5 of panel B, is -0.4 with a standard error of 0.15. That equates to roughly 24 minutes per weekday and is about 35% of average paternal weekday investments during wave I. To put the size of that estimate in perspective, consider Figure C.3 again, which graphs paternal weekday investments during the first wave of the CDS. Average paternal investments to boys and girls in two-parent households generally lie between one and two hours, depending on age. Of course, there is a slight downward trend, and the line drops below one and half hours by

about age seven for both boys and girls. The estimated difference of 24 minutes represents about twenty percent of the average paternal weekday investment in two-parent households at the lower end and forty percent at the higher end. Although the weekday gap drives about eighty percent of the estimated total weekly gap, the gap in paternal weekend investments is also negative, -0.22 hours, but only drives about twenty percent of the gap because of it's smaller magnitude and lower weight in the makeup of the total weekly investment measure.

The difference between the OLS and fixed effects estimates for total paternal time can be explained by examining how each individual contributes to each estimator. In the OLS estimator, those who transition from two-parent to single parent households, or vice versa, are in the omitted group (two-parent household) in one period and in a different group in another. With the fixed effects estimator, they are never in the omitted category. Girls who eventually go through a two-parent to single-mother transition have a lower baseline input than boys who make the same transition, because paternal investments to girls in two-parent households decrease as they get older. Since paternal investments are very low for both genders when their father is not in the household, the within child drop in paternal investments after going from a two-parent to single-mother household is relatively large for boys. From column 3 of Table D.1, boys in two-parent households that eventually split are age 4.6 and get 16.7 hours in weekly paternal investments. On the other hand, girls in the same household structure, from column 7, are of a similar age, 4.9 years old, but receive an average paternal investment of only 13.2 hours per week. When the average paternal investment decreases drastically to near zero for both genders after the transition to a single-mother household, there is more room for a decrease in paternal investments for boys. Another way to see this is to compare average investments for boys in two-parent households that never split with those who eventually split, 16.3 and 16.7, respectively, which are quite similar, despite the age difference in the two groups. On the other hand, girls who are always in a two-parent household receive about 15 hours per week in paternal investments, but those in a two-parent household that eventually splits only receive about 13.2. Finally, consider the similarity in the estimated change for boys, $\hat{\beta}_M$, - 7.5 hours when estimated by OLS and -7.3 when estimated by fixed effects. The difference between the fixed effects and OLS estimates of $\hat{\beta}_{Diff}$ is through a difference in the estimates on the coefficient

for girls, $\hat{\beta}_F$, which are -6.5 and -4.9 when estimated by OLS and fixed effects, respectively.

While the boy-girl differences in paternal investments are both economically and statistically important, it is possible that single-mothers compensate for the extra losses by increasing their investments to boys relative to girls. Columns 1 - 3 of Table D.3 show the estimated differences in total, weekday, and weekend maternal investments. From columns two and three of panel A, estimated weekday and weekend maternal investments are smaller for boys and girls in single-mother households. However, both weekday and weekend maternal investments are relatively low for boys, leading to estimated gaps of -0.19 for weekdays and -0.03 for weekends. From column 1 of panel A, the estimated total weekly gender difference in maternal investments is -1.4, meaning that boys suffer larger investment losses than girls of almost one and one-half hours per week. Adding the full set of control variables changes the estimated coefficients on the interaction terms, but does not lead to significant changes in the estimated gaps. Column 2 of panel B shows that the estimated gender gap in maternal weekday investments grows slightly in magnitude to -0.207. From column 3 of panel B, the estimated gap in weekend investments is now -0.003. The total gap in maternal investments, from column 1 of panel B, is about -1.1, meaning that boys in single-mother homes receive fewer maternal investments than girls in single-mother homes, and the magnitude is similar to the estimated -1.4 hours per week from panel A. The negative estimate on the gap in total maternal investments suggests that mothers do not compensate, but instead decrease their investments to boys relative to girls. However, the standard error for the estimated difference is relatively large, 1.6, so we should not draw any strong conclusions based on that difference.

The analysis in Table D.3 does not restrict the direction of the household transition, including children who go from a two-parent household to a different household type, as well as those who go from not living in a two-parent household to living with both parents. However, these two types of transitions and families could be quite different from each other. One reason that we might expect heterogeneity based on the direction of the transition is that the age that the child is in each household structure is correlated with the direction of the transition. Differential gaps by age could lead to different estimates when splitting the sample by the direction of the transition. Table D.4 shows separate estimates by initial household structure. Columns 1 and 2 show estimates for

total weekly investments for families that were intact in the first wave, and columns 3 and 4 show estimates for children with less than two parents in the household in the first period. From column 2, the estimated gender difference in total paternal investments for those who were in a two-parent household in the first wave is -4.2 hours per week with a standard error of 1.54. Restricting to the subsample of individuals who were in two-parent households in wave one increases the magnitude of the estimated gap in paternal investments by almost eighty percent. However, the estimated gap in maternal investments becomes positive, 0.688, after restricting to those in two-parent households in the first wave. The positive sign on total maternal investments suggests that mothers might compensate for the extra losses that boys suffer in paternal investments, but the estimate is noisy, and smaller in magnitude than the loss in paternal investments. If we take these estimates at face value, then boys lose an additional 4.2 hours per week in paternal investments, about seventy percent of wave one average paternal investments, and mothers partially compensate for the loss with an additional forty minutes per week. While the increase partially offsets the loss, about seventeen percent of it, boys still suffer substantially larger investment losses from transitioning to a single-mother home.

Columns 3 and 4 of Table D.4 display estimates when restricting to the subsample of individuals who were not in a two-parent home in the first wave. There are fewer movers in this direction and the standard errors are relatively large, but the estimates suggest that the direction of the move likely matters. For example, the estimated gap in maternal investments in the subsample who were not in two-parent homes in wave one is -3.5 hours per week. This estimate is of the opposite sign from the subsample of children who were in two-parent homes in the first wave, and relatively large in magnitude. In other words, this subsample is driving the negative estimates on total maternal investments in the full sample. The estimated gap in paternal investments is still negative, -.18, but much smaller in magnitude. Although there appears to be some heterogeneity based on the direction of the transition, the estimates in the sample of children not in two-parent households in the first wave are noisier because they are based on a relatively small number of transitions.

In addition to estimating differences in parental time investments, I include supplemental estimates on the probability of having a positive investment. In this specification, I replace the number

of hours spent with the mother/father with a dummy variable for having any investment, and estimate the differences using linear fixed effects regressions. From column 1 of Table D.5, the estimated change in the probability of receiving a nonzero maternal investment after transitioning to a single-mother household is 0.040 for boys, and 0.068 for girls, leading to an estimated gender gap of -0.028 in favor of girls. In other words, after the transition, girls see a relatively large bump in their probability of having any investment from their mothers. However, the estimate is noisy with a standard error of 0.033. Interestingly, the sign on the weekday and weekend maternal investments estimates are different, with the estimated gap in receiving any weekday investment favoring girls, and the estimated gap in weekend investments favoring boys. For paternal investments, the estimated gap in the probability of receiving any weekday investment is -0.08, meaning that they receive about an 8 percentage point more drastic change in the probability of receving a positive investment from their fathers on weekdays. The analogous weekend estimate is -0.047. However, the estimated gap in the total weekly probability is much smaller in absolute value than both the weekday and weekend estimates, only -0.01. That suggests that boys see a relatively large decrease in the probability of receiving an investment on any given day, but the change in their probability of receiving an investment at some point throughout the week is similar.

Splitting the sample by initial household composition reveals the difference in probability of receiving a positive paternal investment between those who were and were not living with both parents in the first wave. From Panel A of Table D.6, those who lived with both parents in the first wave see a relatively large decrease in the probability of receiving a positive paternal investment throughout the week, with an estimated difference of -0.065. The estimated gap when looking at weekday or weekend investments only is -0.11 in both cases, suggesting that on any given weekday or weekend day boys have an eleven percentage point larger decrease in the probability of receiving a positive paternal investment. On the other hand, the estimated gender gaps in paternal investments are positive for total, weekday, and weekend investments for the subsample that lived with less than both parents in wave I, meaning that boys in single-mother households were relatively well off by this measure. However, the estimates are noisy in that subsample with an estimated weekly gap of 0.094 and a standard error of 0.089.

2.4.1 Heterogeneity in Gender Gaps

Next, I consider potential heterogeneous effects by age and race. Because of the strong correlation between investments and age, and differential age trends by gender and household structure, one might expect that the boy-girl investment gap in single-mother households depends on the age of the child. For example, Figures C.5 and C.6 show how the proportion of total parental investments that come from the father increase with age for boys in two-parents households, which suggests that paternal investments become increasingly important for boys as they get older. Figures C.3 and C.4 provide insight to that trend by showing that gaps arise in boy and girl investments with age, with fathers spending relatively more time with boys. Furthermore, the steep decline in maternal investments from Figures C.1 and C.2 means that paternal investments become increasingly important in the makeup of total investments for boys, but less so for girls. To compare the gender gaps across ages, I group the individuals into four age bins (0 - 5, 6 - 10, 11 - 15, and 16 and over) and estimate the gap for each bin. Panel A of Table D.7 shows the estimates for total maternal investments by age bin. The smallest gap, -0.6 hours per week, is in the 6 - 10 year old bin. From there, the gap size increases to about -1 hour per week in the 11 - 15 year old bin, and then to -2.2 hours per week in the 16 and over bin. There is some suggestive evidence that relative losses for boys increase as they get older, but the estimates are noisy, all with standard errors of 1.8 or higher.

Panel B of Table D.7 shows estimates for paternal investments by age bin. A more clear pattern emerges in the paternal investment gaps, with boys suffering greater investment losses with age. The smallest estimate is -0.4 hours per week for ages 0 - 5, but they increase in magnitude with each age bin. The rest of the estimated gaps are -1.8 hours per week for 6 - 10 year olds, -3.3 for the 11 - 15 year olds, and -3.6 for the 16 and over age group with p-values of 0.12, 0.006 and 0.02, respectively. The pattern in these estimated differences supports the idea that paternal investments become increasingly important for boys as they get older, leading to relatively large investment losses during adolescence.

I also estimate differential gaps by race, focusing on the white and black subsamples, as those two races make up more than ninety percent of the sample of children who made a transition from a two-parent to less than two-parent household. From Table D.8, the estimated gender gap in total

maternal investments for white children in single-mother households is -3.1 hours per week with a standard error of 2.5. The estimated differences for black children in single-mother households is -1.4 hours per week with a standard error of about 2.3. Both of the point estimates are larger in magnitude than the overall estimate, but the estimates are noisy and not statistically different from zero at any standard significance level. The estimated difference in paternal investments for white children in single-mother households is about -4 hours per week, and is strongly statistically significant with a standard error of 1.6. That estimate is similar in magnitude to the overall estimated gap for the subsample of children who were in two-parent households in the first wave. The estimate for the black subsample is about half the size, -2 hours per week, which is similar to the estimated gap across the full sample, and is less statistically significant, with a standard error of 1.4. While all four of the estimated gaps are negative, the differences provide some evidence that the extra losses suffered by boys who are white are larger in terms of both maternal and paternal investments. Parental investments are also larger on average across the sample of white children, which could drive part of the difference. For example, the average weekly maternal investment for white children in two-parent households is about 3 hours more per week than that for black children in two-parent households (21.5 hours per week to 18.2 hours per week), and the difference in weekly paternal investments is similar (14.5 hours per week to 10.9 hours per week).

2.4.2 Composition of the Investment Gaps

The time diary data include descriptions of the activities that children participate in, making it possible to decompose the total gap by activity. Table D.9 displays fixed effects estimates of equation (2.1) by activity category.[11] Column 1 provides estimates on total maternal investments. Estimated gaps in maternal investments are positive on passive leisure (0.22 hours per week), e.g. watching television, tending to needs (0.1 hours per week), and childcare (0.18 hours per week), meaning that

[11]This is not a comprehensive set of the activities, but they account for the majority of the differences. The estimates for the gaps in paternal investments displayed in Table D.9 add up to a litte more than the total difference, -2.5 hours per week compared to -2.3, meaning that the omitted activities add up to a slightly positive number. The displayed estimates for maternal investments add up to about -0.7, less than total gap of -1.1, so the gaps in the omitted activities sum to about -0.4.

mothers increase time with boys relative to girls in those activities after transitioning to a single-mother household. On the other hand, maternal investments of active leisure decrease substantially more for boys in single-mother households. The estimate of -0.85 hours per week in active leisure is the largest in magnitude of any activity. Boys also experience relatively large reductions in paternal investments in leisure, -0.72 and -0.75 hours per week in passive and active leisure, respectively. In general, boys see a much larger dropoff in leisure activity with their parents than girls do after transitioning to a single-mother household. The rest of the estimates for paternal investments displayed in Table D.9 are also negative, but generally of smaller magnitude than the estimates for leisure activity. The next two largest are the differences in tending to needs, -0.48 hours per week, and entertainment, -0.21 hours per week.

2.4.3 Household Structure and Child Behavior

One reason for analyzing gender gaps in the relationship between household structure and parental investments is to assess the possibility that changes in investment levels contribute to non-cognitive skill forumlation, explaining some of the differences we see in outcomes between boys and girls, e.g. behavior and educational attainment. If the documented gaps in time investment losses from transitioning to single-mother households contribute to contemporaneous behavioral issues, then we might expect to see changes in behavior that match the differential investment losses. In other words, if the additional investment decreases that boys suffer is bad for their behavior, then we would expect that to show up in measures of their behavior. I estimate the gender differential changes in behavior from transitioning to single-mother households by replacing the outcome variables in equation (2.1), parental time investments, with measures of child behavior. One major weekness of this exercise is that the behavioral measures, taken from the CDS, are parent-reported. This is problematic because parental perceptions of their child's behavior could change differentially by gender around household structure transitions. Nonetheless, I show results from the analysis to demonstrate the concept.

Table D.10 shows results from fixed effects estimation of equation (2.1), replacing the parental investment outcomes with three parent-rated behavior measures: externalizing behavior, internaliz-

ing behavior, and positive behavior. Each of the ratings are the result of a sequence of questions that the parent answers about the child. There are some minor differences in the questions across the three waves, so I use the raw responses to standardize the set of questions contributing to each behavioral measure. Each of the questions related to externalizing/internalizing behavior are answered on a one through three scale, and the questions for positive behavior are on a one through five scale. I summed the answers and standardized the total scores within each wave. Higher scores in externalizing and internalizing behavior represent more problematic behavior, but a higher score on the positive behaviors measure represents less problematic behavior. Interestingly, The estimated coefficient on the male interaction with single-mother household in the equation for externalizing behavior is negative, -0.11, meaning that parents rate their boys as displaying less externalizing behavior following the transition. On the other hand, the coefficient on the female interaction term is positive, 0.16, meaning that parents state that their girls are acting out more following the transition. The externalizing behavior ratings for boys become much more favorable following the transition to a single-mother household, relative to girls, and the estimated gap in changes in externalizing behavior, from column 1, is -0.27 standard deviations. The result is also statistically significant, with a p-value of 0.02. Apparently, the change in household structure is not leading to more externalizing behavior for boys, according to the parent-rated measure, like we might have expected based on the additional investment losses. The estimated gaps in the other two measures are much smaller in magnitude, -0.007 for internalizing behavior and 0.11 for positive behavior. The estimated gap in externalizing behavior is surprising, but this gap should be estimated using measures that are not based on parent ratings in future work.

2.5 Conclusion

Gender gaps in non-cognitive skills among adolescents have been documented in recent literature, but determining the mechanisms that lead to these gaps is still largely an open question Bertrand and Pan [2013], Jacob [2002]. These gaps could arise for several reasons, including gender differences in returns to and levels of inputs in single-mother and two-parent households. Bertrand and

Pan identified differences in returns to inputs and family structure as major contributing factors in explaining this gap, but suggested that differences in quantity of inputs plays a much smaller role. Because inputs can come in many forms, most of which are correlated with household structure and difficult to measure and interpret, it is difficult to disentagle mechanisms leading to non-cognitive differences. This paper uses data from the Panel Study of Income Dynamics and the Child Development Supplement survey to obtain direct measures of parental time investments and consider gender differences in investments around changes in household composition. One advantage to this approach is that time investments are relatively straightforward in that they are simply measures of the amount of time parents spend with their children, based on twenty-four hour time diaries collected in the CDS.

Finding differential investment losses by gender suggests that living in a single-mother household could have a large negative impact on the quantity of investments that boys receive relative to girls, and in turn could contribute to the gender gap in non-cognitive skills, specifically for those in single-mother households. Using child-level fixed effects, I estimate the differences in time investments that arise from living in a single-mother household. Although both boys and girls see significant reductions in parental investments, boys experience larger decreases. The extra loss in total weekly paternal investments amounts to about 2.3 hours per week, and is strongest through weekday investments, 24 minutes per weekday, which account for roughly eighty percent of the total loss. These findings are economically significant, equating to more than twenty percent of the average paternal investments in the first wave of the data. There is no strong evidence that mothers compensate for the extra loss by increasing investments to boys following a change in household structure, relative to girls. I also find that the additional losses for boys are generally increasing with age, with the estimated gap in total paternal investments over 3.3 hours per week for boys during adolescence. Decomposing the gap in paternal investments shows that the additional losses for boys are largest in leisure activities, which account for about sixty percent of the difference.

All of this evidence considered together with existing research suggests that time investments are another potential mechanism that could help explain the non-cognitive skill gender gap. However, using parent-reported ratings of externalizing and positive behavior from the CDS, I find that parents

rate their girls as having higher levels of externalizing behavior and lower levels of positive behavior when in single-mother households, relative to boys. Future research should consider a more direct link between time investments and outcomes, focusing on measures of non-cognitive skills that are not based on parent ratings.

Chapter 3

Precision for Policy: Calculating Standard Errors in Value-Added Models

3.1 Introduction

Early studies on value-added (VA) focused simply on the possibility of measuring teacher effectiveness Sanders et al. [1997], Raudenbush [2004], Rockoff [2004], Rubin et al. [2004], Rivkin et al. [2005], Aaronson et al. [2007]. These paved the way to several studies investigating the potential for bias in commonly used value-added estimators (e.g. Ballou et al. [2004], McCaffrey et al. [2004], Kane and Staiger [2008], Ballou [2009], Briggs and Weeks [2009], Harris [2009], Ishii and Rivkin [2009], Koedel and Betts [2010], Lockwood and McCaffrey [2009], Reardon and Raudenbush [2009], Goldhaber and Hansen [2013], Rothstein [2008], Sass et al. [2014], Corcoran et al. [2011], Koedel and Betts [2011], Condie et al. [2014],Ballou and Cavalluzzo [2012], Goldhaber and Chaplin [2015], Goldhaber et al. [2014], Kane et al. [2013], Chetty et al. [2014], Guarino et al. [2015]. While much attention has been given to obtaining the value-added estimates, and assessing potential bias in them, there has been less focus on obtaining the correct confidence intervals for the value-added estimates.

Precision of the estimates should be considered when using them in practice, especially in high stakes settings. For instance, if teacher value-added estimates are used to identify the bottom five percent of teachers for sanctions or loss of employment, how confident are we that the teachers in our estimated 6th percentile are actually more effective than those in the 4th percentile? Even with unbiased estimates, we need the correct confidence intervals to address this question and to understand whether teachers in the tails of the effectiveness distribution—which is often the target for policy—are significantly different from those in the middle of the distribution.

Such inference questions have been considered in a limited number of studies. Raudenbush (2009)

discusses inference specifically regarding both the adaptive centering random effects approach proposed in the paper and comparable fixed effects approaches. Two other studies have addressed imprecision as it relates to distinction from the mean teacher effect as well as the ability of estimators to distinguish between any particular ranks (McCaffrey et al. [2004], Lockwood et al. [2002], respectively). Using simulations, Lockwood, Louis, and McCaffrey (2002) use mean squared error as a metric, showing that distinguishing between ranks can be difficult, especially in the tails of the distribution of teacher quality.

Our paper fills a gap in the prior research by comparing different methods for computing standard errors and confidence intervals for the OLS-Lag estimator. Using simulated data on students and teachers, we examine the sensitivity of the standard errors on the teacher value-added estimates to the method of calculation under several student grouping and classroom assignment scenarios. We study how different standard errors behave by comparing the variability of estimates from the simulated data with the size of the computed standard errors. We also evaluate how well confidence intervals calculated using traditional methods perform by counting how frequently the implied confidence intervals around the estimated teacher effect include the true teacher effect.

Our simulations show that relying on standard errors and confidence intervals calculated using traditional methods could be misleading when estimating teacher fixed effects using OLS-Lag. In particular, when students are grouped on unobservable characteristics, even under random assignment of classrooms, traditional standard errors, such as those obtained using a standard cluster robust variance estimator, tend to underreport the estimator variability when value-added estimates are based on a small number of student cohorts per teacher. Consequently, the corresponding confidence intervals over-reject the true teacher effects. Furthermore, the tendency to over-reject the true teacher effects is worse with fewer cohorts because regularly reported confidence intervals use the standard normal distribution to obtain critical values, and this turns out to be a poor approximation with small numbers of cohorts. We propose another method for obtaining confidence intervals that both allows for unrestricted correlation of within classroom errors and uses appropriate critical values from t-distributions. This method works well in the simulations, producing confidence intervals with nearly perfect coverage in all cases with unbiased estimators.

To expand on the simulation evidence, we analyze administrative data from six districts in a large state. For each district, we compare standard errors and confidence intervals from our proposed method with those from the regularly used methods. In two districts, we find that the standard errors computed with our proposed cohort-by-cohort method are systematically larger than those from the other methods, suggesting potential over-rejection from using typical methods. More importantly, we observe consistent divergence in the computed confidence intervals. On average, we find that the confidence intervals from the proposed method are much wider with a small number of cohorts, making it difficult to distinguish any quantile of the teacher effectiveness distribution from another. Consistent with our theoretical predictions and simulation results, average widths of the confidence intervals converge with increasing cohorts, and are similar to traditionally estimated confidence intervals on average with seven cohorts (the maximum in our data). Finally, we illustrate the relevance of confidence intervals in policy settings by comparing how well the different confidence intervals allow us to separate teachers from different points in the VA distribution.

The paper proceeds as follows. The next section provides a brief discussion of the issues that arise when conducting inference on OLS-Lag teacher value-added estimates. The third section describes our simulation design and analysis and presents simulation results. Section 4 describes the administrative data that we use, the methods used to analyze the administrative data, and presents results of that analysis. Section 5 concludes.

3.2 Discussion

The nested nature of student level administrative data and resulting potential for error correlation lead to ambiguity in how to conduct inference on teacher value-added estimates obtained from OLS regressions. There are two issues that arise in conducting inference on teacher value-added estimates as a result of error correlation: obtaining standard errors and using proper critical values. There are several options for calculating robust standard errors to address precision. For instance, researchers might wish to allow error correlation within student, teacher, or school by using a cluster-robust variance estimator. Each of these choices can give different results, and knowing which is correct is

not easy. The second issue, obtaining proper critical values, is particularly relevant when estimating value-added from a small number of classrooms per teacher. If a teacher fixed effect is based on a small number of cohorts, then there is what amounts to a small sample issue Donald and Lang [2007].

When constructing confidence intervals for value-added estimates, we would like to relax the assumptions about error independence to allow for unrestricted correlation within classrooms. There are several reasons one might expect error dependence within a teacher's class, but we focus on the case of nonrandom grouping of students into classrooms. Even if classrooms are randomly assigned to teachers and the estimators are unbiased, grouping students on unobservable characteristics complicates inference by increasing residual variance across classrooms and reducing the amount of independent information available to estimate each teacher fixed effect. In theory, clustering at the level of independence produces consistent standard errors by separating sets of dependent errors into independent pieces of information. For example, if error terms are independent across classrooms, then clustering at the classroom or cohort-school level will result in consistent estimators. However, the asymptotic justification for consistency in this case works as the number of cohorts per teacher increases, which is less useful considering that teacher effectiveness is typically estimated using only a few years of data.

As in the broader grouped data literature, constructing confidence intervals thus is complicated when value-added is estimated from a small number of classrooms per teacher.[1] One key feature of teacher value-added separates it from many other settings — in the value-added context we are estimating a potentially large number of treatments, each applying to a relatively small number of individuals. Therefore, adding more teachers (or schools) is effectively adding new treatments rather than accumulating information on each treatment. This contrasts oft-studied scenarios that involve aggregate treatment cases, where a small number of treatments apply to a large number of individuals. The issue in the value-added case is that using a variance estimator that allows for arbitrary error dependence within group (e.g., within classroom) does not necessarily lead to

[1]Our discussion is aimed at the elementary school setting, where each teacher has one classroom per year. How this carries over to the middle or high school setting, where teachers have several classrooms per year, depends on the independence assumption that the researcher is willing to make.

correct confidence intervals when the number of classrooms per teacher is small. Confidence interval performance could be hindered through both the variance estimates and critical values when the number of classrooms per teacher is small. This is problematic if school districts want to make decisions based on teacher performance over a short period of time, which is often the case.

Our simulations, which we describe in detail in the next section, investigate the performance of standard errors and confidence intervals obtained from several different methods. We generate student level data, estimate teacher value-added using OLS-Lag, and construct multiple confidence intervals around the teacher value-added estimates under several scenarios of student grouping and classroom assignment to teachers. We evaluate the quality of the standard errors by comparing average standard errors with standard deviations of estimated teacher fixed effects (i.e., teacher value-added). We evaluate the quality of the confidence intervals by calculating coverage rates of 95% confidence intervals implied by each method.[2] We use administrative data to check whether patterns found in the simulations carry out in the data, and evaluate the differences in the methods for constructing confidence intervals by examining the ability of the different confidence intervals to distinguish teachers from different percentiles of the value-added distribution.

We compare the performance across five different standard error and confidence interval calculations. The first is the naïve calculation with no correction for heteroskedasticity or within group error correlation. We also calculate three sets of standard errors from a typical cluster robust variance estimator and use standard normal critical values to construct confidence intervals. The clustering in these cases is at the classroom, cohort-school, and school level. Finally, we propose another method that involves calculating a teacher value-added estimate for each class for each teacher using OLS-Lag regressions for each year. Then we use the average of the preliminary estimates as the teacher value-added estimate and construct confidence interval, teacher by teacher, using critical values from the t-distribution with degrees of freedom equal to the number of classrooms (for that teacher) minus one. There are two advantages of the proposed method; it effectively divides the error terms into independent pieces of information (similar to classroom-level clustering) and correctly adjusts the critical values according to the number of independent pieces of information, i.e.,

[2]The coverage rate, discussed in more detail below, refers to the percent of the 500 simulation replications for which the 95% confidence interval of the estimated teacher effect contains the true effect.

the number of classrooms the teacher has taught.

We also include results from school level clustering. Theoretically, though, clustering at the school level is problematic because increasing the number of schools does not increase the amount of information per teacher, but rather increases the number of parameters to estimate. In other words, clustering at the school level assumes fewer independent pieces of information are available than parameters to estimate.

3.3 Simulations

3.3.1 Simulation Design

We start by simulating data on students and teachers, using the following test score data generating process (DGP).

$$A_{i,t} = \lambda A_{i,t-1} + \beta_{i,t} + c_i + cs_i + \varepsilon_{i,t} \tag{3.1}$$

where $A_{i,t}$ and $A_{i,t-1}$ represent the current and lagged test score. All observations are from the same grade, so we observe each student only one time. The teacher effects are represented by $\beta_{i,t}$. An individual student fixed effect, cohort-school fixed effect, and random errors are represented by c_i, cs_i and $\varepsilon_{i,t}$, respectively.

Table E.1 summarizes our simulation parameter choices. Each component of the current test score is generated from a normal distribution. The standard deviations of the baseline score (i.e., the lagged score), teacher fixed effect, and random error are fixed at 0.25, 0.25, and 1, respectively.[3] In the first case presented, there is no cohort-school fixed effect, and the student fixed effects are drawn from a standard normal distribution.[4] In the second case, both the student fixed effect and the cohort-school fixed effect are drawn from a normal distribution with standard deviation of 0.25. The teacher fixed effects are drawn from a normal distribution with a mean of 0.5.[5]

[3]The percent of total variance from teacher fixed effects based on our choices is around the lower end of that considered in Guarino et al. [2015], and the percent from student fixed effects is larger than what they consider in the first case and lower in the second case.

[4]We have done the same analysis with the standard deviation of the student fixed effect equal to 0.5. This changes some results quantitatively, but the general patterns are the same.

[5]Based on these values, the variance is a little over 1 at the least, and over 2 at the most, depending on the

Simulations are restricted to a single grade (with current and lagged student test scores) to simplify the process, and are designed to reflect the elementary school setting. Students do not repeat grades so we have exactly one observation per student. The main results are based on simulated data including five schools, with two teachers per school, and 20 students per class. We focus on a small number of schools and teachers because there is no gain from adding schools to the simulations—recall, adding schools only adds new teachers, or treatments. Information useful for estimating teacher effects accumulates with the number of students the teacher has taught, so more students per class or more cohorts (years) of students, but not with an increase in the number of schools. For this reason, we focus more on how t-statistics change with increasing the number of student cohorts than on how they change as more teachers and schools are added.[6]

We consider three grouping scenarios reflecting how students might be grouped into classrooms within each cohort at each school. These are random grouping (RG), dynamic grouping (DG), and heterogeneity grouping (HG). Dynamic grouping is based on students' lagged test scores, $A_{i,t-1}$, which places high achieving students together (and similarly for low achieving, etc.) to simulate tracking or ability grouping.[7] Heterogeneity grouping (HG) is based on an unobservable student fixed effect, c_i, which might represent an ability level or behavioral profile that is observable to, say, a principal, but is not appropriately captured in prior test scores. We focus on random assignment (RA) of classrooms to teachers, but also consider positive assignment (PA)[8], in which case the classrooms with the higher average $A_{i,t-1}$ for dynamic grouping, or higher average c_i for heterogeneity grouping, are assigned to the teacher with the higher fixed effect. In every scenario we consider, students are grouped into two classrooms based on one of the three possible grouping mechanisms, RG, DG, or HG, and then classrooms are assigned to the two teachers in the school randomly (RA) or positively (PA).[9] In each case presented here there is no correlation between the lagged score and student fixed effect, $corr(A_{i,t-1,}c_i) = 0$.[10]

distributions used for the student and cohort-school fixed effects.

[6]We include results with 50 schools in one case to show that the results do not change.

[7]Grouping on $A_{i,t}$ and c_i is imperfect in the sense that it is not the case that the top half of the distribution is placed in one classroom and the bottom half in the other.

[8]We consider positive assignment to demonstrate what happens with a biased estimator.

[9]Guarino et al. [2015] provide a more thorough discussion of these and other possible grouping mechanisms.

[10]We have done alternate analyses using $corr(A_{i,t}, c_i) = 0.5$, and results are similar.

The two nonrandom grouping mechanisms have different implications for the value-added estimates themselves, as one mechanism involves grouping based on an included control variable ($A_{i,t-1}$), while the other is based on an omitted variable (c_i). Grouping on an omitted variable leads to higher error variance across classrooms, but it does not bias the teacher fixed effects (value-added) estimators if classrooms are randomly assigned to teachers. However, if classroom assignment is nonrandom, grouping based on the omitted student fixed effect does introduce bias Guarino et al. [2015]. In this case, we do not expect the confidence intervals computed using the biased value added estimates to exhibit good coverage rates, but comparing the variability of the estimates with the computed standard errors is still informative.

To summarize, we conduct a simulation for each of the particular grouping and assignment mechanisms under each of the distributional cases described in Table E.1, and for various numbers of cohorts of students. The simulation involves 500 repetitions for each of the particular combinations of grouping, assignment, distributional case, and cohort choices.

3.3.2 Analytic Methods

In each simulation, we estimate teacher effects along with four different types of standard errors: standard errors without any correction for clustering or heteroskedasticity, standard errors with classroom level clustering, cohort-school level clustering, and school level clustering.[11] To estimate the teacher effects, we regress the current test scores on teacher dummy variables using ordinary least squares (OLS), controlling for one lagged test score, which we term OLS-Lag.[12]

$$A_{i,t} = \lambda A_{i,t-1} + (TeacherDummies_{i,t})\beta_0 + X_{i,t}\gamma_0 + \varepsilon_{i,t} \tag{3.2}$$

Where $X_{i,t}$ represents a set of observable student characteristics. In the simulations this is empty, but when estimating value-added using administrative data we include a set of student

[11] Clustered standard errors are obtained using the typical cluster robust sandwich variance estimator.

[12] Guarino et al. [2015] found that the OLS-Lag method was the most robust method, that is, least subject to bias overall, among a set of estimators investigated. We therefore explore the issue of standard errors and confidence intervals using this estimator.

characteristics. OLS-Lag does not restrict the coefficient on the lagged test score (λ), allowing for less than complete persistence in students' test knowledge from one year to the next. For the sake of brevity we present simulation results only for the case where $\lambda = 0.5$, which is in line with the findings of Andrabi et al. [2011].[13]

Our first performance metric is a comparison between each of the four types of standard errors and the actual variability of the estimates. We calculate the actual variability of the estimates by taking the standard deviation of the estimates across the 500 repetitions, as in equation (3.3).

$$SD(\hat{\beta}^{p,q}) = \sqrt{\frac{1}{499}\sum_{r=1}^{500}(\hat{\beta}^{p,r,q} - \bar{\hat{\beta}}^{p,q})^2} \tag{3.3}$$

$\hat{\beta}^{p,r,q}$ represents the value-added estimate of teacher p, from the r^{th} replication, based on q cohorts of simulated data, and $\bar{\hat{\beta}}^{p,q}$ represents the average estimate across all 500 repetitions. We take $SD(\hat{\beta}^{p,q})$ as a measure of the true variability of the estimator. To summarize, for each of 500 simulation replications for a particular scenario, we estimate teacher value-added using OLS-lag, obtaining one value-added estimate for each teacher. Then we calculate the standard deviation of the estimates for each teacher. To assess the performance of the standard errors, we use the average of the particular type of standard error across the 500 repetitions, as shown in equation (3.4).

$$AvgSE_k^{p,q} = \frac{1}{500}\sum_{r=1}^{500} se_k(\hat{\beta}^{p,r,q}) \tag{3.4}$$

The average standard errors are obtained in the following steps for each grouping and assignment scenario: For each of 500 simulation replications, we obtain the four different standard errors for each value-added estimate. Let $se_k(\hat{\beta}^{p,r,q})$ stand for the corresponding standard error reported when the k^{th} type of clustering is used. That is, $k \epsilon \{no\ clustering,\ class\ clustering,\ cohort - school\ clustering,\ school\ clustering\}$. Then we calculate the average of each standard error for each teacher, as shown in equation (3.4). We report the average of these averages, taken across teachers, to summarize the overall results for a particular standard error estimate, and refer to that

[13] Evidence from previous literature suggests that the value of λ is less than one Jacob et al. [2010], Rothstein [2008], Andrabi et al. [2011]. We have also done simulations using $\lambda = 1$, and the results are not sensitive to this change.

as the "Average of the Teacher Average SE." We consider the comparison between the averages from equations (3.3) and (3.4) to be informative, because they tell us how closely the standard errors reflect the actual variability of the estimates.

We refer to the second indicator of performance that we use as the *coverage rate*. The coverage rate for each teacher is the percent of replications for which the 95% confidence interval of the estimated teacher effect contains the true effect. In short, this is computed by comparing computed t-statistics with the critical value. We first construct the t-statistic for the two-sided test with the null hypothesis that the estimated effect equals the known true effect for each replication for each teacher, as in equation (3.5).

$$t_k^{p,r,q} = \frac{\hat{\beta}^{p,r,q} - \beta_{true}^{p}}{se_k(\hat{\beta}^{p,r,q})} \tag{3.5}$$

Again, $\hat{\beta}^{p,r,q}$ represents the estimated teacher effect for the p^{th} teacher based on q cohorts, from the r^{th} simulation replication, and $se_k(\hat{\beta}^{p,r,q})$ is the corresponding standard error computed with clustering method k. Let β_{true}^{p} represent the true effect for teacher p, which does not change across replications. The null hypothesis can be written as $H_0 : \beta_0^p = \beta_{true}^p$. We calculate the absolute value of the t-statistic in equation (3.5) for each teacher, type of standard error, and replication. We then compare this t-statistic to the relevant critical value from the standard normal distribution (1.96 for a 95% confidence interval):

$$CoverageRate_k^{p,r,q} = (\frac{1}{500}) \sum_{r=1}^{500} I[abs(t_k^{p,r,q}) \leq 1.96] \tag{3.6}$$

The *coverage rate* for a given teacher and type of clustering is then calculated as the percent of t-statistics with absolute value less than 1.96, which is identical to forming a 95% confidence interval around each estimate and calculating the percentage of replications in which we fail to reject H_0. Equation (3.6) provides the expression for the coverage rate based on q cohorts, clustering type k, for teacher p, and 500 replications. $I[\cdot]$ stands for the indicator function, and $abs(\cdot)$ stands for the absolute value.

Coverage rates summarize the performance of the confidence intervals by indicating whether

we are actually rejecting the null at a rate that we would expect. With an unbiased estimator we expect that 95% confidence intervals formed using valid standard errors and critical values will include the true effect approximately 95% of the time. However, with a biased estimator coverage rates could be low even when the confidence intervals are the correct size. As we discuss further below, coverage rates also depend on appropriate critical values, which are key to inference and conclusions. Standard errors alone would not reveal this important deviation.

We are thus both interested in the coverage rates and whether the standard errors closely reflect the actual variability that we see in the estimates. Coverage rates highlight the policy implications and practical importance, but depend on bias as well as the width of the confidence intervals. Comparing the standard deviation of the estimates with the average standard error tells us how well the standard errors reflect the true variability, whether the estimator is biased or not. Therefore, it is important to consider coverage rates and standard errors in evaluating the performance of the confidence intervals.

In addition to reporting results based on regressions that pool all years, we report results from a method we refer to as the OLS-Lag cohort-by-cohort approach, or simply cohort-by-cohort. In this approach we estimate value-added separately for each cohort, and consider those estimates as independent pieces of information we use to construct confidence intervals. This approach sidesteps issues related to within classroom grouping, because it does not require an assumption about the within classroom error correlation; rather, the method effectively uses the classrooms as separate pieces of information. We estimate value-added cohort-by-cohort using OLS-Lag, obtaining one teacher effect estimate for each year the teacher is in the data.

$$\hat{\beta}^{p,r,q} = \frac{1}{q}\sum_{t=1}^{q} \hat{\beta}_t^{p,r,q} \tag{3.7}$$

We take the estimated value-added for each teacher as the average of their single-year estimates, as described in equation (3.7). Where $\hat{\beta}_t^{p,r,q}$ represents the estimate for the p^{th} teacher, from the r^{th} replication, and using data from the t^{th} cohort only, and $\hat{\beta}^{p,r,q}$ is the average of the single-year estimates.

$$se_k(\hat{\beta}^{p,r,q}) = \frac{SD(\hat{\beta}_t^{p,r,q})}{\sqrt{q}} \tag{3.8}$$

We use equation (3.8) to obtain a standard error for the estimate, where k now represents the OLS-Lag cohort-by-cohort approach. Calculating the estimated teacher effect and standard error in this way is equivalent to regressing the single-year teacher effect estimates for each teacher on a constant, and using the reported estimate and standard error. Once we have obtained an estimate, $\hat{\beta}^{p,r,q}$, and corresponding standard error, $se_k(\hat{\beta}^{p,r,q})$, for each repetition r, we take the standard deviation of the estimates as the measure of true variability, just as in equation (3.3). Similarly, we take the average standard error, to compare with the true variability, just as in equation (3.4).

Then we construct t-statistics for $H_0 : \beta_0^p = \beta_{true}^p$ using $\hat{\beta}^{p,r,q}$, $se_k(\hat{\beta}^{p,r,q})$, and the true teacher effect, just as we did in equation (3.5) for the previous methods. We then compare these t-statistics with the critical value from the t-distribution with degrees of freedom equal to the number of years the teacher is in the data minus one, i.e., $(q-1)$.[14] Thus, the coverage rate for the cohort-by-cohort method differs from the previous methods in that we use the appropriate critical value now from the t-distribution with $(q-1)$ degrees of freedom:

$$CoverageRate_k^{p,r,q} = (\frac{1}{500}) \sum_{r=1}^{500} I[abs(t_k^{p,r,q}) \leq t_{\alpha=0.025,df=(q-1)}] \tag{3.9}$$

So to obtain the cohort-by-cohort coverage rates we swap out the standard normal critical value for a critical value from a t-distribution, which depends on the number of cohorts.[15] Again this is equivalent to constructing confidence intervals round $\hat{\beta}^{p,r,q}$ teacher-by-teacher, using $se_k(\hat{\beta}^{p,r,q})$, and counting the number of times that we fail to reject true teacher effect.

The coverage rate is particularly important for evaluating one virtue of the cohort-by-cohort method - the critical value adjustment. If the t-distribution with degrees of freedom equal to the number of cohorts minus one is a good approximation for the distribution of the the t-statistic

[14] Teachers have the same number of classrooms in the simulations, but the number of classrooms can differ for each teacher in our analysis of the administrative data.

[15] We have done two robustness checks not included here. One draws the student fixed effect and the random error component from a t distribution with 5 degrees of freedom, and the other uses a chi-squared distribution with 2 degrees of freedom.

obtained in the cohort-by-cohort method, then this adjustment should lead to coverage rates near ninety-five percent. Whereas, assuming that the t-statistic follows a normal distribution would be too liberal.

3.3.3 Simulation Results

In each of the simulation results tables, we show the average standard deviation of the teacher fixed effects estimates, average of the teacher average standard error, and the coverage rate for each method. Each table displays results for the 3, 7, and 20 cohort cases. We also show results with a single cohort for the non-clustered standard errors and confidence intervals. In Table E.2 we show results for the OLS-Lag estimator using the DGP with a student fixed effect drawn from a standard normal distribution, no cohort-school fixed effect, and $\lambda = 0.5$ (referred to as Case 1 in Table E.1). We use three different grouping scenarios: random (RG), dynamic (DG, based on $A_{i,t-1}$), and heterogeneity grouping (HG, based on c_i). We show results for five schools in Panel A and for fifty schools in Panel B.[16] The first column refers to the number of cohorts used in the estimation. So for the rows with one cohort, we simulated data for a single cohort, obtained the valued-added (VA) estimates, standard errors and t-statistics using only that cohort, and then repeated that process for each replication. That means that we used data from only one class for each teacher. Estimates based 3 cohorts include students from three classrooms for each teacher, and so forth.

The second column shows the teacher average standard deviation of the VA estimates averaged across simulation repetitions, which is the average of equation (3.3) across all teachers. This tells us the average variability of the VA estimates. Columns three through six provide the average (across teachers) of equation (3.4) for each method: not clustering, classroom level clustering, cohort-school level clustering, and school level clustering, respectively. We consider the comparison of column 2 with columns 3 – 6 informative because they are two different measures of the variability of the VA estimator. Column 2 estimates estimator variability more directly, whereas, columns 3 – 6 are averages of what practitioners use. We expect that good standard errors will properly reflect the

[16] We only show results with fifty schools in this case. We have performed more simulations with 50 schools, but as previously discussed and shown in Table E.2, adding schools does not change the results.

variability of the estimates, so when the values in columns three through six are close to those in column 2, that method is considered to be performing well by this metric.

Consider column 3 in Panel A, which reports the average standard error without clustering under the random grouping – random assignment (RG-RA) scenario. Because students are grouped randomly in this case, we would expect the standard errors to closely reflect the standard deviation of the estimates (i.e., the variability of the estimator). The similarity of columns 2 and 3 supports this: the average standard error using one cohort and without clustering, reported in column 3, is 0.435, which is very close to the average standard deviation of the estimates reported in column 2, 0.419. Not surprisingly, the average coverage rate of the 95% confidence interval in this case, 0.956 reported in column 7, is almost exactly 95%. Not clustering works well in this case, because the students are grouped randomly, so treating them as independent observations leads to standard errors that reflect the true variability of the estimator and almost exact coverage rates, even with a single cohort of data. Similarly, not clustering gives average standard errors that are similar in magnitude to the average standard deviation of the estimates with any number of cohorts. Comparing columns 2 and 3 when using twenty cohorts, the average standard deviation, 0.095, is similar in magnitude to the average standard error without clustering, 0.097, and the coverage rate, 0.953 from column 7, is again almost exactly 95%. Again, this works well, because the students are randomly grouped in this scenario, so there is no need to account for within group correlation when constructing the confidence intervals.

If we account for within classroom correlation by clustering at the classroom level when grouping is random, we tend to underestimate the variance of the estimator and have low coverage rates with a small number of cohorts. From column 2 of Panel A, with three cohorts and random grouping, the average standard deviation of the VA estimates is 0.253. However, the average standard error under classroom clustering, from column 3 of Panel A, is 0.202. The underestimate of the variance leads to a coverage rate of 0.826, which is well below 95% and means that using this method underestimates the size of the confidence intervals with a small number of cohorts and random grouping. However, unnecessary classroom clustering is not problematic when the number of cohorts increases. The average standard error under classroom clustering with twenty cohorts, 0.095 from column 3 of

Panel A, is very similar to the average standard deviation, 0.095 from column 2 of Panel A. The corresponding coverage rate from column 8 is 0.942, which is much better than the 0.826 with three cohorts. The reason that clustering at the classroom level improves with more cohorts is that it effectively uses one independent piece of information from each classroom to obtain a standard error along with a standard normal critical value to construct the confidence interval. Under the RG-RA scenario, the classroom clustered variance estimator is asymptotically valid with an increasing number of cohorts. However it does not perform well with a small number of classrooms per teacher, because using only three observations to estimate the standard error for each fixed effect tends to underestimate the standard error.

In the second case, also in Panel A of Table E.2, we consider dynamic grouping with random assignment (DG-RA). DG-RA means that the students are grouped into classrooms based on their prior test score, $A_{i,t-1}$, then classrooms are randomly assigned to the teachers. Since we are calculating VA using OLS-Lag, we control for the students' prior exam score. Dynamic grouping does not increase the variance of the estimator, because we are conditioning on the grouping variable. We can see that by comparing the average standard deviations under dynamic grouping with the average standard deviations under random grouping. For example, from column 2 of Panel A, the average standard deviation under dynamic grouping with a single cohort is 0.431, and with twenty cohorts it is 0.095. Similarly, the average standard deviations under random grouping with one and twenty cohorts are 0.431 and 0.099, respectively. This similarity between the average standard errors without clustering for both dynamic and random grouping leads to almost identical coverage rates as well. For example, the average coverage rates under dynamic grouping with one and twenty cohorts are 0.954 and 0.947, respectively. The performance of the classroom clustered standard errors and corresponding confidence intervals are also similar under random and dynamic grouping. The classroom clustered standard errors tend to underestimate the standard deviation of the estimates with a small number of cohorts, 0.202 compared to 0.257 with three cohorts, leading to coverage rates well below 95%, 0.820 with three cohorts. However, the performance of the classroom-clustered standard errors improves with increasing cohorts, and the average coverage rate, 0.935 from column 8, is near 95% with twenty cohorts.

In contrast to the first two cases, when students are grouped based on the student fixed effect, heterogeneity grouping – random assignment (HG-RA), then non-clustered standard errors do not reflect the true variability of the estimator and the coverage rates are substantially lower. The average standard deviation with a single cohort goes from 0.419 under random grouping to 0.888 under heterogeneity grouping. This happens because grouping on an unobservable component increases unobserved variance across classrooms, and not clustering fails to account for that increased variance. The average standard error with a single cohort under heterogeneity grouping is only 0.399. This unobservable component of the student level equation is no longer independent within-classroom, and failing to account for the within-classroom correlation leads to drastically underestimating the variability of the estimator, even when the classrooms are randomly assigned to teachers. When the students in a classroom all have high (or low) average scores, we mistakenly think that we have a relatively precise estimate, based on those twenty observations. This, not surprisingly, leads to a low coverage rate of 0.546, which is much lower than we would hope to get for a 95% confidence interval.

More accurately, we should not consider those twenty student-level observations as independent, because they were grouped on an unobservable determinant of test scores, before the groups were assigned randomly to the teacher. Classroom level clustering can improve the performance of the confidence intervals, but only with a sufficient number of cohorts. With seven cohorts of data, even the classroom-clustered standard errors underreport the standard deviation of the estimates, 0.304 in column 4 compared to 0.340 from column 2. However, the corresponding confidence intervals do contain the teacher effect roughly 90% of the time, which is much better than not clustering. Under heterogeneity grouping, the classroom and cohort-school clustered standard errors perform similarly to the random grouping case, meaning that they underreport variability and cover the true parameter less than 95% of the time with a small number of cohorts, but they improve with an increasing number of cohorts. The reason that using classroom clustered standard errors works with an increasing number of cohorts under heterogeneity grouping is the same reason it works in the first two cases: assuming independence across classrooms is valid in all three cases, but the estimator is asymptotically justified with an increasing number of classrooms per teacher, and doesn't necessarily

work well with a small number of cohorts. Unfortunately, districts often want to make decisions regarding teacher effectiveness based on only a few years of data and waiting for twenty classrooms worth of observations for each teacher is generally not desirable or feasible.

There are two final points to make about Panel A of Table E.2. In all three scenarios presented in Panel A clustering at the cohort-school level performs similarly to classroom clustering. The average standard errors are similar with every number of cohorts. With three cohorts, the average standard error under RA-RG with classroom clustering is 0.202, which is almost identical to the average standard error with cohort-school clustering, 0.204. Similarity in coverage rates follows the similarity in standard errors. The average coverage rates under RA-RG for classroom clustering for three, seven, and twenty cohorts are 0.826, 0.914, and 0.942, which are almost identical to those for cohort-school clustering, 0.824, 0.913, and 0.943. The performance of classroom and cohort-school clustering is also similar under dynamic and heterogeneity grouping. For example, under heterogeneity grouping with 3 cohorts, the average classroom clustered standard error is 0.404 and the average cohort-school clustered standard error is 0.423, leading to coverage rates of 0.822 and 0.828, respectively. Clustering at the classroom level and cohort-school clustering give such similar results because the two calculations are nearly identical. Each teacher dummy is zero for all students outside of that teacher's classrooms, so the inner product of each teacher dummy variable and the residuals is the same whether you take it within classroom or within cohort. The only difference in the two variance-covariance estimators results from taking the inner product of lagged scores and residuals either within classroom or within cohort.[17] Although, we continue to show the results for both classrooms and cohort-school clustering from this point on, we refrain from going into detail about both, as they are generally the same.

Lastly, clustering at the school level severely underestimates the variation in the estimator and leads to the lowest coverage rates of any of the confidence intervals considered. With twenty cohorts, the average coverage rates using school level clustering are 0.020, 0.023, and 0.010 for RG-RA, DG-RA, and HG-RA, respectively, meaning that the confidence intervals only cover the true teacher

[17] Of course, their similarity could be a result of the simplicity of the simulations, as adding more covariates changes the calculation of each variance-covariance matrix. However, in our evaluations of the performance of the classroom and cohort clustered standard errors in administrative data, we find that there are practically no differences in their average performance.

effect about 1 – 2% of the time with twenty cohorts. The theoretical flaw with using school level clustering is that information used to estimate the treatments, i.e. the teacher fixed effects, does not accumulate when adding more schools. The mechanical issue is that teachers are nested in schools, and we are conducting inference on teacher fixed effects. The same issue would arise if we clustered by teacher and tried conducting inference on teacher fixed effects. Because we are estimating the fixed effects by OLS and teachers are nested in schools, within teacher residuals add to zero, and within school residuals also add to zero. This is problematic because the central matrix in the cluster robust variance sandwich estimator involves taking within cluster inner products of the right-hand-side variables and the residuals, but for teacher fixed effects this equates to summing residuals within school. In an extreme case, in which teachers are nested in schools and there are no right-hand-side variables other than the teacher fixed effects, the central matrix of the variance-covariance estimate (the "numerator") is a matrix of zeros. Adding right-hand-side variables or not having teachers nested in schools breaks this mathematical result, but the problem with underestimation of the standard errors appears to persist. Although we continue to show results from school clustering, we do not discuss them in detail from this point on.

To illustrate that increasing the number of schools has no impact on the results, we report results for the same three cases, RG-RA, DG-RA, and HG-RA, with fifty schools in Panel B of Table E.2. The results are strikingly similar to those for 5 schools. Similar to the five-school case, the average standard errors without clustering closely reflect the average standard deviation of the estimates when students are grouped randomly. The average standard error with a single cohort and no clustering, 0.446 from column 3 of Panel B, is similar to the average standard deviation of the estimates with one cohort, 0.454. The resulting coverage rate, 0.947 from column 7, is almost exactly 95%. The average classroom clustered standard errors and corresponding coverage rates with fifty schools and random grouping also perform similarly to the case with five schools, with coverage rates increasing in the number of cohorts from 0.812 to 0.944 for three and twenty cohorts, respectively. Results under heterogeneity grouping are also similar across the five and fifty school cases. The average standard deviation of the estimates with a single cohort and fifty schools increases to 0.894, which is similar to the average standard deviation of the estimates with a single cohort and five

schools, 0.888. Even with fifty schools, the non-clustered standard errors, 0.410 on average, still fail to account for the increased variance from heterogeneity grouping, leading to a low coverage rate of 0.586. Again, similar to the five-school case, increasing the number of cohorts does not improve the coverage rate very much for the confidence intervals that use non-clustered standard errors, as even with twenty cohorts of students the coverage rate without clustering is 0.663. The performance of the standard errors and confidence intervals from clustering at the classroom level improve with more cohorts when there are fifty schools, the same way that they do when there are five. The average standard error, from column 4 in Panel B under HG-RA, when clustering by classroom is 0.404, an underestimate of the average standard deviation of 0.518. The resulting coverage rate is 0.812, substantially lower than 95%. With twenty cohorts, the average classroom-clustered standard error is 0.194, which is similar to the analogous average standard deviation of the estimates of 0.201, and the coverage rate is 0.931. Comparing the results with five and fifty schools illustrates the argument that the number of schools is not important for estimating teacher effects in the simulations. The information on teacher effects accumulates from adding cohorts/classes for the existing teachers, and not from adding schools. For this reason, we only present and discuss results for simulations with fives schools from here on.

Table E.3 reports the results for Case 2, with an unobservable student fixed effect and an unobservable cohort-school fixed effect, each with a standard deviation of 0.25, under RG-RA, DG-RA, and HG-RA. Including a small cohort-school fixed effect influences the results in a similar manner as grouping based on an unobservable characteristic, e.g. the student fixed effect. An unobservable cohort-school fixed effect increases the variance of the estimates, such that standard errors with no clustering fail to accurately reflect the variation in the estimates. For example, under RG-RA with one cohort of data, the average standard deviation of estimates is 0.438 and the average standard error without clustering, from column 3, is 0.317. The non-clustered standard errors fail to account for the extra variance in the estimator from the unobservable fixed effect, and as a result, the average coverage rate with a single cohort of data and random grouping, from column 7 of Table E.3, is only 0.838. Even with more cohorts, the non-clustered standard errors fail to produce a better coverage rate, and with twenty cohorts the average coverage rate is only 0.84. Using cohort-school

(or classroom) clustering, however, does take into account the correlation within each cohort. With a small number of cohorts though, cohort-school clustering underestimates the average standard deviation, much like classroom and cohort-school clustering do in all three scenarios from Table E.2. With three cohorts under RG-RA with a cohort-school fixed effect, the average standard error using cohort-school clustering is 0.208, from column 5 of Table E.3, which is noticeably lower than the average standard deviation of estimates of 0.258. That leads to an average coverage rate of 0.835, much lower than the desired 0.95. However, the average coverage rate with cohort-school clustered standard errors improves with an increasing number of cohorts, and with twenty cohorts the coverage rate is 0.939, as seen in column 9 of Table E.3.

The results with HG-RA and a cohort-school fixed effect are also reported in Table E.3, but they are similar to the RG-RA results with a cohort-school fixed effect. They are similar, because in Table E.3 the student fixed effect has a relatively low variance, and grouping on a low-variance term has a much smaller impact on the variability of the estimates. For example, with three cohorts of data under heterogeneity grouping, the average standard deviation of the estimates is 0.264, from column 2 of Table E.3, but recall that the average standard deviation with three cohorts and heterogeneity grouping on the standard normal student fixed effects is 0.505, from column 2 of Panel A in Table E.2. The difference is that the total variation and variation in the error is larger in Table E.2. Regardless, clustering at the cohort-school level performs similarly for practical purposes in both cases, in that the coverage rates are low with a small number of cohorts, 0.814 with three cohorts, but improve as the number of cohorts increase, 0.937 with twenty. In general, in cases where we wish to account for within classroom dependence, we can do that using clustering when the number of classrooms per teachers is large. However, clustering leads to low coverage rates with a small number of cohorts.

In Panel A of Table E.4, average standard errors and coverage rates from the cohort-by-cohort method are presented under RG-RA, DG-RA, and HG-RA scenarios with the unobservable student fixed effect drawn from a standard normal distribution and no cohort-school fixed effect, i.e. Case 1 from Table E.1.[18] Although the method for estimating the teacher fixed effects is changed slightly

[18]We replicated the analysis from Panel A of Table E.4 under two other scenarios, drawing student fixed effects and random error terms from a t-distribution and a chi-squared distribution. The results are similar to the results using

in this method, it is not practically very different from pooling all years of data and using OLS-lag, so it does not alter the average standard deviations of the estimates very much.[19] From column 2 in Panel A of Table E.4, the average standard deviation of the estimates with three cohorts is 0.253, which is identical to the analogous number from Table E.2. Like clustering at the classroom level with the OLS-Lag estimator, the average standard error in the cohort-by-cohort method underestimates the average standard deviation of the estimates with a small number of cohorts, albeit less severely. Since constructing the confidence interval in the cohort-by-cohort case includes a critical value adjustment, using the critical value from the t-distribution with the number of cohorts minus one as opposed to the standard normal critical value, the coverage rate with three cohorts is 0.949, despite underestimating the variability of the estimator. As the number of cohorts increases, the average standard error and average standard deviation of the estimates become very close, with twenty cohorts they are 0.096 and 0.095, respectively. Much like classroom clustering with OLS-Lag, this method allows for complete dependence of the within classroom student level error terms. The critical value used in this method approaches the standard normal critical value as the number of cohorts increases, so that the adjustment becomes smaller, and with twenty cohorts the confidence intervals cover the true effect at an average rate of 0.959, slightly more conservative than the analogous coverage rate of 0.942 under random grouping with twenty cohorts from column 8 of Table E.2.

Results for the HG-RA case with a student fixed effect drawn from a standard normal distribution are also presented in Panel A of Table E.4. The average standard error in the cohort-by-cohort method reflects some of the increased variability in the estimates, even with a small number of cohorts. The average standard deviation of the estimates with three cohorts is 0.506, almost identical to the analogous number from Table E.2, and the average standard error is 0.470. From column 4 of Table E.4, the coverage rate using the cohort-by-cohort method under HG-RA with three cohorts is 0.934, a vast improvement from any of the coverage rates presented for the same case in Table E.2. For comparison, using OLS-Lag and classroom clustered standard errors with standard normal

standard normal distributions.

[19] Both estimators are unbiased for the teacher fixed effects under random assignment of classrooms to teachers and mechanically similar.

critical values, i.e. column 8 in Table E.2, only covers the true effect 82.2% of the time under HG-RA with three cohorts. With a larger number of cohorts, the cohort-by-cohort confidence intervals continue to achieve almost exact 95% coverage, with average coverage rates of 95.1% with seven cohorts and 95.3% with twenty cohorts.

Panel B shows results using the cohort-by-cohort method for the case with a small student fixed effect and a cohort-school fixed effect, comparable to Table E.3. Using the cohort-by-cohort methods does not lead to lower coverage rates when there is a cohort-school fixed effect, as it does when using OLS-Lag with classroom or cohort-school clustered standard errors. Even under heterogeneity grouping, a case in which we might expect lower coverage rates based on the results in Table E.3, the average coverage rate using the cohort-by-cohort method is 95.1% with 3 cohorts. Overall, despite a tendency of the standard errors from the cohort-by-cohort method to underestimate the average standard deviation of the estimate, the underestimation is less severe than the average standard errors from using clustering with OLS-Lag. The combination of the slightly higher standard errors and the critical value adjustment leads to coverage rates near 95% in every case presented in Table E.4, suggesting that the cohort-by-cohort method outperforms the other options, especially with a small number of cohorts.

Next, in Tables E.5 and E.6, we relax the random assignment aspect of the simulations, and re-evaluate the performance of the different confidence intervals. All simulation results in Tables E.5 and E.6 are based on the DGP in Case 1 of Table E.1, which draws the student fixed effect from a standard normal distribution and does not include a cohort-school fixed effect. In Table E.5, we report results using the OLS-Lag estimator under dynamic grouping with positive assignment (DG-PA) and heterogeneity grouping with positive assignment (HG-PA). Positive assignment means that the classroom with the higher average of the grouping variable, i.e. lagged scores for dynamic grouping and unobservable student fixed effect for heterogeneity grouping, is assigned to the teacher with the higher fixed effect. Under dynamic grouping, positive assignment does not lead to a biased estimator because, once again, the grouping variable is used in the conditioning set when estimating teacher VA. Similar to previous cases, not clustering under DG-PA leads to average standard errors that are similar to the average standard deviation of the estimates, 0.435 compared to 0.437 with a

single cohort, and produces confidence intervals with coverage rates around the 95% mark with any number of cohorts. The coverage rate under DG-PA with a single cohort and no clustering, from column 7 of Table E.5 is 0.953.

All of the estimators we have discussed up until this point are unbiased estimators for the teacher fixed effects. However, grouping on unobservable heterogeneity and positively assigning (HG-PA) classrooms to teachers leads to biased teacher value-added estimators. However, positively assigning classrooms with heterogeneity grouping does not lead to the substantial increase in the average standard deviation of the estimates that we saw previously. In fact, the average standard deviation of the estimates with a single cohort under heterogeneity grouping and positive assignment is 0.417, which is much lower than the 0.888 under HG-RA from Table E.2. The reason that heterogeneity grouping does not inflate the average standard deviation of the estimates with positive assignment relative to random assignment, is that higher (lower) VA teachers are consistently assigned to the classroom with the higher (lower) average student fixed effect, rather than being randomly assigned to some classrooms with a high and some with a low average student fixed effect. The fact that the average non-clustered standard errors are similar to the average standard deviation of the estimates, which are 0.399 and 0.417, respectively with one cohort, is more a coincidence than a sign of good performance. Because the estimators are biased though, coverage rates decrease with an increasing number of cohorts, and the average coverage rates for both the non-clustered and classroom clustering methods are 0.376 with twenty cohorts of data.

Finally, we present simulation results for the cohort-by-cohort method under DG-PA and HG-PA in Table E.6. Much like the OLS-Lag standard errors and confidence intervals under positive assignment, the coverage rates are low and do not improve with increasing cohorts when students are grouped on the unobservable student fixed effect, because grouping on an unobservable and non-random assignment produces a biased estimator. Coverage rates under HG-PA from Table E.6 are only 0.621 and 0.385 with three and twenty cohorts, respectively. With a biased estimator, such as the estimators we consider with HG-PA, the tighter confidence intervals with increasing cohorts are misleading.

Our simulation results show a number of key results. We compare the average standard errors

obtained from estimating teacher VA by OLS-Lag and not clustering, clustering at the classroom level, cohort-school clustering, and school clustering with the average standard deviation of the estimates across simulations repetitions. We find that under random or dynamic grouping of students and random assignment of classrooms to teachers, no clustering generally works well with any number of cohorts, which is not surprising since the student level errors are independent within classroom in these cases. Similarly, even with a single cohort, 95% confidence intervals from this method can obtain coverage rates around 95%. Clustering at the classroom or cohort-school level, however, tends to underestimate the variability of the estimator with a small number of cohorts, leading to coverage rates around 80 to 83% under most scenarios. But the performance of the confidence intervals based on classroom and cohort-school clustered standard errors improves with increasing cohorts, and they reach coverage rates around 92 – 94% in most scenarios. On the other hand, grouping students on an unobservable fixed effect and randomly assigning classrooms to teachers increases the estimator variance, which non-clustered standard errors fail to account for and thus performs poorly. While clustering at the classroom level produces standard errors that accurately reflect the variance of the estimator with a large number of cohorts under heterogeneity grouping, they still tend to underestimate it with a small number of cohorts. Consequently, the coverage rates of confidence intervals produced from classroom or cohort-school clustering are low with three cohorts, around 82%, but improve with increasing cohorts and reach over 90% with twenty cohorts.

In comparison, when we construct confidence intervals using our proposed cohort-by-cohort method, we find that even under heterogeneity grouping the confidence interval coverage rates are near 95%. The improved performance of the cohort-by-cohort method over clustering by classroom with OLS-Lag estimation is through both an increase in the average standard error and a critical value adjustment. The critical value depends on the number of cohorts used in the estimate, which corresponds to the pieces of independent information. With three cohorts under heterogeneity grouping and random assignment with a student fixed effect drawn from a standard normal distribution, the cohort-by-cohort method has a 93.4% coverage rate, which is much closer to 95% than the coverage rates using OLS-Lag with clustering and only three cohorts under heterogeneity grouping.

However, when students are grouped nonrandomly and classrooms are assigned nonrandomly,

the resulting bias in the teacher VA estimates is problematic for all methods. In this case, the cohort-by-cohort method does not improve the confidence interval performance when grouping and classroom assignment are based on an unobservable student fixed effect, as that leads to biased estimators.

While the simulations are helpful in understanding the relative performance of methods under known conditions, we turn to student-level administrative data in the next section to see how the methods compare in actual data. Much like the simulations, we can compare the size of the standard errors and confidence intervals under each method. Furthermore, we demonstrate what the differences in the confidence intervals means for separating teachers across the value-added distribution.

3.4 Confidence Intervals in Practice

3.4.1 Data

Using administrative data on six districts from a large state for years 2001 - 2007, we compare differences in standard errors and confidence intervals for value-added estimates produced using the previously discussed methods. All results in this section are based on estimating math value-added for fourth grade teachers. We restrict the dataset to fourth grade only for simplicity and to make comparisons with the simulations straightforward. We also restrict the sample to students with no missing data for any of the variables included in the regressions, and who are in classrooms with 10 or more students total.[20]

We estimate value-added separately by district, as in equation (3.2), where $X_{i,t}$ represents a set of student level characteristics, including race/ethnicity indicators, number of absences, and indicators for female, disability, limited English proficient and free- and reduced-price lunch. The largest of the six districts has 138,913 fourth grade students assigned to 2,580 teachers in the estimation sample.[21]

[20]The main results do not restrict to teachers with a certain number of classrooms. A second analysis, which can be found in the appendix, restricts to teachers with all seven years of data available to avoid changes in sample composition.

[21]See Table E.7 for additional student sample sizes and average characteristics by district.

3.4.2 Methods

We focus on conducting inference in four different ways. The first three are from estimating teacher value-added using OLS-Lag by district for all teachers and pooling all years. For each estimate, we report three standard errors: no clustering, cohort-school clustering, and school clustering.[22] Then we construct confidence intervals using standard normal critical values.

The fourth method is based on the previously discussed OLS-Lag cohort-by-cohort method. For the cohort-by-cohort method, we first calculate OLS-Lag value-added estimates for each year in the data. Then for each teacher we use the average of the single-year estimates as the estimated teacher value-added, and the standard error of that mean to construct confidence intervals using the critical value from the t-distribution with degrees of freedom equal to the number of years in the data minus one.

For each method, we provide the average standard errors and confidence interval width. Lastly, we demonstrate practical implications of the differences, by comparing the percentages of teachers in different parts of the distribution with upper (lower) bounds that are under (above) different percentiles of the distribution under each method. For example, we show the percent of teachers with estimates in the top 10% of the VA distribution who also have a lower bound above the 90^{th} percentile for each confidence interval. This sheds light on how policy prescriptions depend on the method used to calculate confidence intervals.

3.4.3 Results

Figure F.1 compares the average standard errors for each method with 2 – 7 cohorts for six districts. There is some variation across districts in the ordering of the non-clustered, cohort-school clustered, and cohort-by-cohort standard errors. In District A, B, and E, the cohort-by-cohort standard errors are larger for every number of cohorts, and in some cases the differences are substantial. Across all of our simulations, from Tables E.2, E.3, and E.5, the standard errors obtained from cohort-school clustering underestimate the standard deviation of the estimates with a small number of

[22] We omit the classroom clustering case in our reporting, because in most cases they are almost identical to cohort-school clustered standard errors on average.

cohorts, and the average cohort-by-cohort standard errors are generally larger with a small number of cohorts. This is perhaps the difference that we observe in Districts A, B, and E in Figure F.1. For District A the average cohort-school clustered standard error with two cohorts is about 20 points, and the average cohort-by-cohort clustered standard error is more than double that, around 50 points with two cohorts.[23] The difference in the two averages shrinks with increasing cohorts, and with seven cohorts the average school-cohort clustered standard error is a little less than 20 points, and the average cohort-school clustered standard error is around 35 points. Their convergence is more pronounced in District E, for which they only differ by about five points with seven cohorts.

On the other hand, in Districts C, D, and F, the order is switched, with the average cohort-school clustered standard errors larger than the cohort-by-cohort standard errors for most cohorts. The largest gap, in District C, appears to be increasing with cohorts, and with seven cohorts the average cohort-school clustered standard error is a little less than 40 points, while the average cohort-by-cohort standard error is a little over 20, leaving a gap of about 15 points. The fact that the average cohort-school clustered standard error is larger than the average non-clustered standard error is consistent with heterogeneity grouping, as well as any scenario with an observed cohort-school fixed effect. However, based on the simulations, we might expect the cohort-by-cohort standard errors to be larger, relative to the cohort-school clustered standard errors. Despite the average cohort-by-cohort standard error being smaller than the cohort-clustered standard error in District C, it does not mean that the cohort-by-cohort confidence intervals will be smaller, as they use a larger critical value.

In Figure F.2 we compare the average width of confidence intervals constructed from the different methods. It is encouraging that, as the number of cohorts increases, the sizes of the cohort-by-cohort and OLS-Lag confidence intervals become similar. However, with a small number of cohorts, the cohort-by-cohort confidence intervals tend to be much more conservative than the clustering methods. Of course, part of the difference is due to the larger critical values used in the cohort-

[23]To put these numbers in perspective, the average score for the estimation sample in one of the districts is 1592 with a standard deviation of 258. Value-added estimates in that district for teachers in the sample for at least 2 years range from -161 to 306. The average estimated teacher effect is about 16 and the standard deviation of teacher effect estimates is almost 56. The average cohort clustered standard error is 41.7, whereas the average of the non-clustered standard errors is 29.5.

by-cohort method. In some districts though, such as Districts A, B, and E, the cohort-by-cohort standard errors are also larger than any other method, leading to a larger gap between the confidence intervals. For example, in District E, for which the average standard error is over 40 points with three cohorts under the cohort-by-cohort method but closer to 30 with three cohorts for the no clustering and cohort-school clustering methods, the cohort-by-cohort confidence intervals appear especially conservative with a small number of cohorts. Despite the large difference with three cohorts, the gap in the confidence intervals is much smaller with seven cohorts, because the gaps in the average standard errors and the difference in the critical values both decline with the number of cohorts.

In other districts, where the cohort-by-cohort average standard error is smaller than the cohort-school clustered standard errors, such as Districts D and F, the cohort-by-cohort confidence intervals are still noticeably larger with a small number of cohorts due to the increased critical value. By seven cohorts, the difference disappears and the confidence interval widths are virtually identical.

Finally, in Table E.8 we consider how frequently we can distinguish teachers from their peers based on VA. This is of particular interest if districts use value-added estimates to guide employment decisions. The teacher value-added and standard errors are estimated district-by-district, but the table reports average results across all six districts. Panel A shows the percentage of teachers from the bottom 10% of the VA distribution who have an upper bound on their confidence interval below the 10^{th}, 25^{th}, 50^{th}, and 90^{th} percentiles. From column 2 of Panel A, using OLS-Lag and no clustering, 7.4% of teachers in the bottom decile have an upper bound that is also below the tenth percentile. In other words, 7.4% of teachers with VA estimates in the bottom ten percent have confidence intervals that are completely contained in the bottom decile when we do not use any clustering. For comparison, when using cohort-school clustering, 23.5% of the bottom decile teachers have confidence intervals that are completely contained in the bottom decile. The fact that cohort-clustering leads to more teachers in the bottom decile is in line with some of the simulation results. As we saw in the simulations, even if assuming independence across classrooms (or cohorts) is correct, simply clustering tends to underestimate the variance with a small number of cohorts. Using the cohort-by-cohort method, also often underestimated the variance with a small number of cohorts in the simulations, but the critical value adjustment leads to higher coverage rates. Using the cohort-

by-cohort method in the administrative data, only 0.8% of bottom decile teachers have a confidence interval with an upper bound that is also below the tenth percentile. Similarly, not clustering and cohort-school clustering assign confidence intervals that are completely contained below the 25^{th} percentile for 44.1% and 77.9% of the teachers, respectively. However, using the cohort-by-cohort method only produces confidence intervals that are completely contained in the bottom quartile of the VA distribution for 1.6% of teachers in the bottom decile. In other words, using the cohort-by-cohort method produces much more conservative conclusions about the separation of teachers across the VA distribution. In fact, from column 5 of Table E.8, the cohort-by-cohort method suggests that only 29% of bottom decile teachers have a 95% confidence interval with an upper bound below the 90^{th} percentile. If we take that at face value, it suggests that it is difficult to exclude 71% of the teachers in the bottom decile from having a true effect in the top decile.

Fortunately, the cohort-by-cohort confidence intervals are better at distinguishing teachers in the upper end of the VA distribution. In Panel C of Table E.8, we show the percentage of teachers in the top decile with lower bounds above the 10^{th}, 25^{th}, 50^{th}, and 90^{th} percentiles. From column 9, we find that 8.1% of teachers in the top 10% of the VA distribution have lower bounds that are also above the 90^{th} percentile when using the cohort-by-cohort method. This indicates much better separation than the less than one percent of the bottom decile teachers with confidence intervals in the bottom decile. Furthermore, we find that the cohort-by-cohort method assigns confidence intervals to 18.5% and 45.2% of top decile teachers with lower bounds above the 75^{th} and 25^{th} percentiles, respectively. The better separation of top decile teachers can come from a combination of sources. For example, confidence intervals, especially from the cohort-by-cohort method, shrink with more cohorts. In general, if we expect that the cohort-by-cohort method generates substantially better coverage rates than the other options, then simply using OLS-Lag with, or without clustering, will lead to greatly over-estimating the degree of separation between teachers in the VA distribution.

3.5 Conclusion

In exploring the precision of teacher value-added estimates, we first look at four easily calculable methods of constructing confidence intervals for OLS-Lag estimates of teacher value-added: no clustering, clustering at the classroom level, cohort-school level, or school level. Using simulations, we find that in certain cases—particularly when grouping students based on unobservable heterogeneity or under the presence of an unobservable cohort fixed effect—not clustering is problematic as it fails to account for within classroom correlation. Furthermore, standard variance estimators that are meant to account for within group correlation, classroom clustering and cohort-school clustering in this case, can perform well with a large number of cohorts, but underestimate the variance with a small number of cohorts.

We propose a different method that we call OLS-Lag cohort-by-cohort, or simply cohort-by-cohort, which includes a critical value adjustment and, like classroom and cohort-school clustering, does not rely on any within classroom independence assumption. The OLS-Lag cohort-by-cohort method also tends to underestimate the variance of the estimator with a small number of cohorts, but produces coverage rates near 95% under all grouping scenarios with random assignment, because of the critical value adjustment used in constructing the confidence intervals. There are several cases where this method outperforms the others. Specifically, with a small number of cohorts and when students are grouped non-randomly, the confidence intervals from the cohort-by-cohort method perform much better than those from not clustering and from clustering by classroom or cohort-school. Furthermore, with heterogeneity grouping or cohort-school fixed effects, using the cohort-by-cohort method outperforms the confidence intervals from not clustering even as the number of cohorts increases, because not clustering completely fails to account for any within class correlation. The confidence intervals obtained from the cohort-by-cohort method and the classroom and cohort-school clustering methods do converge with increasing cohorts, but using a large number of cohorts is unrealistic in the teacher-value added context.

We also use student level administrative data to compare the performance of the different methods in six districts from a large state. We find that there is some heterogeneity in relative size of the cohort-by-cohort standard errors and the other standard errors considered, but that in all

cases the cohort-by-cohort confidence intervals are wider with a small number of cohorts, in part because of the critical value adjustment. We find that if we use the cohort-by-cohort method, only 29% of teachers in the bottom decile of the VA distribution have a 95% confidence interval with an upper bound below the 90^{th} percentile of the VA distribution. This number is much lower than that produced using OLS-Lag estimation with cohort-school clustering or no clustering at all. Both of these result in 100% of bottom decile teachers having an upper bound below the 75^{th} percentile. We also find that, while cohort-school clustering assigns 75.8% of the top decile teachers a lower bound that is higher than the 75^{th} percentile, placing them safely in the top quartile of the VA distribution, the cohort-by-cohort method only assigns 18.5% of the top decile teachers a confidence interval that is completely contained in the top quartile. This shows that the calculation of confidence intervals has important implications for identifying teachers in the tails of the distribution. The cohort-by-cohort method is conservative relative to the other methods considered, suggesting that using the other methods could lead districts and policymakers to drastically overestimate the degree of separation between teachers in the VA distribution. In conclusion, we cannot prescribe a single method that will necessarily provide accurate coverage rates of the true teacher effects under all possible data scenarios, bur rather, we advise researchers and practitioners to exercise caution when drawing inferences from these estimates.

APPENDICES

APPENDIX A

Figures for Chapter 1

Figure A.1: Average Test Scores by DL Attendance

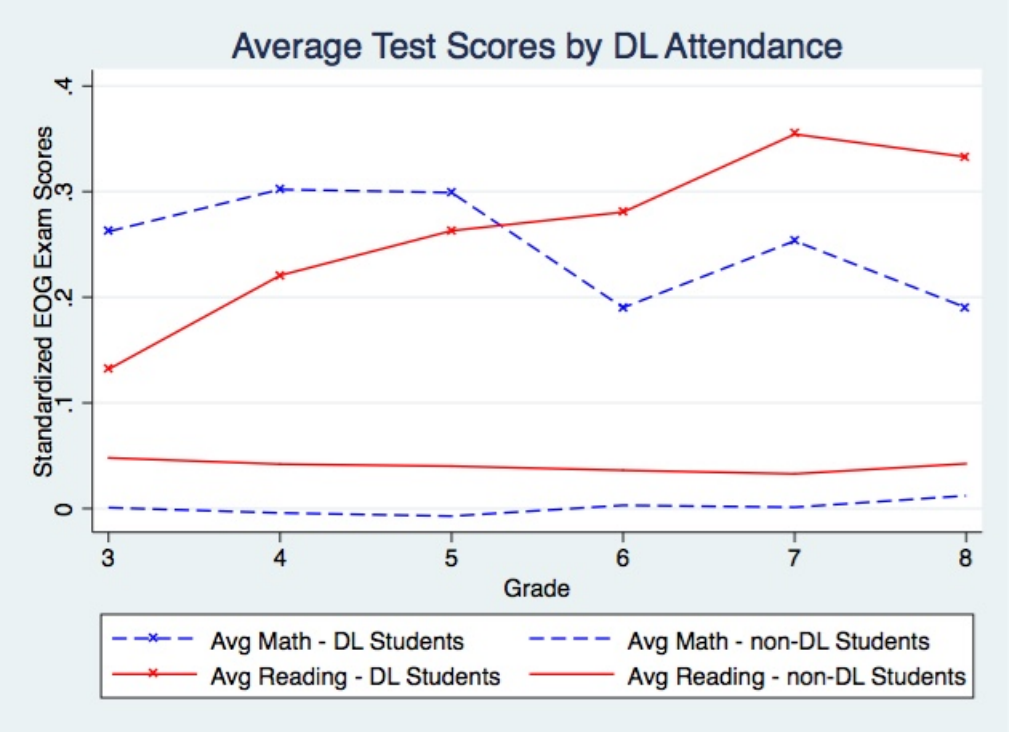


Figure A.2: LEP Average Test Scores by DL Attendance

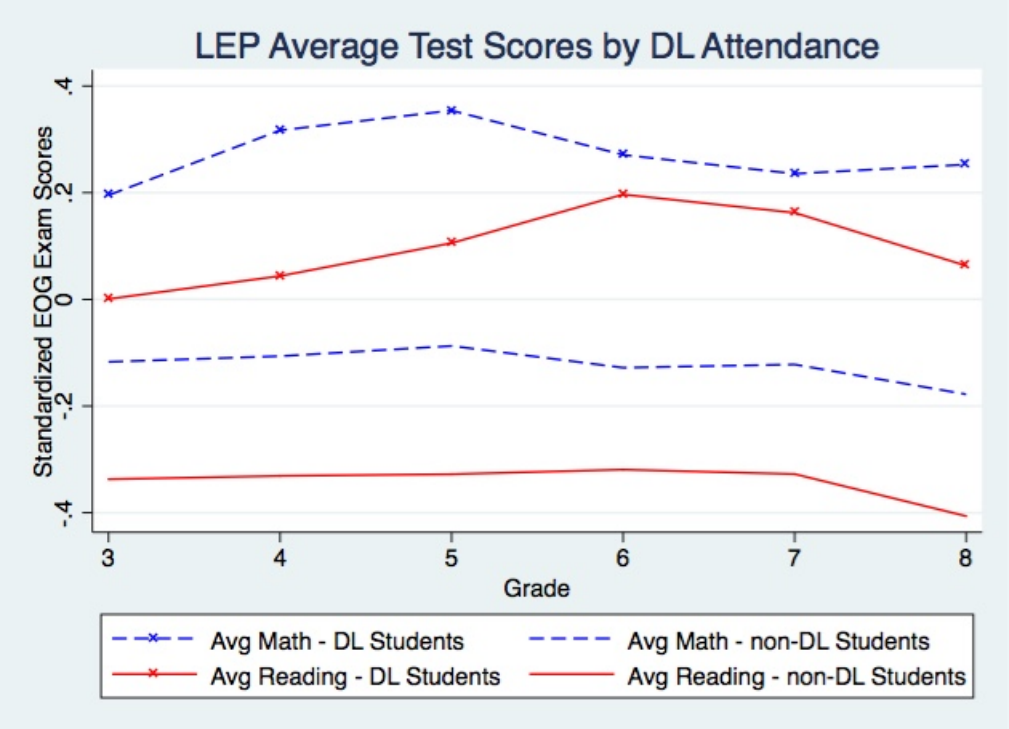


Figure A.3: Average Test Scores by LEP Status

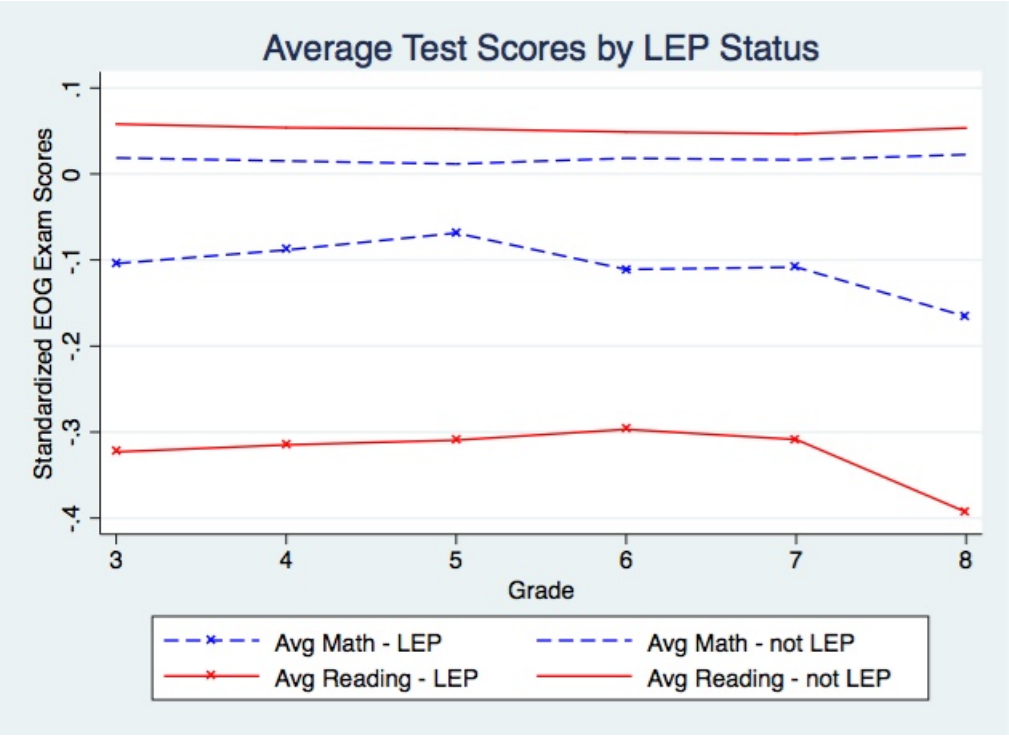
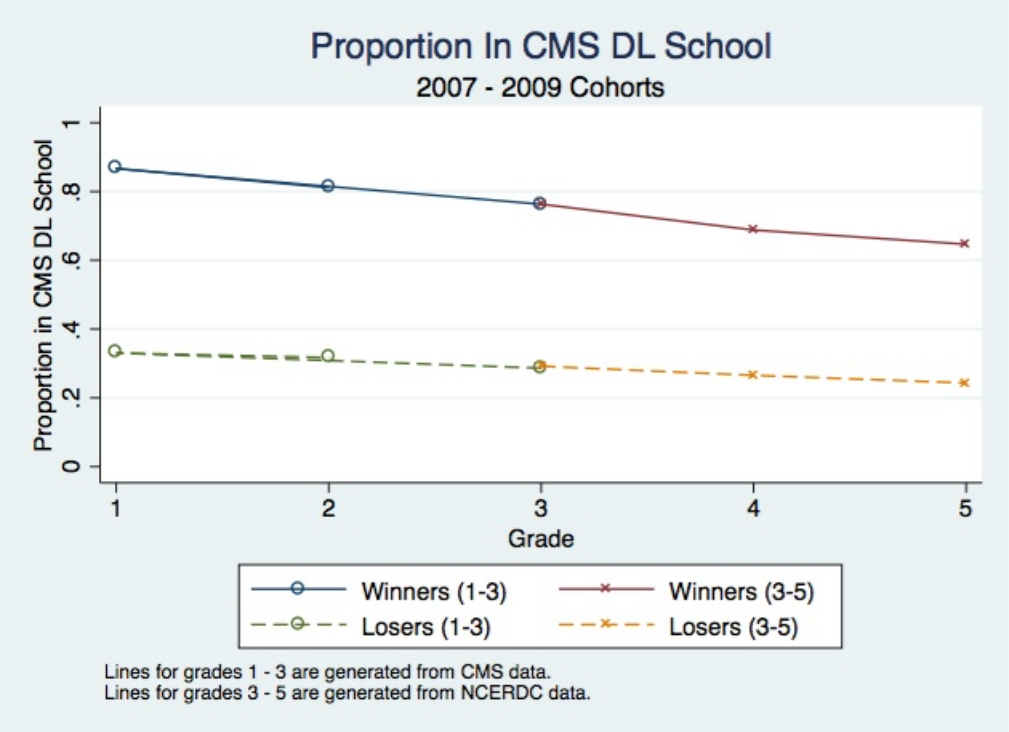


Figure A.4: Proportion in CMS DL School



APPENDIX B

Tables for Chapter 1

Table B.1: Application Numbers and Neighborhood School Characteristics

	One-Way School	Two-Way Schools		
	Waddell	**Collinswood**	**Oaklawn**	**Other Applicants**
	[1]	[2]	[3]	[4]
Applicants				
Sibling Placement	0.296	0.232	0.218	0.230
Won First Choice	0.782	0.561	0.951	0.589
DL Applications	2.081	1.248	1.283	0.093
Neighborhood School				
White	0.315	0.250	0.119	0.197
Black	0.383	0.389	0.607	0.501
Hispanic	0.216	0.281	0.194	0.220
FRPL	0.574	0.652	0.760	0.688
LEP	0.180	0.232	0.175	0.194
EOG Exam Scores				
Math	-0.010	-0.080	-0.318	-0.186
Reading	-0.032	-0.136	-0.357	-0.224
N	1,147	1,112	533	13,071

*Notes: The first three rows display of the type and number of applications submitted for all those submitting applications in the CMS school choice lottery. The rest of the table shows mean characteristics of the neighborhood schools that applicants are assigned to weighted by the number of applicants from each school. Everything is based on first choice school.

Table B.2: Second and Third Choices

	Second Choice DL [1]	**Third Choice DL** [2]
Attend DL School	0.105	0.087
Second Choice DL	1.000	0.219
Third Choice DL	0.203	1.000
Assignment		
Collinswood	0.000	0.000
Waddell	0.080	0.072
Oaklawn	0.077	0.075
Any DL Choice	0.157	0.147
Choice		
Collinswood	0.339	0.245
Smith	0.332	0.475
Oaklawn	0.329	0.279
Won		
First Choice	0.490	0.426
Second Choice	0.143	0.094
Third Choice	0.070	0.117
Any Choice	0.703	0.638
Observations	286	265

Table B.3: Dual Language and Neighborhood School Characteristics

	Waddell		**Collinswood**		**Oaklawn**		
	Area [1]	Waddell [2]	Area [3]	Collinswood [4]	Area [5]	Oaklawn [6]	Other Magnets [7]
Teaching Experience							
0 - 3 Years	0.329	0.320	0.294	0.501	0.298	0.542	0.299
11+ Years	0.300	0.272	0.367	0.255	0.359	0.195	0.374
FRPL	0.806	0.334	0.759	0.573	0.900	0.688	0.568
AYP Targets	0.869	0.986	0.871	1.000	0.779	1.000	0.873
Pct at Grade Level							
Reading	0.612	0.822	0.612	0.880	0.496	0.758	0.718
Math	0.698	0.878	0.705	0.944	0.531	0.753	0.733
KG Class Size	18.057	21.400	19.049	22.333	18.200	19.333	18.736

*Note: Average characteristics at each dual language school, for the neighborhood schools with zones contiguous to each dual language school, and all other magnet schools.

Table B.4: Summary and Balance - English Proficient Sample

	Application Sample			Estimation Sample		
	Won	**Lost**	**Test**	**Won**	**Lost**	**Test**
	[1]	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**
Attended (K/First)	0.870	0.358	0.509***	0.899	0.375	0.521***
	(0.337)	(0.480)	(0.042)	(0.302)	(0.485)	(0.050)
Won Any Choice	1.000	0.364	0.645***	1.000	0.367	0.649***
	(0.000)	(0.482)	(0.050)	(0.000)	(0.483)	(0.049)
Won Any DL Choice	1.000	0.113	0.880***	1.000	0.117	0.879***
	(0.000)	(0.317)	(0.036)	(0.000)	(0.322)	(0.036)
Female	0.546	0.513	0.012	0.533	0.508	0.003
	(0.499)	(0.501)	(0.045)	(0.500)	(0.501)	(0.050)
Black	0.304	0.354	-0.043	0.303	0.387	-0.075*
	(0.461)	(0.479)	(0.042)	(0.460)	(0.488)	(0.041)
White	0.372	0.268	0.017	0.373	0.262	0.027
	(0.484)	(0.444)	(0.043)	(0.484)	(0.441)	(0.047)
Hispanic	0.112	0.215	-0.038	0.111	0.206	-0.048
	(0.316)	(0.412)	(0.039)	(0.315)	(0.405)	(0.041)
Second Choice DL	0.407	0.328	0.074	0.408	0.315	0.081
	(0.492)	(0.470)	(0.044)	(0.492)	(0.465)	(0.049)
Third Choice DL	0.245	0.212	0.054	0.247	0.210	0.037
	(0.431)	(0.409)	(0.036)	(0.432)	(0.408)	(0.049)
Non-missing Test Scores	0.847	0.821	0.000	1.000	1.000	
	(0.361)	(0.384)	(0.035)	(0.000)	(0.000)	
FRPL	0.181	0.373		0.174	0.351	
	(0.374)	(0.467)		(0.380)	(0.478)	
EOG Math Score				0.524	0.281	0.269**
				(1.004)	(0.979)	(0.110)
EOG Reading Score				0.452	0.249	0.136
				(0.941)	(0.897)	(0.102)
Lottery FE			X			X
FRPL-Year Dummies			X			X
Neighborhood School FE			X			X
Observations	339	302	641	287	248	535
Number of Clusters			46			44

Table B.5: Summary and Balance - ESL/LEP Sample

	Application Sample			Estimation Sample		
	Won	Lost	Test	Won	Lost	Test
	[1]	[2]	[3]	[4]	[5]	[6]
Attended (K/First)	0.891	0.275	0.642***	0.929	0.290	0.669***
	(0.312)	(0.448)	(0.067)	(0.258)	(0.455)	(0.068)
Won Any Choice	1.000	0.298	0.735***	1.000	0.297	0.711***
	(0.000)	(0.459)	(0.061)	(0.000)	(0.458)	(0.060)
Won Any DL Choice	1.000	0.023	0.976***	1.000	0.021	0.987***
	(0.000)	(0.152)	(0.013)	(0.000)	(0.143)	(0.013)
Female	0.457	0.444	-0.046	0.442	0.455	-0.052
	(0.500)	(0.498)	(0.074)	(0.499)	(0.500)	(0.096)
Black	0.054	0.053	0.016	0.062	0.055	0.026
	(0.227)	(0.224)	(0.027)	(0.242)	(0.229)	(0.025)
White	0.054	0.029	0.040	0.044	0.034	0.030
	(0.227)	(0.169)	(0.030)	(0.207)	(0.183)	(0.031)
Hispanic	0.829	0.895	-0.099**	0.823	0.890	-0.105**
	(0.378)	(0.308)	(0.043)	(0.383)	(0.314)	(0.046)
Second Choice DL	0.202	0.205	-0.107*	0.195	0.234	-0.136**
	(0.403)	(0.405)	(0.058)	(0.398)	(0.425)	(0.065)
Third Choice DL	0.147	0.111	0.003	0.142	0.117	-0.028
	(0.356)	(0.315)	(0.034)	(0.350)	(0.323)	(0.036)
Non-missing Test Scores	0.876	0.848	0.055	1.000	1.000	
	(0.331)	(0.360)	(0.046)	(0.000)	(0.000)	
FRPL	0.679	0.726		0.690	0.724	
	(0.462)	(0.441)		(0.464)	(0.448)	
EOG Math Score				0.082	-0.199	0.154
				(0.802)	(0.888)	(0.141)
EOG Reading Score				-0.074	-0.313	0.223
				(0.833)	(0.797)	(0.141)
Lottery FE			X			X
FRPL-Year Dummies			X			X
Neighborhood School FE			X			X
Observations	113	145	300	113	145	258
Number of Clusters			39			36

Table B.6: Impact of Attending a Dual Language School on Achievement

Panel A: English Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.464***	0.042*		0.024	
			(0.064)	(0.022)		(0.015)	
Attend DL School	-0.004	-0.022*			0.089*		0.053*
	(0.018)	(0.011)			(0.047)		(0.032)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	1,472	1,472	1,472	1,472	1,472	1,472	1,472
Number of Clusters	44	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.667***	0.052**		0.042*	
			(0.069)	(0.024)		(0.024)	
Attend DL School	0.052***	0.065***			0.078**		0.064**
	(0.019)	(0.019)			(0.034)		(0.032)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	809	809	809	809	809	809	809
Number of Clusters	36	36	36	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with years of treatment (grade plus one). Standard errors are clustered by lottery.

Table B.7: Impact of Attending a Dual Language School on Achievement

Panel A: English Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.471***	0.040*		0.027*	
			(0.081)	(0.020)		(0.014)	
Attend DL School	0.011	-0.010			0.086**		0.057*
	(0.017)	(0.014)			(0.043)		(0.032)
Observations	1,472	1,472	1,472	1,472	1,472	1,472	1,472
Number of Clusters	44	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.635***	0.040**		0.044**	
			(0.084)	(0.019)		(0.020)	
Attend DL School	0.039**	0.065***			0.063**		0.069**
	(0.019)	(0.023)			(0.031)		(0.030)
Observations	809	809	809	809	809	809	809
Number of Clusters	36	36	36	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, and year of exam. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with years of treatment (grade plus one). Standard errors are clustered by lottery.

Table B.8: Attrition and Weighting

Panel A: Summary of Probabilities of Staying

	Full Sample		English Sample		ESL/LEP Sample	
	Winners	**Losers**	**Winners**	**Losers**	**Winners**	**Losers**
	[1]	[2]	[3]	[4]	[5]	[6]
Average Pr(Stay)	0.855	0.831	0.847	0.821	0.876	0.848
SD Pr(Stay)	0.031	0.038	0.043	0.060	0.040	0.047
APE	0.024		0.015		0.019	
(SE)	(0.024)		(0.030)		(0.042)	
N	468	473	339	302	129	171

Panel B: Non-Random Attrition

	Coefficients on Indicator for Winning Lottery		
	Full Sample	**English Sample**	**ESL/LEP Sample**
	[1]	[2]	[3]
No Controls	0.024	0.026	0.028
(SE)	(0.024)	(0.030)	(0.040)
Controls + Lottery FE	0.019	-0.003	0.047
(SE)	(0.024)	(0.029)	(0.036)
+ Neighborhood School FE	0.025	-0.002	0.052
(SE)	(0.029)	(0.038)	(0.049)
N	941	641	300

*Notes: Panel A summarizes estimated probabilites of having at least one set of test scores available. The estimates are based on logit regressions including gender, race, frpl-year, and an indicator for winning the lottery. Panel B shows estimated coefficients from OLS regressions of having at least one set of test scores available on winning the lottery. The baseline controls are gender, race, and frpl-year. The second set of OLS estimates also condition on lottery fixed effects, and the third set include neighborhood school fixed effects.

Table B.9: Impact of Attending a Dual Language School on Achievement - Weighted

Panel A: English Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.468***	0.042*		0.024	
			(0.065)	(0.022)		(0.015)	
Attend DL School	-0.004	-0.022*			0.089*		0.052*
	(0.018)	(0.012)			(0.046)		(0.031)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	1,472	1,472	1,472	1,472	1,472	1,472	1,472
Number of Clusters	44	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.673***	0.052**		0.043*	
			(0.071)	(0.025)		(0.025)	
Attend DL School	0.053***	0.066***			0.078**		0.064**
	(0.019)	(0.020)			(0.033)		(0.032)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	809	809	809	809	809	809	809
Number of Clusters	36	36	36	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with years of treatment (grade plus one). Regression are weighted by the inverse probability of having test scores available in the data. Weights were generated from logit regressions. Standard errors are clustered by lottery.

Table B.10: Heterogeneous Effects

Panel A: English Sample

	Gender		School Type		Race/Ethnicity		
	Female	Male	One-Way	Two-Way	White	Black	Hispanic
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Math	0.106**	0.068	0.081	0.090*	0.190**	0.046	0.090*
	(0.054)	(0.046)	(0.127)	(0.046)	(0.076)	(0.053)	(0.048)
Reading	0.069**	0.030	0.032	0.054**	0.084	0.034	0.115***
	(0.034)	(0.043)	(0.104)	(0.028)	(0.055)	(0.034)	(0.037)
Neighborhood School FE	X		X			X	
Observations	1,472		1,472			1,472	
Number of Clusters	44		44			44	

Panel B: ESL/LEP Sample

	Gender		School Type		Race/Ethnicity		
	Female	Male	One-Way	Two-Way	White	Black	Hispanic
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Math	0.051	0.090**	-0.018	0.079**	-	-	0.083***
	(0.048)	(0.038)	(0.176)	(0.034)	-	-	(0.031)
Reading	0.011	0.087**	-0.055	0.065**	-	-	0.062**
	(0.045)	(0.039)	(0.157)	(0.031)	-	-	(0.031)
Neighborhood School FE	X		X			X	
Observations	809		809			809	
Number of Clusters	36		36			36	

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with years of treatment (grade plus one). Heterogeneous effects are estimated using interactions with the assignment variable and instrumenting for interactions with the attendance variable. Standard errors are clustered by lottery.

Table B.11: Effects by Grade

Panel A: English Sample

	OLS		ITT		LATE	
	Math	Reading	Math	Reading	Math	Reading
	[1]	[2]	[3]	[4]	[5]	[6]
Grade						
Third	-0.011	-0.152*	0.189**	0.108	0.374**	0.215
	(0.120)	(0.082)	(0.092)	(0.079)	(0.171)	(0.147)
Fourth	0.042	-0.086	0.211	0.070	0.446	0.151
	(0.112)	(0.078)	(0.143)	(0.093)	(0.284)	(0.188)
Fifth	-0.009	-0.115	0.139	0.216**	0.264	0.431**
	(0.138)	(0.094)	(0.150)	(0.096)	(0.276)	(0.205)
Sixth	-0.107	-0.107	0.327**	0.136	0.721**	0.298
	(0.121)	(0.091)	(0.141)	(0.135)	(0.346)	(0.295)
Neighborhood School FE	X	X	X	X	X	X
Observations	1,472	1,472	1,472	1,472	1,472	1,472
Number of Clusters	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS		ITT		LATE	
	Math	Reading	Math	Reading	Math	Reading
	[1]	[2]	[3]	[4]	[5]	[6]
Grade						
Third	0.055	0.312***	0.244**	0.244*	0.393**	0.390**
	(0.097)	(0.101)	(0.109)	(0.128)	(0.174)	(0.187)
Fourth	0.374**	0.254*	0.413**	0.153	0.657**	0.238
	(0.138)	(0.134)	(0.171)	(0.167)	(0.274)	(0.248)
Fifth	0.397***	0.370***	0.320*	0.208	0.471**	0.302
	(0.131)	(0.100)	(0.163)	(0.137)	(0.229)	(0.188)
Sixth	0.417***	0.408***	0.367**	0.364**	0.542**	0.531***
	(0.134)	(0.145)	(0.171)	(0.140)	(0.231)	(0.183)
Neighborhood School FE	X	X	X	X	X	X
Observations	809	809	809	809	809	809
Number of Clusters	36	36	36	36	36	36

Robust standard errors in parentheses

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

*Notes: Each regression includes lottery fixed effects and controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and intearacted with each exam grade. Instruments are interactions between grade of exam and the indicator for winning the lottery. Standard errors are clustered by lottery.

Table B.12: Impact of Dual Language Schooling on LEP Status

	Limited English Proficient			
	OLS		LATE	
	[1]	[2]	[3]	[4]
Attend DL School				
Grade 3	-0.025	0.021	-0.032	0.045
	(0.047)	(0.055)	(0.084)	(0.101)
Grade 4	-0.148**	-0.120	-0.159	-0.107
	(0.071)	(0.080)	(0.137)	(0.160)
Grade 5	-0.192**	-0.176*	-0.071	-0.051
	(0.079)	(0.093)	(0.118)	(0.144)
Grade 6	-0.210***	-0.196**	-0.168*	-0.141
	(0.073)	(0.076)	(0.086)	(0.111)
Neighborhood School FE		X		X
Observations	809	809	809	809
Number of Clusters	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of observation, and year of observation. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with grade dummy variables. Estimates are from OLS and 2SLS interaction terms. Standard errors are clustered by lottery.

APPENDIX C

Figures for Chapter 2

Figure C.1: MTI - Weekday

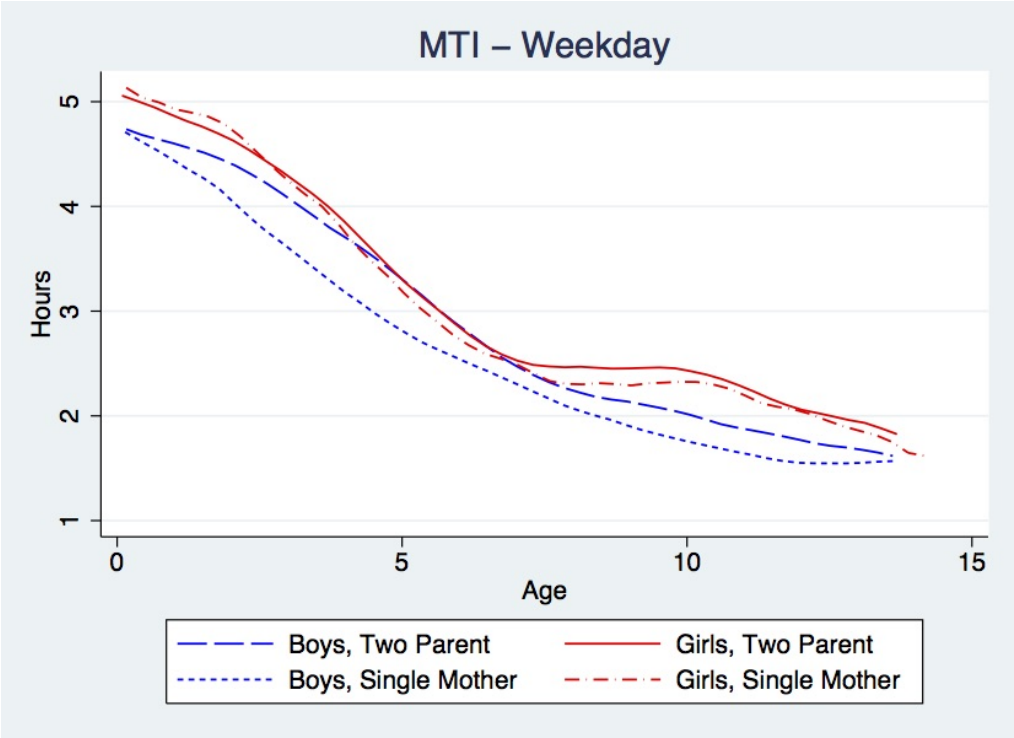


Figure C.2: MTI - Weekend

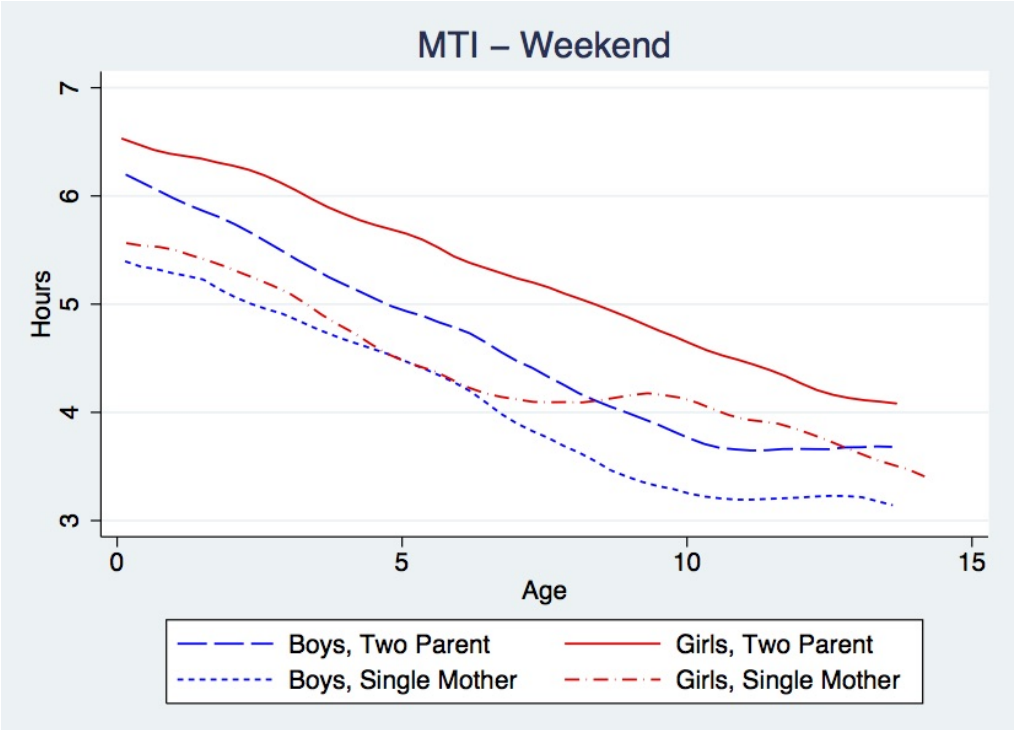


Figure C.3: FTI - Weekday

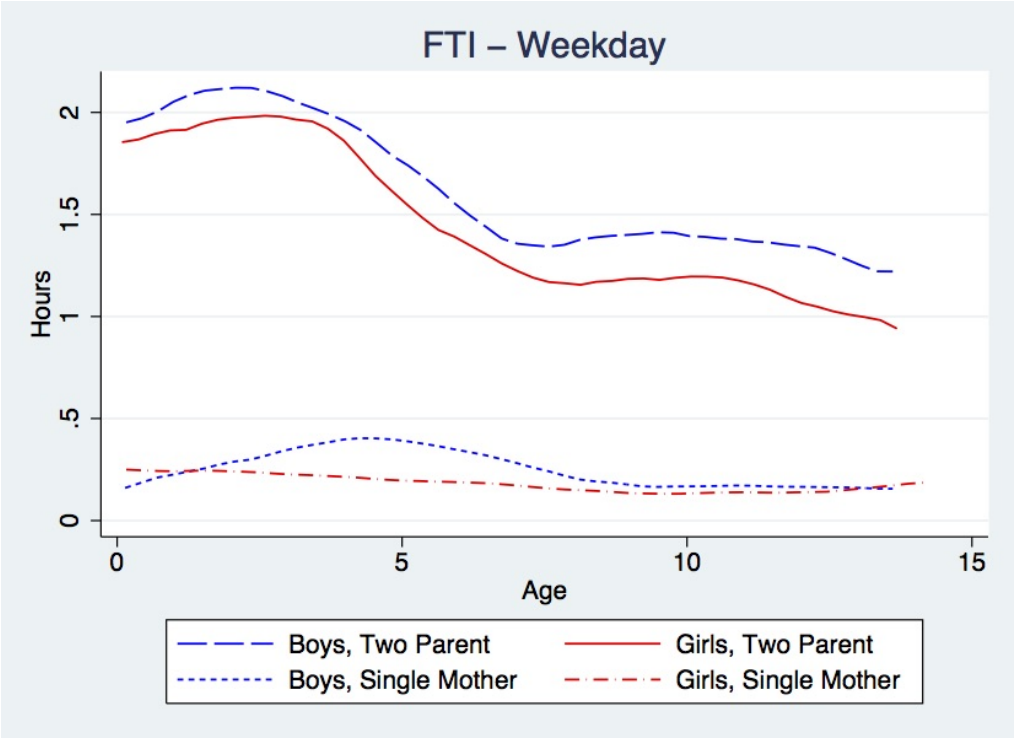


Figure C.4: FTI - Weekend

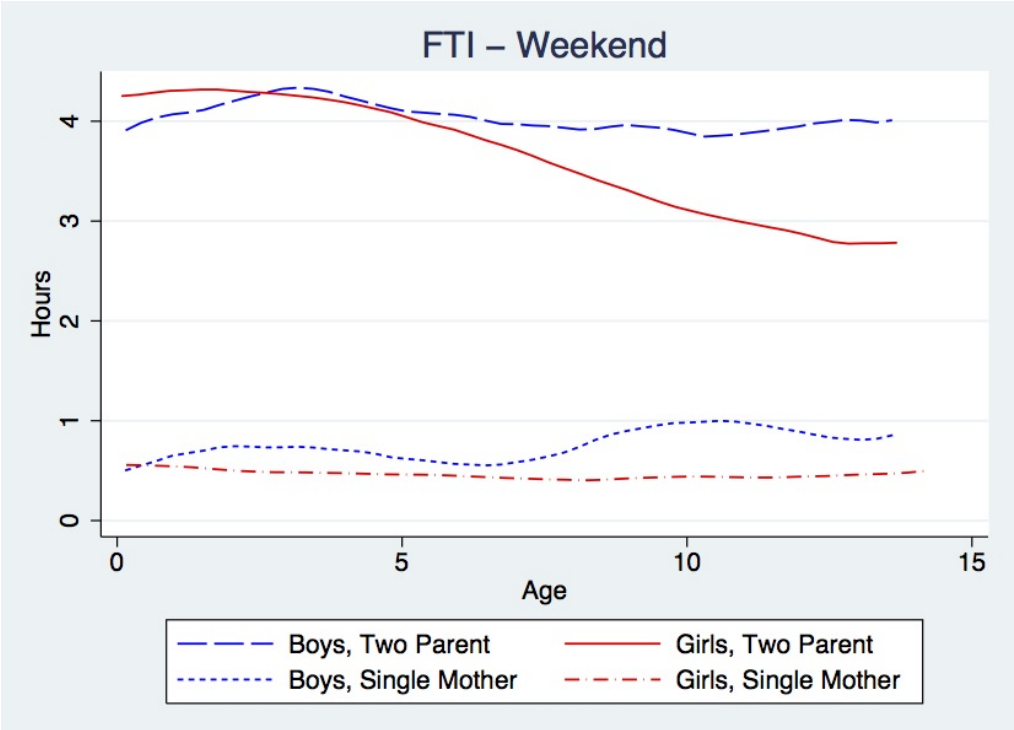


Figure C.5: Investments from Mother/Father - Boys

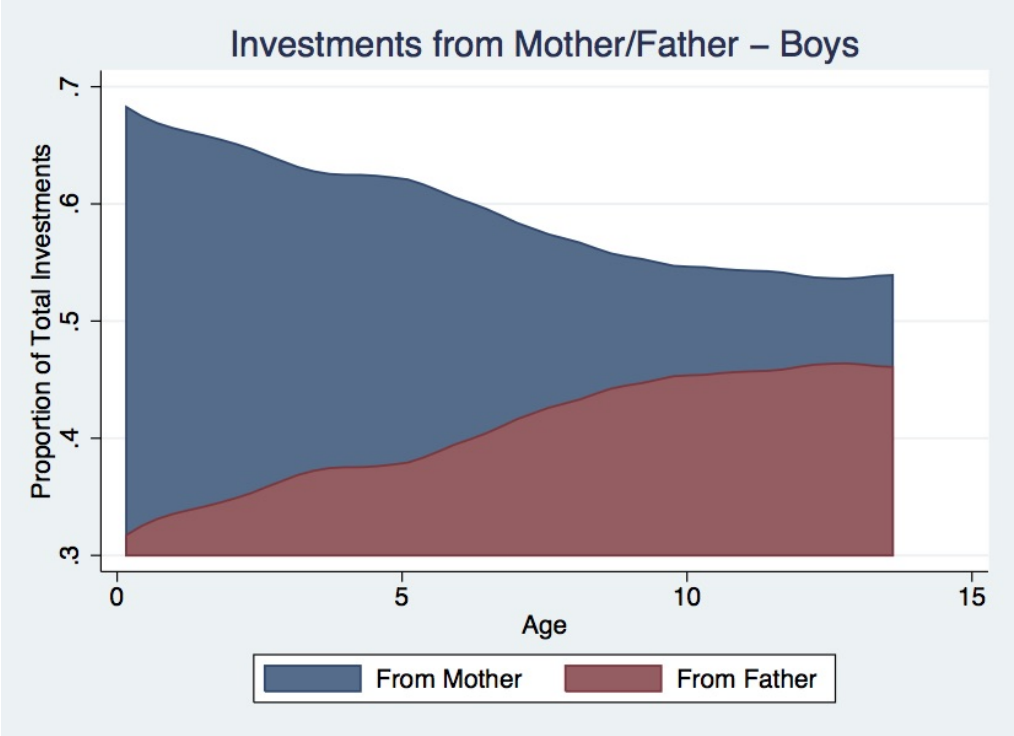
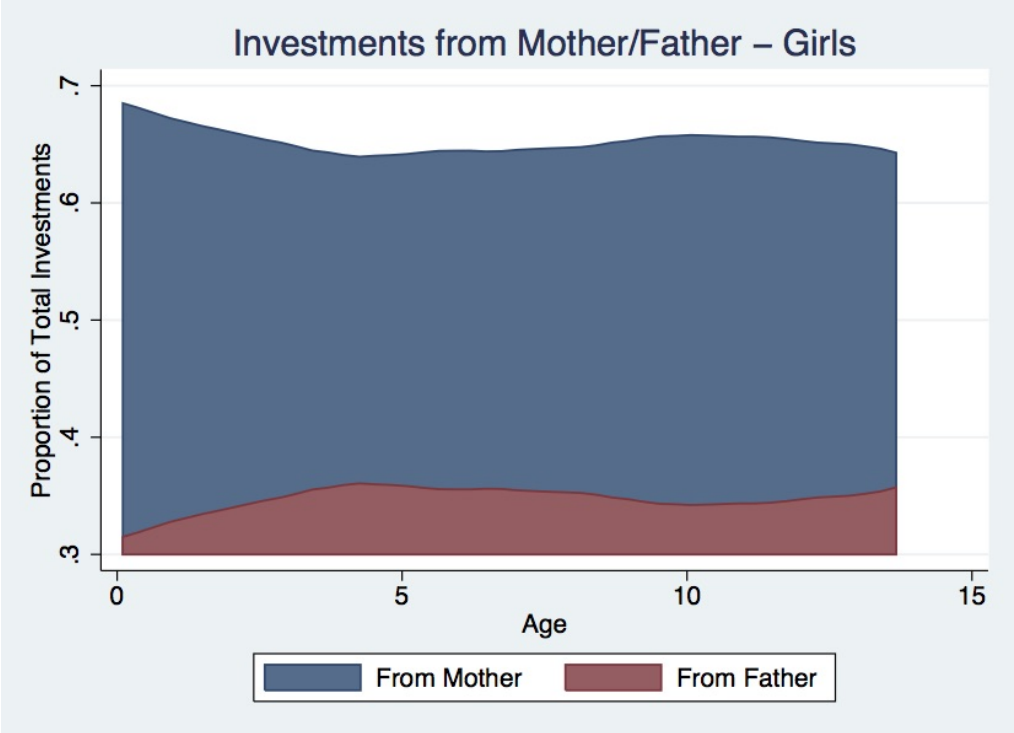


Figure C.6: Investments from Mother/Father - Girls



APPENDIX D

Tables for Chapter 2

Table D.1: Wave I Summary by Gender and Household Structure

	Boys-No Change		Boys-Change		Girls-No Change		Girls-Change	
	Both	**<2**	**Both**	**<2**	**Both**	**<2**	**Both**	**<2**
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Mother's Investments								
Total	24.802	18.847	25.568	18.116	26.608	21.694	27.455	21.946
	(15.206)	(15.226)	(15.264)	(17.029)	(15.486)	(19.368)	(15.495)	(18.872)
Weekday	3.073	2.289	3.100	2.238	3.199	2.705	3.301	2.803
	(2.415)	(2.291)	(2.372)	(2.554)	(2.477)	(3.074)	(2.566)	(2.606)
Weekend	4.719	3.701	5.035	3.464	5.306	4.085	5.476	3.964
	(3.206)	(3.266)	(3.156)	(3.368)	(3.192)	(3.504)	(3.195)	(3.681)
Father's Investments								
Total	16.371	3.531	16.771	5.056	15.061	2.218	13.246	5.160
	(11.397)	(7.945)	(13.147)	(9.518)	(11.440)	(6.621)	(10.919)	(8.650)
Weekday	1.618	0.351	1.872	0.498	1.493	0.241	1.382	0.586
	(1.615)	(1.047)	(2.125)	(1.269)	(1.712)	(0.927)	(1.697)	(1.123)
Weekend	4.141	0.887	3.705	1.284	3.797	0.506	3.167	1.116
	(3.284)	(2.139)	(3.001)	(2.717)	(3.158)	(1.686)	(2.910)	(2.112)
Age	6.533	7.288	4.617	6.283	6.848	7.366	4.931	5.540
	(3.850)	(3.626)	(3.149)	(3.214)	(3.847)	(3.774)	(3.430)	(3.429)
# Bio Sibs In HH	1.309	1.191	1.071	1.116	1.272	1.163	1.138	1.000
	(1.102)	(1.145)	(0.968)	(1.011)	(0.983)	(1.074)	(1.094)	(1.035)
White	0.630	0.257	0.490	0.279	0.672	0.258	0.478	0.292
Black	0.219	0.646	0.471	0.581	0.164	0.653	0.434	0.597
Hispanic	0.094	0.047	0.006	0.093	0.106	0.036	0.038	0.028
Stepmother								
In HH		0.036		0.035		0.014		0.028
Out of HH		0.090		0.105		0.093		0.042
Stepfather								
In HH		0.126		0.151		0.127		0.083
Out of HH		0.032		0.023		0.043		0.083
Parents in HH and Married	0.968		0.897		0.968		0.862	
Observations	754	444	155	86	720	418	159	72

*Notes: This table displays the averages by gender and household type with standard deviations in parentheses. Columns labeled No Change contain numbers for children who have no change in household structure at some point in the sample, and those labeled Change contain numbers for children who undergo some change throughout the sample period. Columns labeled Both display numbers for children who have both biological parents in the household during the first wave, and those labeled <2 are for those with less than both biological parents in the household during the first wave.

Table D.2: OLS Estimates of Gender Gaps in Time Investments

	Mother			Father		
	Total	**Weekday**	**Weekend**	**Total**	**Weekday**	**Weekend**
	[1]	[2]	[3]	[4]	[5]	[6]
Single Mother HH X						
Male	1.243	0.009	0.620	-7.567	-0.774	-1.791
	(0.963)	(0.158)	(0.214)	(0.743)	(0.102)	(0.209)
Female	1.139	0.058	0.430	-6.548	-0.699	-1.489
	(1.001)	(0.163)	(0.221)	(0.725)	(0.099)	(0.209)
Other	-11.513	-1.537	-1.881	-2.069	-0.135	-0.607
	(0.999)	(0.159)	(0.229)	(0.936)	(0.127)	(0.238)
Male - Female	0.103	-0.049	0.190	-1.019	-0.076	-0.302
(Standard Error)	(0.699)	(0.106)	(0.164)	(0.441)	(0.0596)	(0.126)
Observations	6,699	6,786	6,726	6,699	6,786	6,726
R-squared	0.280	0.233	0.163	0.269	0.157	0.229

Robust standard errors in parentheses

*Notes: This table displays estimates from OLS regressions of time investments on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. The fourth row displays the difference in the male and female interactions. Each regression includes controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, gender interactions with age and age-squared, and a male dummy variable. Standard errors are clustered by individual and displayed below the estimates.

Table D.3: Fixed Effects Estimates of Gender Gaps in Time Investments

Panel A: No Controls

	Mother			Father		
	Total	**Weekday**	**Weekend**	**Total**	**Weekday**	**Weekend**
	[1]	[2]	[3]	[4]	[5]	[6]
Single Mother HH X						
Male	-5.550	-0.707	-0.905	-10.346	-1.222	-2.077
	(1.229)	(0.186)	(0.266)	(0.933)	(0.131)	(0.232)
Female	-4.146	-0.509	-0.872	-8.425	-0.839	-2.032
	(1.388)	(0.209)	(0.302)	(0.767)	(0.114)	(0.200)
Other	-15.549	-2.023	-2.611	-6.057	-0.711	-1.291
	(1.451)	(0.203)	(0.332)	(1.116)	(0.151)	(0.301)
Male - Female	-1.405	-0.198	-0.033	-1.921	-0.383	-0.044
(Standard Error)	(1.799)	(0.270)	(0.390)	(1.166)	(0.167)	(0.297)
Observations	6,699	6,786	6,726	6,699	6,786	6,726
R-squared	0.031	0.022	0.019	0.061	0.037	0.041
Individuals	3,283	3,308	3,287	3,283	3,308	3,287

Panel B: With Controls

	Mother			Father		
	Total	**Weekday**	**Weekend**	**Total**	**Weekday**	**Weekend**
	[1]	[2]	[3]	[4]	[5]	[6]
Single Mother HH X						
Male	1.273	0.007	0.558	-7.306	-0.903	-1.332
	(1.713)	(0.274)	(0.358)	(1.235)	(0.173)	(0.318)
Female	2.442	0.214	0.562	-4.941	-0.503	-1.105
	(1.863)	(0.281)	(0.396)	(1.135)	(0.159)	(0.312)
Other	-7.433	-1.036	-1.109	-2.260	-0.342	-0.314
	(1.948)	(0.296)	(0.409)	(1.365)	(0.191)	(0.354)
Male - Female	-1.169	-0.207	-0.003	-2.364	-0.400	-0.227
(Standard Error)	(1.612)	(0.243)	(0.372)	(1.084)	(0.158)	(0.289)
Observations	6,699	6,786	6,726	6,699	6,786	6,726
R-squared	0.260	0.226	0.136	0.118	0.070	0.094
Individuals	3,283	3,308	3,287	3,283	3,308	3,287

Robust standard errors in parentheses

*Notes: This table displays estimates from fixed effects regressions of time investments on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. The fourth row of each panel displays the difference in the male and female interactions. Regressions in Panel A do not use any control variables. Each regression in Panel B includes controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.4: Gender Gaps in Investments by Initial HH Structure

	Both in Wave 1		<2 in Wave 1	
	Mother	**Father**	**Mother**	**Father**
	[1]	[2]	[3]	[4]
Single Mother HH X				
Male	1.948	-9.034	3.197	0.344
	(2.137)	(1.644)	(3.259)	(1.737)
Female	1.259	-4.800	6.738	0.530
	(2.167)	(1.543)	(3.682)	(1.487)
Other	-9.022	-2.865	-4.197	3.323
	(2.761)	(2.119)	(3.319)	(1.728)
Male - Female	0.688	-4.234	-3.541	-0.186
(Standard Error)	(2.028)	(1.540)	(2.876)	(1.668)
Observations	4,284	4,284	2,415	2,415
R-squared	0.288	0.146	0.219	0.068
Individuals	1,998	1,998	1,285	1,285

Robust standard errors in parentheses

*Notes: This table displays estimates from fixed effects regressions of total weekly time investments on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. Columns 1 and 2 restrict to the sample of children who were in a two-parent household in the first wave, and columns 3 and 4 restrict to those who were not living with both parents in the first wave. The fourth row displays the difference in the male and female interactions. Each regression controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.5: FE Gender Gaps in Probability of PTI>0

	Mother			Father		
	Total	Weekday	Weekend	Total	Weekday	Weekend
	[1]	[2]	[3]	[4]	[5]	[6]
Single Mother HH X						
Male	0.040	0.047	0.055	-0.419	-0.377	-0.331
	(0.032)	(0.043)	(0.044)	(0.054)	(0.054)	(0.054)
Female	0.068	0.107	0.041	-0.409	-0.297	-0.285
	(0.033)	(0.044)	(0.044)	(0.052)	(0.056)	(0.050)
Other	-0.435	-0.414	-0.364	-0.187	-0.094	-0.071
	(0.052)	(0.060)	(0.057)	(0.059)	(0.065)	(0.057)
Male - Female	-0.028	-0.060	0.014	-0.010	-0.080	-0.047
(Standard Error)	(0.033)	(0.039)	(0.044)	(0.049)	(0.050)	(0.049)
Observations	6,699	6,786	6,726	6,699	6,786	6,726
R-squared	0.152	0.155	0.133	0.202	0.117	0.178
Individuals	3,283	3,308	3,287	3,283	3,308	3,287

Robust standard errors in parentheses

*Notes: This table displays estimates from fixed effects regressions of a dummy variable for having a positive time investment on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. The fourth row displays the difference in the male and female interactions. Each regression controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.6: Gender Gaps in Probability of PTI>0 by Initial HH Structure

Panel A: Both Parents in HH in Wave 1

	Mother			Father		
	Total	Weekday	Weekend	Total	Weekday	Weekend
	[1]	[2]	[3]	[4]	[5]	[6]
Single Mother HH X						
Male	0.009	0.049	-0.007	-0.595	-0.507	-0.449
	(0.035)	(0.052)	(0.052)	(0.058)	(0.068)	(0.064)
Female	0.012	0.074	-0.023	-0.530	-0.397	-0.338
	(0.034)	(0.051)	(0.049)	(0.060)	(0.069)	(0.061)
Other	-0.452	-0.417	-0.432	-0.138	0.016	-0.042
	(0.081)	(0.092)	(0.075)	(0.079)	(0.096)	(0.081)
Male - Female	-0.003	-0.025	0.016	-0.065	-0.110	-0.110
(Standard Error)	(0.034)	(0.049)	(0.053)	(0.058)	(0.065)	(0.061)
Observations	4,284	4,323	4,293	4,284	4,323	4,293
R-squared	0.142	0.147	0.146	0.304	0.149	0.226
Individuals	1,998	2,006	1,996	1,998	2,006	1,996

Panel B: Less Than Two Parents in HH in Wave 1

	Mother			Father		
	Total	Weekday	Weekend	Total	Weekday	Weekend
	[1]	[2]	[3]	[4]	[5]	[6]
Single Mother HH X						
Male	0.126	0.066	0.125	0.004	-0.022	0.031
	(0.070)	(0.087)	(0.093)	(0.101)	(0.087)	(0.099)
Female	0.159	0.166	0.111	-0.090	-0.025	-0.035
	(0.076)	(0.094)	(0.091)	(0.098)	(0.097)	(0.088)
Other	-0.358	-0.393	-0.276	0.051	0.072	0.140
	(0.077)	(0.092)	(0.102)	(0.097)	(0.093)	(0.092)
Male - Female	-0.033	-0.101	0.013	0.094	0.0026	0.066
(Standard Error)	(0.076)	(0.081)	(0.084)	(0.089)	(0.089)	(0.084)
Observations	2,415	2,463	2,433	2,415	2,463	2,433
R-squared	0.172	0.173	0.129	0.102	0.080	0.112
Individuals	1,285	1,302	1,291	1,285	1,302	1,291

Robust standard errors in parentheses

*Notes: This table displays estimates from fixed effects regressions of a dummy variable for having a positive time investment on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. Panel A restricts to the sample of children who were in a two-parent household in the first wave, and Panel B restricts to those who were not living with both parents in the first wave. The fourth row of each panel displays the difference in the male and female interactions. Each regression controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.7: Gender Gaps by Age

Panel A: Mothers

	Mother - Total Weekly Investment			
	Ages 0 - 5	**Ages 6 - 10**	**Ages 11 - 15**	**16 and Over**
	[1]	[2]	[3]	[4]
Single Mother HH X				
Male	1.093	2.112	2.235	-2.451
	(2.096)	(1.815)	(1.833)	(2.175)
Female	2.486	2.725	3.295	-0.205
	(2.484)	(1.975)	(1.976)	(2.403)
Male - Female	-1.393	-0.612	-1.060	-2.246
(Standard Error)	(2.671)	(1.844)	(1.843)	(2.526)
Observations	6,699	6,699	6,699	6,699
Individuals	3,283	3,283	3,283	3,283

Panel B: Fathers

	Father - Total Weekly Investment			
	Ages 0 - 5	**Ages 6 - 10**	**Ages 11 - 15**	**16 and Over**
	[1]	[2]	[3]	[4]
Single Mother HH X				
Male	-9.701	-7.477	-5.321	-4.231
	(1.428)	(1.274)	(1.311)	(1.577)
Female	-9.290	-5.610	-1.991	-0.633
	(1.251)	(1.209)	(1.177)	(1.325)
Male - Female	-0.411	-1.867	-3.331	-3.598
(Standard Error)	(1.399)	(1.211)	(1.203)	(1.590)
Observations	6,699	6,699	6,699	6,699
Individuals	3,283	3,283	3,283	3,283

Robust standard errors in parentheses

*Notes: Each panel displays estimates from a fixed effects regression of total weekly time investments on child gender interacted with a dummy for being in a single-mother household for four different age groups. The omitted category is a household with both biological parents present. The third row of each panel displays the differences in the male and female interactions. Each regression controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.8: Gender Gaps by Race

	Mother		Father	
	White	Black	White	Black
	[1]	[2]	[3]	[4]
Single Mother HH X				
Male	-0.300	1.635	-9.461	-6.390
	(2.091)	(2.018)	(1.561)	(1.417)
Female	2.866	3.085	-5.462	-4.373
	(2.352)	(2.244)	(1.362)	(1.294)
Male - Female	-3.166	-1.450	-3.999	-2.017
(Standard Error)	(2.500)	(2.296)	(1.658)	(1.426)
Observations	6,699	6,699	6,699	6,699
Individuals	3,283	3,283	3,283	3,283

Robust standard errors in parentheses

*Notes: This table displays estimates from a fixed effects regression of total weekly time investments on child gender interacted with a dummy for being in a single-mother household by race. The omitted category is a household with both biological parents present. The third row of each panel displays the differences in the male and female interactions. Each regression controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.9: Gender Gaps by Activity Type

	Mother [1]	**Father** [2]
Passive Leisure	0.222	-0.723
	(0.749)	(0.461)
Active Leisure	-0.858	-0.753
	(0.565)	(0.426)
Entertainment	-0.060	-0.212
	(0.369)	(0.330)
Tending to Needs	0.100	-0.483
	(0.567)	(0.365)
Obtaining Goods and Services	-0.182	-0.195
	(0.445)	(0.265)
Household Activity	-0.175	-0.121
	(0.222)	(0.097)
Childcare	0.183	-0.005
	(0.171)	(0.015)
Observations	6,699	6,699
Individuals	3,283	3,283

Robust standard errors in parentheses

*Notes: Each estimate is from a fixed effects regression of weekly time investments for a specific activity category on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. Each estimate is from the difference in the male and female interaction terms. Each regression controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

Table D.10: Child Behavior and HH Structure

	Externalizing Behavior [1]	**Internalizing Behavior** [2]	**Positive Behavior** [3]
Single Mother HH X			
Male	-0.113	0.009	-0.008
	(0.134)	(0.136)	(0.129)
Female	0.160	0.016	-0.117
	(0.136)	(0.145)	(0.137)
Other	-0.163	-0.092	0.146
	(0.143)	(0.157)	(0.148)
Male - Female	-0.273	-0.007	0.110
(Standard Error)	(0.117)	(0.123)	(0.132)
Observations	6,060	6,058	6,078
R-squared	0.026	0.027	0.010
Number of Individuals	3,200	3,193	3,197

Robust standard errors in parentheses

*Notes: This table displays estimates from fixed effects regressions of parent rated child behaviors on child gender interacted with a dummy for being in a single-mother household. The omitted category is a household with both biological parents present. The fourth row of displays the difference in the male and female interactions. Each regression includes controls for the number of biological siblings in the household, the CDS wave, a marriage indicator for parents in the same household, dummy variables for having stepparents in/out of the household, and gender interactions with age and age-squared. Standard errors are clustered by individual and displayed below the estimates.

APPENDIX E

Tables for Chapter 3

Table E.1: Summary of Simulation Design

Element		Distribution(Mean, SD) Case 1	Case 2
Score base	$A_i t - 1$	N(0, 0.25)	
Teacher fixed effect	$\beta_{i,t}$	N(0.5, 0.25)	
Student fixed effect	c_i	N(0, 1)	N(0, 0.25)
Cohort fixed effect	cs_i	.	N(0, 0.25)
Random error	$\epsilon_i t$	N(0, 1)	

Table E.2: Average Standard Errors and Coverage Rates
(Student FE - N(0,1), no Cohort-School FE)

		Average of Teacher Average SE				Average Coverage Rate			
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Number of	Avg St Dev of	*None*	*Class*	*Cohort-school*	*School*	*None*	*Class*	*Cohort-school*	*School*
PANEL A - Sample contains 5 schools									
Random Grouping - Random Assignment									
1	0.419	0.435				0.956			
3	0.253	0.251	0.202	0.204	0.008	0.946	0.826	0.824	0.047
7	0.160	0.164	0.149	0.150	0.003	0.956	0.914	0.913	0.032
20	0.095	0.097	0.095	0.095	0.001	0.953	0.942	0.943	0.020
Dynamic Grouping - Random Assignment									
1	0.431	0.435				0.954			
3	0.257	0.251	0.202	0.204	0.010	0.944	0.820	0.818	0.064
7	0.159	0.164	0.149	0.150	0.004	0.957	0.908	0.908	0.045
20	0.099	0.097	0.095	0.095	0.002	0.947	0.935	0.933	0.023
Heterogeneity Grouping - Random Assignment									
1	0.888	0.399				0.546			
3	0.505	0.245	0.404	0.423	0.008	0.662	0.822	0.828	0.024
7	0.340	0.162	0.304	0.317	0.003	0.641	0.898	0.907	0.015
20	0.204	0.097	0.188	0.195	0.001	0.646	0.918	0.928	0.010
PANEL B - Sample contains 50 schools									
Random Grouping - Random Assigment									
1	0.454	0.446				0.947			
3	0.260	0.258	0.198	0.198	0.003	0.946	0.812	0.812	0.015
7	0.168	0.169	0.154	0.154	0.001	0.952	0.908	0.908	0.010
20	0.097	0.100	0.097	0.097	0.000	0.957	0.944	0.944	0.006
Dynamic Grouping - Random Assignment									
1	0.451	0.446				0.948			
3	0.254	0.258	0.198	0.199	0.004	0.954	0.823	0.822	0.023
7	0.165	0.169	0.153	0.154	0.001	0.955	0.914	0.914	0.014
20	0.101	0.100	0.097	0.097	0.001	0.947	0.934	0.934	0.008
Heterogeneity Grouping - Random Assignment									
1	0.894	0.410				0.586			
3	0.518	0.251	0.404	0.406	0.003	0.647	0.812	0.812	0.008
7	0.335	0.167	0.309	0.310	0.001	0.668	0.906	0.907	0.005
20	0.201	0.099	0.194	0.194	0.000	0.663	0.931	0.932	0.003

*Notes: For all simulations, lambda is set to 0.5. The teacher effects are estimated using pooled dynamic OLS. Columns 3 - 6 display the average standard errors by clustering type. The types of clustering are no clustering, classroom level clustering, cohort-school level clustering, and school level clustering. Columns 7 - 10 display the average coverage rate for each type of clustering.

Table E.3: Average Standard Errors and Coverage Rates
(Student FE - N(0,0.25), Cohort-School FE - N(0,0.25))

		Average of Teacher Average SE				Average Coverage Rate			
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Number of	Avg St Dev of	*None*	*Class*	*Cohort-school*	*School*	*None*	*Class*	*Cohort-school*	*School*
Random Grouping - Random Assignment									
1	0.438	0.317				0.838			
3	0.258	0.186	0.213	0.208	0.006	0.846	0.835	0.821	0.032
7	0.173	0.123	0.160	0.156	0.002	0.838	0.905	0.900	0.025
20	0.100	0.073	0.100	0.097	0.001	0.840	0.942	0.939	0.013
Dynamic Grouping - Random Assignment									
1	0.448	0.318				0.831			
3	0.263	0.187	0.213	0.208	0.008	0.832	0.829	0.819	0.044
7	0.173	0.123	0.158	0.154	0.003	0.843	0.906	0.898	0.033
20	0.102	0.073	0.101	0.098	0.001	0.836	0.940	0.936	0.019
Heterogeneity Grouping - Random Assignment									
1	0.452	0.317				0.821			
3	0.264	0.186	0.215	0.211	0.006	0.828	0.828	0.814	0.036
7	0.177	0.123	0.160	0.156	0.003	0.830	0.899	0.891	0.024
20	0.103	0.073	0.101	0.099	0.001	0.836	0.937	0.931	0.012

*Notes: For all simulations, lambda is set to 0.5. The teacher effects are estimated using pooled dynamic OLS. Columns 3 - 6 display the average standard errors by clustering type. The types of clustering are no clustering, classroom level clustering, cohort-school level clustering, and school level clustering. Columns 7 - 10 display the average coverage rate for each type of clustering.

Table E.4: Average Standard Errors and Coverage Rates, Cohort-by-cohort Estimation

[1] Number of Cohorts	[2] Avg St Dev of Estimates	[3] Avg of Teacher Avg SE	[4] Average Coverage Rate
Panel A: Student FE - N(0,1), no Cohort-School FE			
Random Grouping - Random Assignment			
3	0.253	0.228	0.949
7	0.163	0.157	0.955
20	0.095	0.096	0.959
Dynamic Grouping - Random Assignment			
3	0.258	0.227	0.944
7	0.165	0.157	0.944
20	0.099	0.097	0.953
Heterogeneity Grouping - Random Assignment			
3	0.506	0.470	0.934
7	0.332	0.330	0.951
20	0.204	0.198	0.953
Panel B: Student FE - N(0,0.25), Cohort-School FE - N(0,0.25)			
Random Grouping - Random Assignment			
3	0.258	0.233	0.951
7	0.175	0.167	0.952
20	0.100	0.099	0.957
Dynamic Grouping - Random Assignment			
3	0.263	0.235	0.952
7	0.172	0.167	0.953
20	0.102	0.099	0.951
Heterogeneity Grouping - Random Assignment			
3	0.264	0.237	0.951
7	0.176	0.168	0.949
20	0.103	0.100	0.952

*Notes: For all simulations, lambda is set to 0.5. The teacher effects are estimated using pooled OLS.

Table E.5: Average Standard Errors and Coverage Rates
(Student FE N(0,1) , no Cohort-School FE)

		Average of Teacher Average SE				Average Coverage Rate			
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Number of	Avg St Dev of	*None*	*Class*	*Cohort-school*	*School*	*None*	*Class*	*Cohort-school*	*School*
Dynamic Grouping - Positive Assignment									
1	0.437	0.435				0.953			
3	0.246	0.252	0.200	0.203	0.015	0.954	0.836	0.833	0.089
7	0.157	0.165	0.151	0.152	0.009	0.966	0.926	0.926	0.089
20	0.096	0.097	0.094	0.095	0.005	0.951	0.935	0.937	0.085
Heterogeneity Grouping - Positive Assignment									
1	0.417	0.399				0.547			
3	0.239	0.231	0.192	0.194	0.007	0.388	0.346	0.349	0.014
7	0.156	0.151	0.144	0.145	0.003	0.374	0.365	0.366	0.014
20	0.093	0.089	0.090	0.090	0.001	0.376	0.376	0.376	0.006

*Notes: For all simulations, lambda is set to 0.5. The teacher effects are estimated using pooled dynamic OLS. Columns 3 - 6 display the average standard errors by clustering type. The types of clustering are no clustering, classroom level clustering, cohort-school level clustering, and school level clustering. Columns 7 - 10 display the average coverage rate for each type of clustering.

Table E.6: Average Standard Errors and Coverage Rates,
Cohort-by-cohort Estimation

[1]	[2]	[3]	[4]
Number of Cohorts	Avg St Dev of Estimates	Avg of Teacher Avg SE	Average Coverage Rate
Dynamic Grouping - Positive Assignment			
3	0.246	0.227	0.959
7	0.164	0.158	0.949
20	0.097	0.096	0.954
Heterogeneity Grouping - Positive Assignment			
3	0.239	0.217	0.621
7	0.160	0.152	0.382
20	0.093	0.091	0.385

*Notes: For all simulations, lambda is set to 0.5. The teacher effects are estimated using pooled OLS.

Table E.7: Student Characteristics, by District

	A	**B**	**C**	**D**	**E**	**F**
Math score	1557.70	1575.64	1497.20	1491.17	1539.38	1523.40
Lagged Math score	1423.54	1464.70	1355.97	1354.15	1420.15	1385.79
Days absent	7.19	6.72	6.52	7.56	7.18	6.97
Asian (%)	0.03	0.02	0.01	0.03	0.03	0.04
Black (%)	0.35	0.10	0.27	0.42	0.20	0.28
Hispanic (%)	0.24	0.05	0.60	0.05	0.24	0.24
American Indian (%)	0.00	0.00	0.00	0.00	0.00	0.00
Multi-racial (%)	0.03	0.03	0.01	0.04	0.05	0.03
Other race/ethnicity (%)	0.03	0.03	0.02	0.04	0.06	0.04
Female (%)	0.49	0.50	0.50	0.51	0.50	0.49
Disability (%)	0.14	0.16	0.11	0.11	0.15	0.15
LEP (%)	0.20	0.01	0.51	0.06	0.19	0.20
FRL (%)	0.45	0.21	0.71	0.53	0.50	0.53
Num. students	104,019	12,805	138,913	48,592	67,071	61,043

*Notes: LEP refers to student classified as having Limited English Proficiency, and FRL refers to students eligible for free- or reduced-price lunch. The number of students is the total number of students in grade 4 during years 2001-2007.

Table E.8: Percent of 95% Confidence Intervals Above/Below Cutoffs, By Region of Value-Added Distribution

	Percent of 95% CI Upper Bounds Under Percentile				Percent of 95% CI Lower Bounds Over Percentile			
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Method	*Under 10th*	*Under 25th*	*Under 75th*	*Under 90th*	*Over 10th*	*Over 25th*	*Over 75th*	*Over 90th*
Panel A: Bottom 10%								
Pooled OLS-Lag								
No clustering	0.074	0.441	1.000	1.000				
Cohort-school clustering	0.235	0.779	1.000	1.000				
School Clustering	0.706	0.971	1.000	1.000				
OLS-Lag cohort-by-cohort	0.008	0.016	0.210	0.290				
Panel B: 25th - 75th percentile								
Pooled OLS-Lag								
No clustering			0.241	0.759	0.812	0.264		
Cohort-school clustering			0.507	0.896	0.902	0.540		
School Clustering			0.890	0.994	0.991	0.889		
OLS-Lag cohort-by-cohort			0.042	0.169	0.180	0.058		
Panel C: Top 10%								
Pooled OLS-Lag								
No clustering					1.000	1.000	0.600	0.200
Cohort-school clustering					0.992	0.983	0.758	0.333
School Clustering					1.000	1.000	0.975	0.792
OLS-Lag cohort-by-cohort					0.589	0.452	0.185	0.081

*Notes: These percentages are averages taken over the six districts in from Table E.7, and include 432,443 fourth grade students and 9,102 fourth grade teachers from seven cohorts.

APPENDIX F

Figures for Chapter 3

Figure F.1: Average Standard Errors, by District

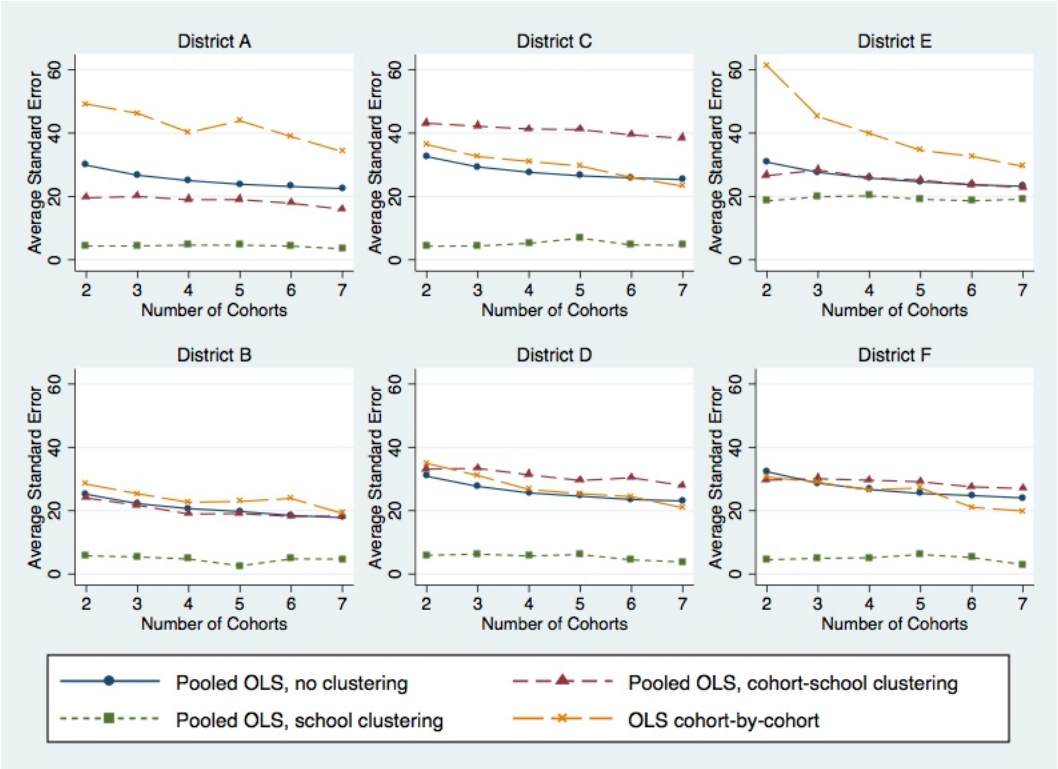
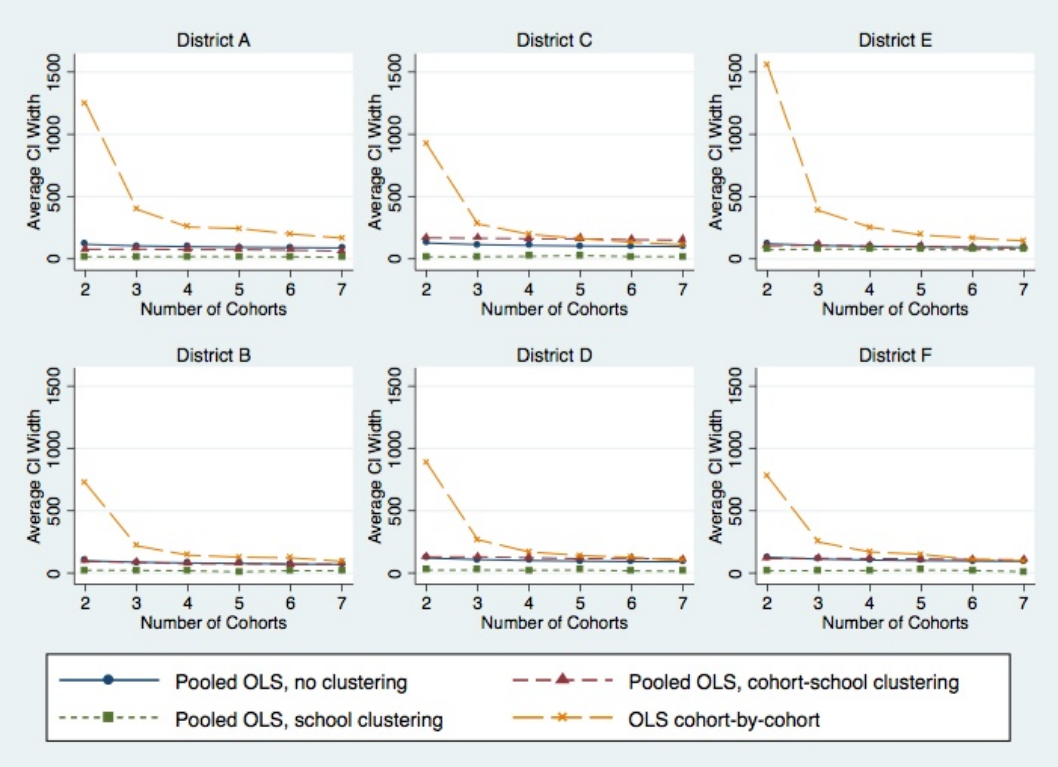


Figure F.2: Average Confidence Interval Widths, by District



APPENDIX G

Supplemental Tables for Chapter 1

Table G.1: Impact of Dual Language Education - Constant Effect

Panel A: English Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.492***	0.209*		0.120	
			(0.053)	(0.115)		(0.077)	
Attended (K/First)	-0.016	-0.127*			0.426*		0.245
	(0.100)	(0.063)			(0.231)		(0.155)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	1,472	1,472	1,472	1,472	1,472	1,472	1,472
Number of lotfe	44	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.653***	0.296**		0.226	
			(0.063)	(0.136)		(0.138)	
Attended (K/First)	0.266**	0.348***			0.453**		0.346*
	(0.111)	(0.109)			(0.199)		(0.193)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	809	809	809	809	809	809	809
Number of Clusters	36	36	36	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in kindergarten or first grade. The treatment and attendance variables are not interacted with years of treatment in this specification. Standard errors are clustered by lottery.

Table G.2: Impact of Dual Language Education - 3rd Grade Attendance Measure

Panel A: English Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.293***	0.209*		0.120	
			(0.062)	(0.115)		(0.077)	
Attended (3rd)	0.043	-0.012			0.715*		0.411
	(0.142)	(0.094)			(0.414)		(0.274)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	1,472	1,472	1,472	1,472	1,472	1,472	1,472
Number of lotfe	44	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.560***	0.296**		0.226	
			(0.067)	(0.136)		(0.138)	
Attended (3rd)	0.278**	0.347**			0.528**		0.403*
	(0.136)	(0.163)			(0.248)		(0.233)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	809	809	809	809	809	809	809
Number of clusters	36	36	36	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in third grade. The treatment and attendance variables are not interacted with years of treatment in this specification. Standard errors are clustered by lottery.

Table G.3: Grades Three Through Five Only

Panel A: English Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.503***	0.034		0.023*	
			(0.054)	(0.024)		(0.014)	
Attend DL School	-0.001	-0.023			0.068		0.046*
	(0.022)	(0.013)			(0.045)		(0.027)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	1,172	1,172	1,172	1,172	1,172	1,172	1,172
Number of Clusters	44	44	44	44	44	44	44

Panel B: ESL/LEP Sample

	OLS			Math		Reading	
	Math	Reading	First Stage	ITT	LATE	ITT	LATE
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Won First Choice			0.652***	0.055*		0.028	
			(0.061)	(0.031)		(0.030)	
Attend DL School	0.049**	0.058**			0.084*		0.043
	(0.023)	(0.022)			(0.047)		(0.043)
Neighborhood School FE	X	X	X	X	X	X	X
Observations	623	623	623	623	623	623	623
Number of Clusters	36	36	36	36	36	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with years of treatment (grade plus one). Standard errors are clustered by lottery.

Table G.4: Cohort Interactions

	English Sample		ESL/LEP Sample	
	Math	Reading	Math	Reading
	[1]	[2]	[3]	[4]
Attend DL School				
2007 Cohort	0.401**	0.200*	0.105***	0.133***
	(0.157)	(0.116)	(0.038)	(0.031)
2008 Cohort	-0.009	-0.014	0.048	-0.012
	(0.035)	(0.030)	(0.063)	(0.028)
2009 Cohort	0.074	0.132**	0.032	0.045
	(0.109)	(0.058)	(0.050)	(0.086)
2010 Cohort	0.151**	0.153***	0.286	0.239
	(0.069)	(0.040)	(0.197)	(0.166)
2011 Cohort	0.131***	0.029	0.067	0.061
	(0.044)	(0.034)	(0.086)	(0.096)
Neighborhood School FE	X	X	X	X
Observations	1,471	1,471	809	809
Number of Lottery FE	44	44	36	36

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1
*Notes: Each regression includes lottery fixed effects (priority-year-program) as well as controls for female, race, frpl-year, exceptionality, grade of exam, year of exam, and neighborhood school fixed effects. Reported coefficients are on interactions between the attendance variable and cohort. Attendance is measured by whether the student attended a DL school in kindergarten or first grade and interacted with years of treatment (grade plus one). Standard errors are clustered by lottery.

Table G.5: Attrition and Weighting (Panel)

Panel A: Summary of Probabilities of Testing

	Full Sample		English Sample		ESL/LEP Sample	
	Winners	**Losers**	**Winners**	**Losers**	**Winners**	**Losers**
	[1]	[2]	[3]	[4]	[5]	[6]
Average Pr(Test)	0.853	0.832	0.843	0.825	0.878	0.841
SD Pr(Test)	0.022	0.029	0.032	0.047	0.035	0.045
APE	0.019		0.011		0.027	
(SE)	(0.029)		(0.033)		(0.038)	
N	1532	1433	1105	816	427	617

Panel B: Non-Random Attrition

	Coefficients on Indicator for Winning Lottery		
	Full Sample	**English Sample**	**ESL/LEP Sample**
	[1]	[2]	[3]
No Controls	0.021	0.018	0.037
(SE)	(0.022)	(0.033)	(0.032)
Controls + Lottery FE	0.021	-0.009	0.049
(SE)	(0.031)	(0.038)	(0.038)
+ Neighborhood School FE	0.030	-0.003	0.057
(SE)	(0.032)	(0.046)	(0.042)
N	2965	1921	1044

*Notes: Panel A summarizes estimated probabilites of having test scores available. This table uses an expanded dataset, relative to Table 6, where each student has an observation for each grade that they could have tested in if they passed each grade. The estimates are based on logit regressions including gender, race, frpl-year, and an indicator for winning the lottery. Panel B shows estimated coefficients from OLS regressions of having at least one set of test scores available on winning the lottery. The baseline controls are gender, race, and frpl-year. The second set of OLS estimates also condition on lottery fixed effects, and the third set include neighborhood school fixed effects.

APPENDIX H

Supplemental Tables for Chapter 2

Table H.1: FE Estimates of Gender Gaps with Day of Week FEs

	Mother			Father		
	Total	Weekday	Weekend	Total	Weekday	Weekend
	[1]	[2]	[3]	[4]	[5]	[6]
Single-Mother HH X						
Male	1.250	0.010	0.539	-7.339	-0.901	-1.357
	(1.711)	(0.273)	(0.359)	(1.245)	(0.175)	(0.317)
Female	2.370	0.211	0.535	-5.131	-0.509	-1.139
	(1.865)	(0.279)	(0.396)	(1.142)	(0.160)	(0.310)
Other	-7.493	-1.044	-1.144	-2.311	-0.345	-0.358
	(1.941)	(0.293)	(0.408)	(1.381)	(0.193)	(0.353)
Male - Female	-1.120	-0.200	0.004	-2.207	-0.392	-0.218
(Standard Error)	(1.617)	(0.244)	(0.372)	(1.084)	(0.159)	(0.288)
Observations	6,699	6,786	6,726	6,699	6,786	6,726
R-squared	0.263	0.230	0.138	0.126	0.078	0.099
Number of Individuals	3,283	3,308	3,287	3,283	3,308	3,287

*Notes: This is a replication of the main fixed effects estimates from Table 3, Panel B, but includes indicators for day of the week. For the total weekly time regression, there is a dummy included for each weekday-weekend day combination, with one combination excluded. Standard errors are clustered by individual and displayed below the estimates.

Table H.2: Gender Gaps in Investments
by Initial HH Structure (Day of Week FEs)

	Both in Wave 1		**<2 in Wave 1**	
	Mother	**Father**	**Mother**	**Father**
	[1]	[2]	[3]	[4]
Single-Mother HH X				
Male	1.849	-9.417	3.246	0.329
	(2.145)	(1.642)	(3.252)	(1.725)
Female	1.192	-5.289	6.861	0.483
	(2.162)	(1.549)	(3.743)	(1.465)
Other	-9.097	-3.034	-4.231	3.302
	(2.700)	(2.161)	(3.349)	(1.724)
Male - Female	0.657	-4.128	-3.615	-0.154
(Standard Error)	(2.037)	(1.534)	(2.896)	(1.642)
Observations	4,284	4,284	2,415	2,415
R-squared	0.292	0.160	0.224	0.080
Number of Individuals	1,998	1,998	1,285	1,285

Robust standard errors in parentheses

*Notes: This is a replication of the main fixed effects estimates by initial household structure from Table 4, but includes indicators for day of the week. For the total weekly time regression, there is a dummy included for each weekday-weekend day combination, with one combination excluded. Standard errors are clustered by individual and displayed below the estimates.

APPENDIX I

Supplemental Tables for Chapter 3

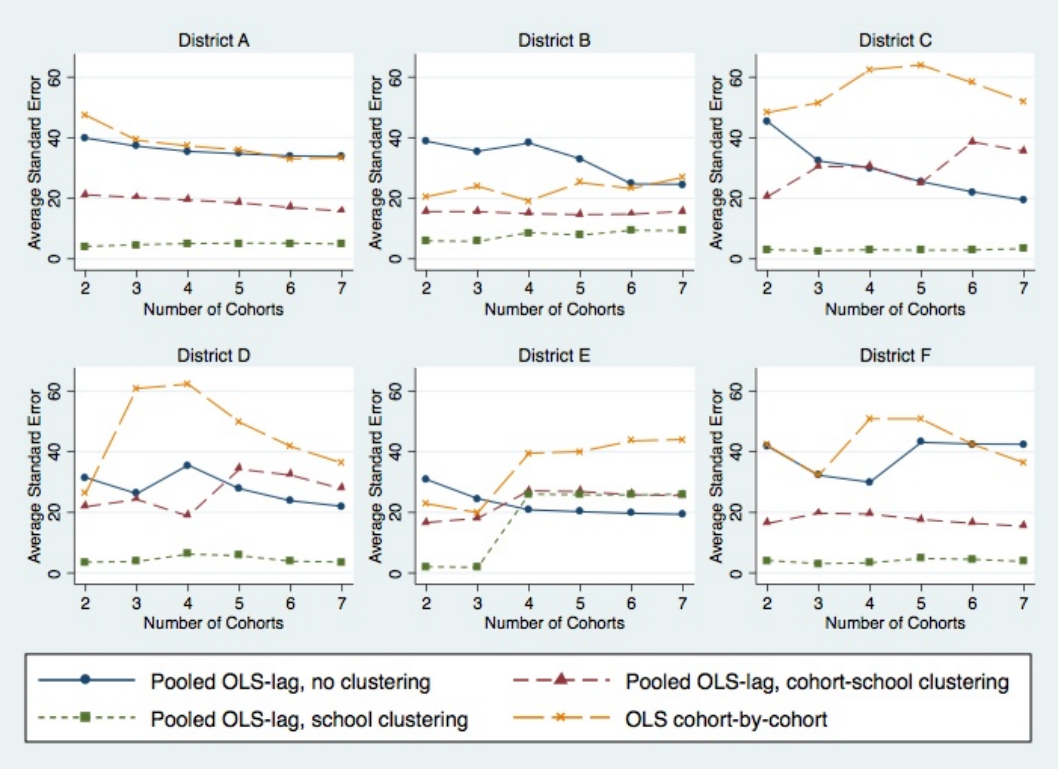
Table I.1: Estimation Sample Sizes and Graph Sample Sizes

	Estimation sample		Graph sample - Alt. 1		Graph sample - Alt. 2	
[1]	[2]	[3]	[4]	[5]	[6]	[7]
Cohorts	*Teachers*	*Students*	*Teachers*	*Students*	*Teachers*	*Students*
District A						
2	884	28,794	435	19,119	113	5,167
3	1,089	43,014	315	20,361	113	7,609
4	1,373	59,105	241	20,584	113	9,887
5	1,562	74,083	195	20,543	113	12,053
6	1,798	89,853	137	17,126	113	14,217
7	1,955	104,019	113	16,267	113	16,267
District B						
2	105	3,352	51	2,261	16	742
3	133	4,997	37	2,452	16	1,096
4	153	6,652	28	2,402	16	1,401
5	184	8,607	24	2,566	16	1,718
6	221	10,727	19	2,350	16	2,009
7	256	12,805	16	2,344	16	2,344
District C						
2	1,170	41,453	589	28,602	129	6,565
3	1,338	58,494	346	26,634	129	10,254
4	1,623	79,141	262	27,405	129	13,736
5	1,947	99,727	204	26,073	129	16,754
6	2,279	120,314	160	23,904	129	19,435
7	2,580	138,913	129	21,966	129	21,966
District D						
2	497	15,936	269	11,373	43	1,846
3	617	22,758	174	10,758	43	2,672
4	739	29,476	126	10,112	43	3,477
5	869	36,217	91	8,976	43	4,231
6	985	42,841	67	7,744	43	4,954
7	1,110	48,592	43	5,645	43	5,645
District E						
2	590	18,176	278	11,891	57	2,488
3	743	27,428	190	12,162	57	3,706
4	940	36,589	127	10,639	57	4,824
5	1,110	46,196	92	9,465	57	5,908
6	1,288	56,531	75	9,218	57	6,985
7	1,523	67,071	57	7,969	57	7,969
District F						
2	660	17,835	323	11,883	70	2,572
3	761	26,086	228	12,728	70	3,883
4	933	35,267	172	12,936	70	5,201
5	1,096	44,179	120	11,277	70	6,496
6	1,263	52,739	84	9,388	70	7,787
7	1,408	61,043	70	8,904	70	8,904

APPENDIX J

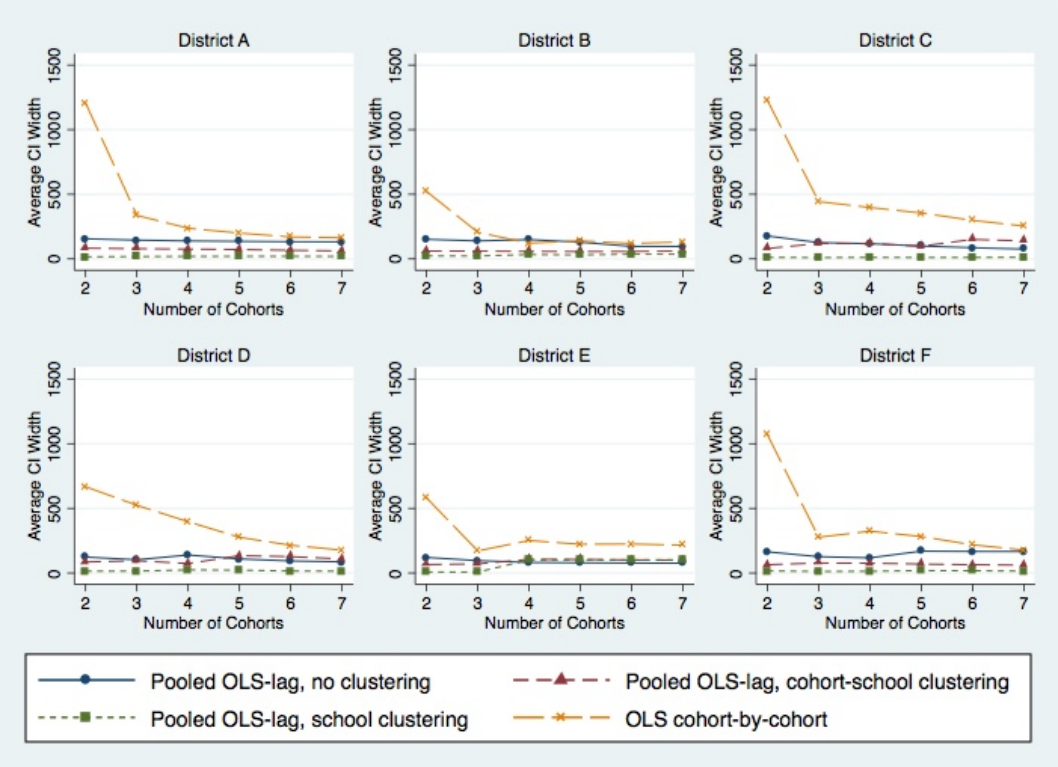
Supplemental Figures for Chapter 3

Figure J.1: Average Standard Errors, by District



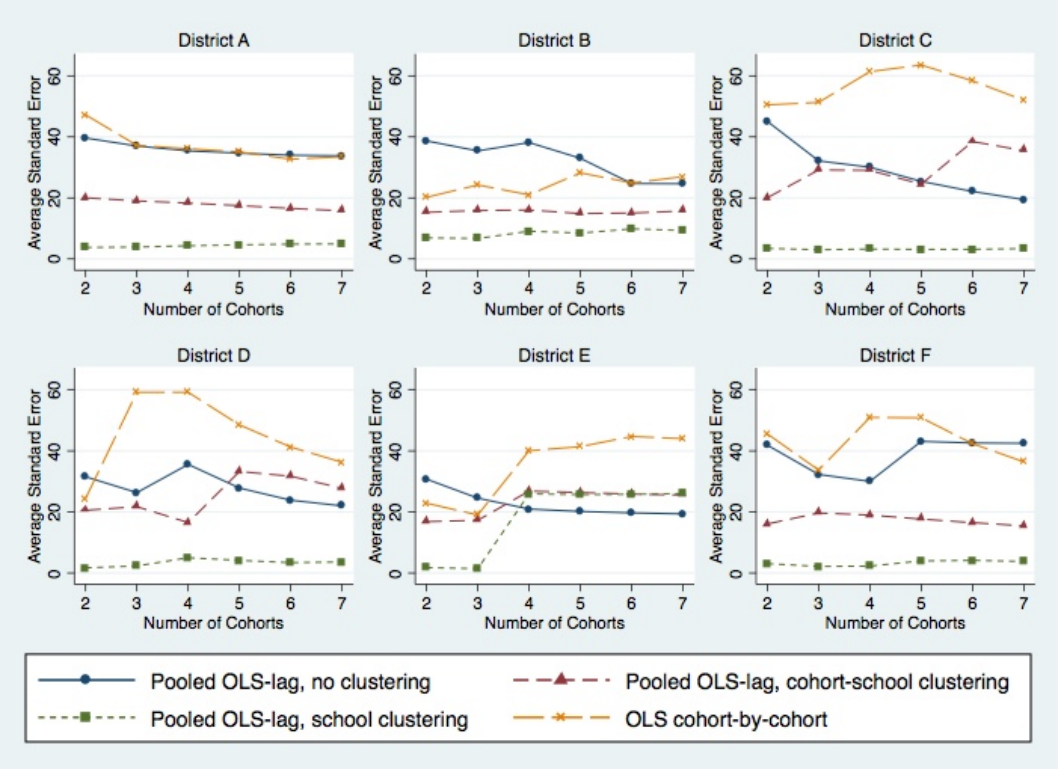
Notes: For each district, teacher effects and standard errors are calculated using an estimation sample of all 4th grade teachers and their students in cohorts $\leq$ c. For each graph data point corresponding to c cohorts, the standard errors are averaged over only the subsample of teachers with exactly c cohorts. The number of teachers in each estimation sample and graph sample are listed in Appendix Table I.1.

Figure J.2: Average Confidence Interval Widths, by District



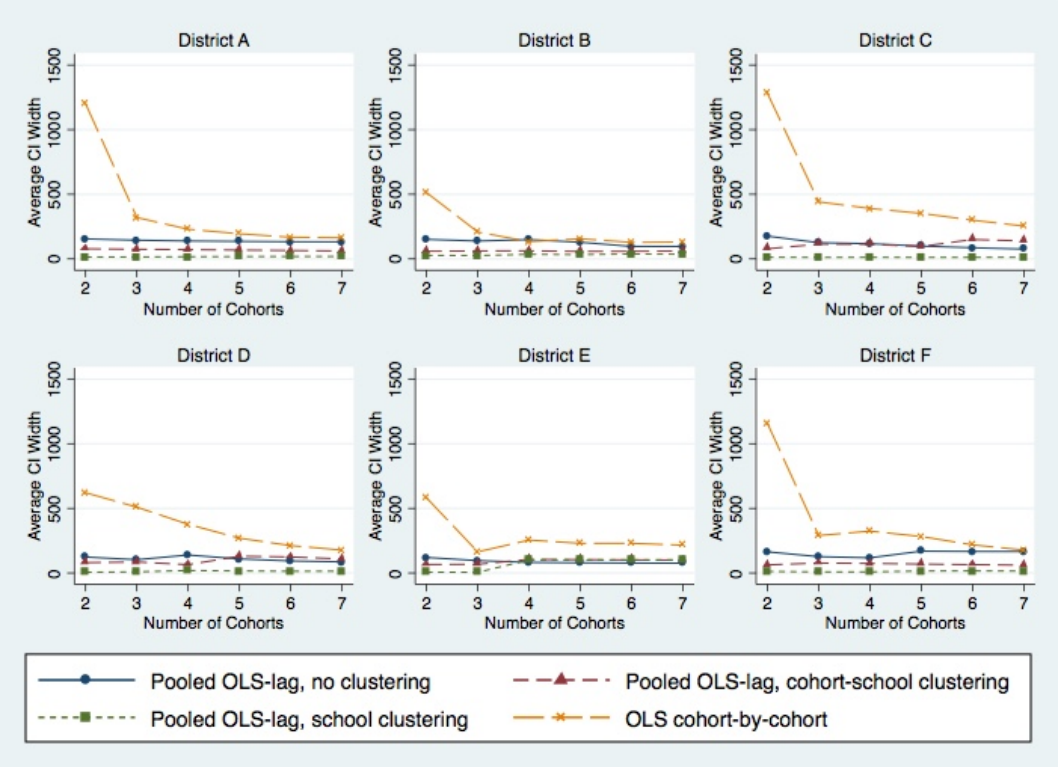
Notes: For each district, teacher effects and standard errors (and confidence interval widths) are calculated using an estimation sample of all 4th grade teachers and their students in cohorts $\leq$ c. For each graph data point corresponding to c cohorts, the confidence interval widths are averaged over only the subsample of teachers with exactly c cohorts. The number of teachers in each estimation sample and graph sample are listed in Appendix Table I.1.

Figure J.3: Average Standard Errors, by District



Notes: For each district, teacher effects and standard errors are calculated using an estimation sample of all 4th grade teachers and their students in cohorts $\leq$ c. For each graph data point corresponding to c cohorts, the standard errors are averaged over only the subsample of teachers with exactly 7 cohorts. The number of teachers in each estimation sample and graph sample are listed in Appendix Table I.1.

Figure J.4: Average Confidence Interval Widths, by District



Notes: For each district, teacher effects and standard errors (and confidence interval widths) are calculated using an estimation sample of all 4th grade teachers and their students in cohorts $\leq$ c. For each graph data point corresponding to c cohorts, the confidence interval widths are averaged over only the subsample of teachers with exactly 7 cohorts. The number of teachers in each estimation sample and graph sample are listed in Appendix Table I.1.

REFERENCES

REFERENCES

Daniel Aaronson, Lisa Barrow, and William Sander. Teachers and student achievement in the chicago public high schools. *Journal of labor Economics*, 25(1):95–135, 2007.

Tracy Packiam Alloway. *Improving working memory: Supporting students' learning*. Sage, 2010.

Tahir Andrabi, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics*, 3(3): 29–54, 2011.

Alan Baddeley. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839, 2003.

Alan D Baddeley and Graham J Hitch. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

Michael Baker and Kevin Milligan. Boy-girl differences in parental time investments: Evidence from three countries. *National Bureau of Economic Research (No. w18893)*, 2013.

Dale Ballou. Test scaling and value-added measurement. *Education*, 4(4):351–383, 2009.

Dale Ballou, William Sanders, and Paul Wright. Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics*, 29(1):37–65, 2004.

Mokher Christine G Ballou, Dale and Linda Cavalluzzo. Using value-added assessment for personnel decisions: How omitted variables and model specification influence teachers' outcomes. *Unpublished Manuscript*, 2012.

Marianne Bertrand and Jessica Pan. The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64, 2013.

Derek C Briggs and Jonathan P Weeks. The sensitivity of value-added modeling to the creation of a vertical score scale. *Education*, 4(4):384–414, 2009.

M Cazabon, W Lambert, and G Hall. Two-way bilingual education: A report on the Amigos Program. *Washington, DC: Center for Applied Linguistics*, 1999.

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *The American Economic Review*, 104(9):2593–2632, 2014.

Rosa Minhyo Cho. Are there peer effects associated with having english language learner (ell) classmates? evidence from the early childhood longitudinal study kindergarten cohort (ecls-k). *Economics of Education Review*, 31(5):629–643, 2012.

Elisavet Chrysochoou, Zoe Bablekou, and Nikokaos Tsigilis. Working memory contributions to reading comprehension components in middle childhood children. *The American journal of psychology*, 124(3):275–289, 2011.

Brian Cobb, Diego Vega, and Cindy Kronauge. Effects of an elementary dual language immersion school program on junior high achievement. *Middle Grades Research Journal*, 1(1):27–48, 2009.

Scott Condie, Lars Lefgren, and David Sims. Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, 40:76–92, 2014.

Sean P Corcoran, Jennifer L Jennings, and Andrew A Beveridge. Teacher effectiveness on high-and low-stakes tests. *Society for Research on Educational Effectiveness*, 2011.

Daniela Del Boca and Anna Laura Mancini. Parental time and child outcomes: Does gender matter? *Bank of Italy Occasional Paper*, (187), 2013.

Daniela Del Boca, Chiara Monfardini, Cheti Nicoletti, et al. Children‘s and parent‘s time-use choices and cognitive development during adolescence. *Human Capital and Economic Opportunity Working Group working paper*, 6, 2012.

David J Deming, Justine S Hastings, Thomas J Kane, and Douglas O Staiger. School Choice, School Quality, and Postsecondary Attainment. 104(3):991–1013, 2014.

Stephen G Donald and Kevin Lang. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233, 2007.

Mia Dufva, Pekka Niemi, and Marinus JM Voeten. The role of phonological memory, word recognition, and comprehension skills in reading development: From preschool to grade 2. *Reading and Writing*, 14(1-2):91–117, 2001.

Charlotte Geay, Sandra Mcnally, and Shqiponja Telhaj. Non-native speakers of English in the classroom: What are the effects on pupil performance? *Economic Journal*, 123(November 2010): 281–307, 2013.

Dan Goldhaber and Duncan Dunbar Chaplin. Assessing the ‘rothstein falsification test‘: Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, 8(1):8–34, 2015.

Dan Goldhaber and Michael Hansen. Is it just a bad class? assessing the long-term stability of estimated teacher performance. *Economica*, 80(319):589–612, 2013.

Dan Goldhaber, Joe Walch, and Brian Gabele. Does the model matter? exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, 1 (1):28–39, 2014.

JP Greene. A meta-analysis of the effectiveness of bilingual education. *Claremont, CA: Thomas Rivera Policy Institute*, 1998.

Cassandra M Guarino, Mark D Reckase, and Jeffrey M Wooldridge. Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, 2015.

Jonathan Guryan, Erik Hurst, and Melissa Kearney. Parental education and parental time use. *Journal of Economic Perspectives*, 22(3), 2008.

Douglas N Harris. Would accountability based on teacher value added be smart policy? an examination of the statistical properties and policy alternatives. *Education*, 4(4):319–350, 2009.

James J Heckman and Stefano Mosso. The economics of human development and social mobility. *National Bureau of Economic Research (No. w19925)*, 2014.

Elizabeth Howard and Julie Sugarman. Two way immersion programs: Features and statistics. *Occassional Reports UC Berkeley*, (March):10–13, 2001.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Jun Ishii and Steven G Rivkin. Impediments to the estimation of teacher value added. *Education*, 4(4):520–536, 2009.

Brian A Jacob. Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review*, 21(6):589–598, 2002.

Brian A Jacob, Lars Lefgren, and David P Sims. The persistence of teacher-induced learning. *Journal of Human resources*, 45(4):915–943, 2010.

Thomas J Kane and Douglas O Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research (No. w14067)*, 2008.

Thomas J Kane, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger. Have we identified effective teachers? validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation*. Citeseer, 2013.

Cory Koedel and Julian Betts. Value added to what? how a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, 5(1):54–81, 2010.

Cory Koedel and Julian R Betts. Does student sorting invalidate value-added models of teacher effectiveness? an extended analysis of the rothstein critique. *Education*, 6(1):18–42, 2011.

JR Lockwood and Daniel F McCaffrey. Exploring student-teacher interactions in longitudinal achievement data. *Education*, 4(4):439–467, 2009.

JR Lockwood, Thomas A Louis, and Daniel F McCaffrey. Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, 27(3):255–270, 2002.

Lesli A Maxwell. Dual classes see growth in popularity. *Education Week*, 31(26):16–17, 2012.

Lesli A. Maxwell. Dual Language Programs Take Root in N.C. *Education Week*, 34(8):1, October 2014.

Daniel F McCaffrey, JR Lockwood, Daniel Koretz, Thomas A Louis, and Laura Hamilton. Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, 29 (1):67–101, 2004.

public use dataset Panel Study of Income Dynamics. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, 2014.

Stephen W Raudenbush. What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1):121–129, 2004.

Sean F Reardon and Stephen W Raudenbush. Assumptions of value-added models for estimating school effects. *Education*, 4(4):492–519, 2009.

Steven G Rivkin, Eric A Hanushek, and John F Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

Jonah E Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2):247–252, 2004.

Jesse Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2008.

Cecilia Elena Rouse. Private school vouchers and student achievement: An evaluation of the milwaukee parental choice program. *The Quarterly Journal of Economics*, 113(2):553–602, 1998.

Donald B Rubin, Elizabeth A Stuart, and Elaine L Zanutto. A potential outcomes view of value-added assessment in education. *Journal of educational and behavioral statistics*, 29(1):103–116, 2004.

William L Sanders, S Paul Wright, and Sandra P Horn. Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of personnel evaluation in education*, 11(1):57–67, 1997.

Tim R Sass, Anastasia Semykina, and Douglas N Harris. Value-added models and the measurement of teacher productivity. *Economics of Education Review*, 38:9–23, 2014.

R. E. Slavin and a. Cheung. A Synthesis of Research on Language of Reading Instruction for English Language Learners. *Review of Educational Research*, 75(2):247–284, 2005.

Jennifer L. Steele, Robert O. Slater, Gema Zamarro, Trey Miller, Jennifer Li, and Susan Burkhauser. Effects of Dual-Language Immersion on Students' Academic Performance. *American Educational Research Journal (Forthcoming)*, 2016.

Wayne P Thomas and V.P. Collier. English Learners in North Carolina, 2009 The North Carolina Context. *Fairfax VA: George Mason University. A Research Report Provided to the North Carolina Department of Public Instruction*, 2009.

WP Thomas, V.P. Collier, and K. Collier. English learners in North Carolina, 2010. *Fairfax, VA: George Mason University. A Research Report Provided to the North Carolina Department of Public Instruction*, 2010.

I. M. Umansky and S. F. Reardon. Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal*, 51(5):879–912, 2014.

Rachel A Valentino and Sean F Reardon. Effectiveness of four instructional programs designed to serve english learners. *Educational Evaluation and Policy Analysis*, 37(4):612–637, 2015.

Teresa Watanabe. Dual Language Immersion Programs Growing in Popularity. *Los Angeles Times*, May 8, 2011.

Ann C Willig. A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of educational research*, 55(3):269–317, 1985.