Effectiveness of four instructional programs designed to serve English language learners: Variation by ethnicity and initial English proficiency

Rachel A. Valentino

Sean F. Reardon

Stanford University Graduate School of Education

March, 2014

Acknowledgements:

This research was supported by grant award #R305A110670 from the Institute for Education Sciences (IES), U.S. Department of Education. Preparation of this manuscript by Rachel A. Valentino was also supported in part by the Institute for Education Sciences (IES), U.S. Department of Education, through grant #R305B090016 to Stanford University. The authors acknowledge the substantive contributions made by their district partners to help clean and acquire data and for providing valuable feedback to help interpret research findings. The authors also thank Sandy Nader and Ilana Umansky for their invaluable help compiling the data, and to Ilana Umansky, Camille Whitney, Christopher A. Candelaria, & Lindsay Fox for their valuable feedback on earlier versions of this paper. Please direct questions to rachel.valentino@stanford.edu.

Keywords: English language learners, bilingual education, dual immersion, English immersion, Chinese, Latino, academic growth

Effectiveness of four instructional programs designed to serve English language learners: Variation by ethnicity and initial English proficiency

Abstract

In this paper we provide a descriptive and quasi-experimental analysis of the relationship between four elementary school instructional programs designed to serve English learners (ELs) and EL students' longitudinal academic outcomes in English language arts and math through middle school. We also consider differential program effectiveness by child ethnicity and initial English proficiency. Although bilingual education has been well studied, little research has examined the effectiveness of programs longitudinally, most has focused on academic outcomes only in literacy, and most research from the U.S. has exclusively focused on Spanish-speaking ELs. In this paper we find considerable differences in program effects between programs (i.e. transitional bilingual, developmental bilingual, dual immersion, and English immersion), between students of different ethnicities (i.e. Chinese and Latino), and across academic subjects.

In the past 30 years, while the overall population of school aged children increased by approximately 10 percent, the population of children speaking a language other than English at home increased by about 140 percent (NCES, 2009). On average, English learners (ELs) perform far worse than non-ELs on academic tests. For instance, on both the math and reading sections of the National Assessment of Educational Progress, the gap between ELs and non-ELs is roughly one standard deviation – about the same size as the white-black achievement gap (NCES, 2011). While the size of these gaps may in part be confounded by socioeconomic status, there are still strong associations between language status and academic performance even after controlling for socioeconomic status (Reardon & Galindo, 2009; Kieffer, 2010; Fuligni, 1997).

Given these patterns, it is critical to determine what the best and most effective instructional methods are for ELs. Despite a large body of research on the topic, the longrunning debate over whether bilingual education (in contrast to English-only instruction) is beneficial for ELs' academic development continues. As a result, there is much variability across states and school districts in the kinds of programs available to ELs (Goldenberg, 2008). Some offer several instructional options such as bilingual education or English immersion instruction, while others have effectively banned the use of bilingual education altogether (Rolstad, Mahoney, & Glass, 2005).

On one side of the debate, some data and theory suggest that ELs benefit most from being immersed in English-only classrooms, because spending more time on task practicing English results in quicker English language development (Porter, 1990; Baker 1998; Rossell & Baker, 1996). On the other side, some theory and evidence suggest that in order to learn a new language, children require a fundamental literacy base in their first language, and that fostering the continued development of children's first language will later transfer to the development of the

second language because languages share common underlying proficiencies (Cummins, 1979; 2000; Goldenberg, 1996). This perspective also stresses that academic content may be lost in translation when instruction is not in students' first language.

Generally, there is more empirical support for the latter argument, which suggests that bilingual education is superior to English-only instruction for ELs (Slavin & Cheung, 2005; Rolstad et al., 2005; Goldenberg, 2008; Greene, 1998; Willig, 1985), however, the evidence is not conclusive. Most research on English immersion versus bilingual education is not based on randomized experiments or rigorous quasi-experiments; most looks at short-term rather than long-term outcomes (for exception see: Slavin et al., 2010); and much of it is based on studies conducted in Canada or exclusively with Spanish-speaking ELs. Further, "bilingual" instruction is implemented differently in different studies, complicating any synthesis of results. For example, some bilingual models serve only ELs in the same classroom, separate from native English-speaking students, while others serve both ELs and non-ELs in the same classroom with the goal of creating biliteracy among both groups.

In this paper, we address these gaps in the literature by using quasi-experimental methods to answer two main research questions: (1) What are the differential effects of four EL instructional programs (Transitional Bilingual, Developmental Bilingual, Dual Immersion, English Immersion) on ELs' academic growth in English Language Arts (ELA) and math through middle school? and (2) Do do these growth effects by program vary by the ethnicity or initial English proficiency of the EL student?

Review of the Literature

Theoretical Perspective

Each side of the ongoing debate about the benefits and drawbacks of bilingual education is grounded in a different theoretical perspective. One the one hand, some evidence suggests that spending more time-on-task with maximum exposure to English language instruction results in quicker acquisition of and better performance in English (Porter, 1990; Baker 1998; Rossell & Baker, 1996). On the surface, this argument seems logical, and is often widely accepted because it is what EL students in the U.S. have done for generations. However, the time-on-task hypothesis is mainly supported by the effectiveness of French immersion programs in Canada, results of which may not translate to the effectiveness of English immersion programs in the U.S. (Cummins, 1999).

On the other hand, research indicates that bilingual education and the use of a student's home language is essential to fostering English language acquisition and continued academic development in other subject areas (Goldenberg, 1996; Cummins, 1979). This finding is further supported by existing research which has documented a positive relationship between time spent in bilingual instruction and academic achievement (Goldenberg, 1996; Rolstad et al., 2005; Collier & Thomas, 2004; Greene, 1997).

There are two main reasons why this positive relationship between bilingual education and EL students' academic performance might exist. The first is that if students are immersed in English-only instruction but have not developed a minimum level of competency in English, there will likely be a discrepancy between what is taught and what is understood (Goldenberg, 1996). Further, children need a knowledge base to be effective readers and speakers. They may be able to continue expanding that knowledge base more quickly if they are taught in a language that they are more familiar with than if they are learning in a language that they do not fully understand.

The second is that the continued development of children's first language may facilitate acquisition of the second language, as academic language skills may be developmentally linked to similar underlying proficiencies that are common across languages (Cummins, 1979; 2000; Goldenberg, 1996). For instance, Collier & Thomas (1989) find evidence that immigrant students with two to three years of initial schooling in their country of origin tend to perform better academically than those who start in a new country. These findings are consistent with the idea that children should learn to read in their home language first, rather than attempting to both learn to read in general, and read in a new language simultaneously (Cummins, 1999).

Types of Bilingual Education

Although there is some suggestive evidence that bilingual education is more beneficial for second-language learners than is English immersion, there are a number of different models of two-language instruction and there is not conclusive evidence to suggest that each model provides equally beneficial effects. There are three main models of instruction that utilize a two-language model in the classroom: transitional bilingual, developmental bilingual, and dual immersion instruction.

Transitional bilingual classrooms serve only ELs, separate from their non-EL peers. Instruction starts primarily in students' home language in kindergarten, and increases in the amount of English used for instructional purposes at a rapid pace in the early elementary years. The intention of these programs is to transition students into mainstreamed classrooms (i.e. English immersion classrooms) quickly – usually by grade two or three. These programs use ELs' home languages to ensure that they receive adequate access to core academic content, but

do not necessarily ensure that students are completely proficient in their home language before transitioning them to a mainstreamed English-only classroom.

Developmental bilingual education programs are similar to transitional bilingual programs in that they incorporate EL students' home language into classrooms and exclusively enroll ELs, but these programs are longer term, often lasting through the fifth grade or later, and have the goal of helping students develop competency in English while maintaining and continuing to develop competency in their native language.

Finally, dual immersion programs are more similar to developmental than transitional bilingual instruction because they hold a goal of facilitating biliteracy through a program that lasts longer, however, they differ slightly in approach to instruction. Dual immersion programs enroll both native English speakers and ELs in the same classroom. A goal of the program is for both groups to emerge bilingual and biliterate. In early years the majority of instruction occurs in the ELs' home language (often referred to as the target language), but by late elementary school, about half of the instructional time is spent in the target language and the other half is spent in English.

These three different two-language instructional models can be contrasted with English immersion programs, which teach all students primarily in English, in classrooms with some number of English-speaking peers. Often, but not always, school districts offer additional English as a second language (ESL) pull-out services during the day to provide instruction targeted at students' level of English proficiency and that specifically focuses on their English language development. No where in this approach is the child's home language used.

Effectiveness of Bilingual Education

There is a sizable body of literature documenting the effects of bilingual education compared to English immersion instruction on ELs' academic performance. A handful of meta-analyses have tried to summarize the literature, but the conclusions of these meta-analyses vary, depending on the study inclusion criteria they use. The two meta-analyses that used the most stringent inclusion criteria conclude that ELs who attended bilingual programs outperformed their peers who attended English immersion programs by anywhere from 0.18 to 0.33 standard deviations per year in academic subjects. Further, when restricted to only randomized experiments or only studies conducted in the U.S., effect sizes were on the higher end of this range (about 0.3 standard deviations per year) in each case (Greene, 1997; Slavin & Cheung, 2005). While the size of these effects is encouraging, most of the studies included in these two meta-analyses only tracked outcomes for two to three years at best. Further, they do little to tease apart the differential effectiveness of specific two-language instructional models (i.e. transitional bilingual vs. developmental bilingual vs. dual immersion), making it difficult to disentangle which components make bilingual programs so effective.

There are a handful of high quality studies conducted in the U.S. that have attempted to evaluate specific or multiple bilingual instructional approaches. One in particular randomly assigned students to one of two programs: a transitional bilingual program or an English immersion program. They then tracked Spanish-speaking EL students' outcomes from kindergarten through fourth grade. They found that in the early grades, ELs that were enrolled in an English immersion program outperformed their peers who attended transitional bilingual programs in English, but by fourth grade, no significant differences on the English assessments emerged. (Slavin, Madden, Calderon, Chamberlain, & Hennessy, 2010). These findings suggest the possibility that in early grades, some forms of bilingual instruction may slow the process of

English language development, simply because much instructional time is spent on home language development, but that ultimately transfer may occur from the home language to English, which is why ELs in bilingual instruction ultimately catch up. Among other things, the findings point to the importance of long-term follow-up to determine "effectiveness".

Other research has compared more than one bilingual instructional model within the same study. Ramirez, Yuen, Ramey, and Pasta (1991) compared both transitional bilingual and developmental bilingual programs to English immersion programs among Spanish-speaking ELs. Similar to Slavin et al (2010), the authors found that in early grades, students attending transitional and developmental bilingual programs performed worse in English than their peers enrolled in English immersion classrooms, but by second grade this significant difference disappeared. The findings from this study should, however, be interpreted with caution, as the authors' matching algorithm did not account for students' pre-test scores. Still, between both papers, the findings suggest that at the very least, bilingual educational does not hinder English language development and performance in the medium term.

The above studies do little, however, to shed light on the potential benefits of dual immersion instructional programs, which to date have not been extensively researched. Two noteworthy studies do consider the effects of such programs on students' outcomes. The first is the only paper to our knowledge that examines the same three instructional models that we do.

Thomas and Collier (2002) found that across five large school districts, ELs attending dual immersion programs almost always outperformed their peers in transitional and developmental bilingual programs. Further, in all districts, the students attending the developmental bilingual programs always performed at least as well as and in some districts better than those in the transitional bilingual programs. It is possible that the variation across

districts is in part accounted for by differences in how transitional and developmental bilingual programs were implemented. This study provides good descriptive evidence of differences in EL students' performance across programs, but did not include any control variables. It is possible that the observed differences across programs were due to the fact that students enrolling in different types of programs differ systematically on characteristics such as SES.

The second study randomly assigned preschool students to either dual immersion or English-only preschool classrooms and found that through first grade, there were no differences in English language, literacy, and math outcomes between students attending these two programs (Barnett, Yarosz, Thomas, Jung, & Blanco, 2007). The study also found significant gains in the Spanish language development of both EL students and native English speaking children without loss to English language development. However, this particular study takes place in pre-K, before language minority children are tested for classification as "English Learners." Also, since assignment only lasted for the preschool year, we do not know what kinds of programs these students attended in kindergarten or first grade. It is possible that some students continued to attend some type of dual immersion instruction, while others were in English immersion.

Motivation for the current study

Although there is a sizable body of literature comparing the effectiveness of bilingual education to English immersion instruction among ELs, there are still many gaps in the literature. First, the overwhelming majority of studies tracking elementary-aged ELs exclusively consider outcomes for one to three years after initial program attendance, and even the few exceptions to this still only track differences in academic abilities through fourth (Slavin et al., 2005) or fifth grades (Maldonado, 1977; Collier & Thomas, 2004). Tracking outcomes beyond these grades is

particularly important in light of the fact that children initially enrolled in bilingual programs need time to develop English skills (Hakuta, Butler, & Witt., 2000) and may actually realize the largest gains from program attendance in the longer term. Further, most current studies almost exclusively consider outcomes in English and/or ELs' home languages, without considering the impact of bilingual instruction on academic development in other core subjects (for exceptions see Ramirez et al., 1991; Willig, 1985; Barnett et al., 2007). In this study, we add a longitudinal and multi-subject perspective by looking at outcomes from kindergarten through late middle school in English language arts (ELA) and math.

In addition, most research that has been conducted on the topic in the U.S. exclusively focuses on the effectiveness of different instructional programs for Spanish-speaking ELs, or perhaps worse yet clusters all ELs into one category, without considering differences in initial language spoken and initial level of English proficiency. First, although generally evidence suggests that supporting a child's home language development can ultimately transfer to second language proficiency because some features of language, such as reading comprehension, are universal across languages (Goldenberg, 2008), other research also indicates that the degree of transfer across languages may vary depending on the structures of the languages in question. When languages are typologically distant (such as English and many character-based East-Asian languages), procedural literacy skills may be less likely to transfer (Genesee, Geva, Dressler, & Kamil, 2006; Lado, 1964). More specifically, different cognitive resources become dominant when developing procedural literacy skills in different languages. For example, visual processes may be more dominant when learning to read a character-based language like Japanese, than when learning an alphabetic language such as English or Spanish (Geva, 2006). When there are typological language differences, it is thus unlikely that all features of learning language such as

letter-sound correspondence, phonological awareness, and reading comprehension will be identical across languages (a reality that is more likely between typologically similar languages). In other words, there may be fewer underlying proficiencies that transfer from one's first language to the second when the languages are typologically different than when they are similar.

Motivated by this background research, we disaggregate findings by Chinese and Latino ELs. Because Spanish and English have many structural similarities across languages, we hypothesize that Latino ELs in two language programs, particularly those that foster continued development of one's home language (i.e. developmental bilingual and dual immersion) will do significantly better than their latino peers who are enrolled in English immersion programs. However, because Chinese and English have very different phonological structures and distinct alphebets, we hypothesize that Chinese ELs in English immersion programs will perform better than their Chinese peers in bilingual programs.

In addition to considering findings by ethnicity, we also disaggregate our findings by initial English proficiency to consider whether some programs are more or less effective for students of different language backgrounds. To our knowledge there is little research to date on differential benefits of bilingual instruction by the initial level of English proficiency of students. For this reason, we make no a priori hypotheses about the differential effectiveness of initial English proficiency on program effectiveness.

Finally, very few studies in the extant literature adjusts for potential selection bias. In our analyses, we account for a robust set of controls and use methods that allow for a stronger causal inference than much of the existing literature. Importantly, because the district of study employs a choice system where families rank their program preferences prior to program entry (after

which they are assigned to schools based on a complex algorithm, which includes some randomness to break ties), in one of our approaches we control for the factors used in the assignment algorithm and for parental school and instructional program preferences in our models. Our ability to control explicitly for the parental preferences used in the algorithm provides a relatively strong causal warrant to our results. At the very least it provides a stronger warrant than if we could control only for observable student characteristics. We also present on results that use an exploratory instrumental variables approach in appendix A.

Taken together, our study will add to the literature by considering long term program impacts, differences in program effects by student characteristics, and differences across a diverse set of instructional approaches all while attempting to use more stringent methods than much of the research has used to date.

Data and Methods

Data

The data used in the current study comes from a large urban district that serves a sizable EL population. Our analytic sample follows 13,750 EL students who entered the district in kindergarten sometime between the 2001-2002 and 2009-2010 academic years. In the entry year, students are assessed to determine initial English proficiency and thus EL status. It is also the point at which students were assigned to their initial program of attendance. The district of study implements a choice model for school selection, where families rank program (i.e. instructional model and school) preferences prior to program entry, after which they are assigned based on an algorithm. The district's algorithm attempts to give applicants their highest possible choice, but uses a number of "tie-breakers" to determine who gets into programs that have more applicants

than slots (which many do). The algorithm gives first priority to students who have a sibling at a given school. The next tie-breaker gives preference to students who are currently enrolled in the district's preschool or transitional kindergarten program in the same attendance zone as their listed preference, followed by a tie-breaker that gives preference to students living in neighborhoods in the city where average test scores are low. The final tie-breaker is given to students who live in the school's attendance zone. Among students with the same priority rankings, ties are broken using random assignment. The tie-breaker process adds some randomness, through which we can compare students who had the same school-by-program preferences, but attended different programs and/or schools.

Our outcome data come from the state standardized tests in English language arts (ELA) and math that students took each year from second through eighth grade. While we consider outcomes in ELA through eighth grade, we only analyze outcomes through sixth grade for math. We do this because starting in seveth grade, students may take a subject-specific math test (e.g. general math vs. Algebra). This means that any two students enrolled in the same grade may have taken different state tests if they were enrolled in differe math classes. We wanted to ensure that our analyses only compared students who took the same test. We standardize these ELA and math outcomes relative to the state mean within each grade and year. Of our analytic sample, approximately 33% were Latino ELs, approximately 45% were Chinese ELs, and the remaining were ELs of a variety of other ethnic backgrounds, including approximately 5% of Japanese, Korean, or Filipino backgrounds.

[Insert Table 1 about here]

Initial Program: We identify the programs ELs in the district initially attended: English Immersion (EI), Transitional Bilingual (TB), Developmental Bilingual (DB), and Dual

Immersion (DI). Program definitions, including the mission of each program, the population of students served, and the amount of instructional time spent in English versus the target language can be found in Table 1. We use students' initial program attended, which is stable within students, to estimate program effectiveness. Although on the surface it may appear as though initial program only measures a one year "treatment", we interpret observed effects of program attendance for the full number of years intended by the specific program model. We do this because the majority of our sample (95.2%) attend the same program for at least four years, from kindergarten through third grade. This indicates that there is little movement in and out of programs once ELs enroll in a particular program during their kindergarten year. After third grade, the proportion of students who are enrolled in the same program that they were initially enrolled in begins to differentially drop depending on the initial program attended. For instance, as can be seen in Table 2, because TB programs are designed to reclassify students as fluent English proficient and transition them into English immersion programs more quickly than the DB and DI programs, lower proportions of ELs who were initially enrolled in TB are still enrolled in that same program in 4^{th} and 5^{th} grades than is true of ELs who were initially enrolled in the DB and DI programs. This change is simply an artifact of the program design rather than reflecting a lack of compliance. As is also reflected in Table 2, programs are typically available to students through 5^{th} grade, at which point, across models, students are generally transitioned into EI programs.

[Insert Table 2 about here]

Sample Descriptives. The majority of students in our sample (57%) are initially enrolled in EI programs. Approximately 21% of ELs in EI are Latino, while approximately 47% are Chinese. About equal proportions of EL students are enrolled in the TB and DB programs – 20%

and 17%, respectively. More specifically, approximately 37% of those initially attending the TB programs are Latino ELs and 56% are Chinese, while these figures are 50% and 43%, respectively in the DB program. The DI program enrolled the smallest portion of ELs in our sample (8%), in part because there are fewer of such programs available and in part because up to half of the slots in DI programs are reserved for non-EL students. Latino ELs make up the majority of ELs enrolled in DI (71%), followed by Chinese (14%) ELs. Students initially enrolled in the DB and DI programs have the lowest initial English proficiency in the fall of kindergarten, while students initially enrolled in the TB program have the highest scores. A more detailed description of sample demographics by initial program attended can be found in Table 3.

[Insert Table 3 about here]

Table 4 describes the proportion of students initially enrolled in each program in our sample by their year of entry into kindergarten. Although we have outcome data available through the 2011-2012 academic year, our sample includes only students who entered kindergarten in fall 2009 or earlier, as academic outcomes are measured from grade two onward. Approximately 1500 EL students entered our sample in each year. The proportion of students initially enrolled in EI remained fairly constant over each year in consideration, while the proportion of students initially enrolled in TB and DB decreased slightly, and the proportion initially enrolled in DI increased over time, largely because the district expanded the number of DI programs during this time period.

[Insert Table 4 about here]

Methods

Research question 1. In order to answer the first research question of interest regarding the differential effect of each instructional program on ELs' academic growth through middle school, we estimate three separate three-level growth curve models in a Hierarchical Linear Model (HLM) framework: the first without student controls, the second with added student controls, and the third with added student controls and fixed effects for student/parent preferences. In this three level approach we model observations over grades *t* (level 1), which are nested within students, *i* (level 2), who attend programs *p* (level 3). At the level two, we allow students' individual intercepts and slopes to vary across grades. This allows a set of more flexible assumptions than that of standard regression, which assumes that average effects are constant across students. In addition, at the third level, students' intercepts and slopes are allowed to vary by 191 distinct school-by-instructional program combinations. For instance, if two EL instructional models, TB and an EI are offered in school A, and the same two models are also available in school B, this would represent four rather than two distinct programs. We estimate three-level models like this to obtain correct standard errors that account for clustering within school-by-program cells. These models take the following form:

$$
\begin{aligned}
Level\ 1: \quad & Y_{tip} = \alpha_{0ip} + \alpha_{1ip} G_{tip} + e_{tip} \\
Level\ 2: \quad & \alpha_{0ip} = \beta_{00p} + \mathbf{X}_{ip}\mathbf{B}_0 + r_{0ip} \\
& \alpha_{1ip} = \beta_{10p} + \mathbf{X}_{ip}\mathbf{B}_1 + r_{1ip} \\
Level\ 3: \quad & \beta_{00p} = \gamma_{000} + \mathbf{P}_i\mathbf{\Gamma}_0 + u_{00p} \\
& \beta_{10p} = \gamma_{100} + \mathbf{P}_i\mathbf{\Gamma}_1 + u_{10p}
\end{aligned}
\qquad (1)
$$

Y_{tip} represents the ELA or math score for student *i* in grade *t* in program *p*. The variable G_{tip} indicates a student's grade, centered at grade 2, so that α_{0ip} and α_{1ip} indicate student i's test score in grade two and the average rate of change of his or her test scores from grades 2 to 8 (or to grade 6, in the case of math), respectively. The student-specific intercepts and slopes are

modeled as linear functions of a vector of student characteristics and initial program preferences ($\mathbf{X}_{ip}$) and a vector of indicator variables indicating the student's initial program type ($\mathbf{P}_i$), plus student- and school-by-program random effects. The coefficients of interest here are the vectors $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$, which indicate the differences among instructional program types in the intercepts and slopes, respectively, of EL students' test score trajectories.

In our first model (Model 1 in Table 5) we fit Equation (1) without any student-level covariates (no vector $\mathbf{X}_{ip}$) to provide a baseline descriptive model. In our second model (Model 2 in Tables 5), we include a vector of grand mean centered stable student/family control variables, $\mathbf{X}_{ip}$, at level two in the model. In Model 2, this vector includes the students' gender, ethnicity, and initial English proficiency score[1].

While Model 2 adjusts for a set of observable student characteristics that are undoubtably related to students' academic growth trajectories and students' choice of programs to attend, alone they may not fully account for student selection into programs. However, because we have data on parental preferences for the type and location of the EL instructional program that they wanted their child to attend, we are able to explicitly control for these pre-treatment preferences. Specifically, we include a vector of dummy variables indicating which of 191 school-by-program options parents listed first on their school-entry application. We add this set of additional school-by-program preferences to our existing vector of student/family controls, $\mathbf{X}_{ip}$, at level two. Because families can, and often do, list multiple ranked choices on their school-entry application, we also ran these models using various different specifications of "preferences," including one that controlled for students' top three choices for instructional

[1] Note, due to the Family Education Rights and Privacy Act (FERPA) we were unable to control for student free and reduced price lunch (FRPL) status. However, we note that we received information from our district partners indicating that the proportion of ELs qualifying for FRPL status does not vary across ethnicity-by-pathway combinations. In fact, most ELs in the district qualify for FRPL. Thus, controlling for such a variable is unlikely to change our results.

program. Our findings are robust to all specifications, so for the sake of parsimony, we present on just those controlling for students' first school-by-program preference. Results controlling for student preferences are presented in Model 3 of Table 5.

Because school-choice data are only available for students who entered the district in kindergarten starting in 2004, we only analyze academic outcomes through 7th grade in ELA for these models to ensure that we have adequate sample sizes in all grades. Because of this, and also the fact that we have to restrict our sample to only students for whom we have preferences data, the sample in Model 3 is roughly half the size of the sample in Models 1 and 2. To ensure that any differences between Models 2 and 3 are not due to the difference in samples, we also fit Model 2 using the smaller sample used in Model 3. These are presented as "Model 2: Restricted Sample" in our results tables.

We also attempted to answer this question using an exploratory instrumental variable (IV) approach in a two-stage least squares (2SLS) framework. We use the initial addresses of students to construct three instrumental variables of the geographic proximity of students to each instructional program type. We consider this approach exploratory because we do not have a means of testing our exclusion restriction and the standard errors in our models sizably increase. For this reason, we do not present these results in the main paper, but a description of these methods and results can be found in Appendix A.

Research question 2. In order to answer the second research question as to whether program effects vary by ethnicity or initial English proficiency, we add interactions between program type and both a set of dummy variables for students' ethnicity ethnicity and standardized initial English language proficiency score to our full three-level program model (Equation 1 above). Because of limited power, we do not estimate these by-subgroup models in

our IV framework. Also, for the sake of parsimony, and because of the remarkable consistency of results answering question one across Models 1 through 3, we remove our descriptive model (1) from Table 6. This leaves Model 2 (including the full sample of years), which controls for stable student characteristics but not preferences; Model 2: Restricted, which restricts the sample of Model 2 to be that of Model 3 for the sake of comparing results across the same sample; and finally, Model 3, which is a smaller sample, but adds controls for program preferences.

It should be noted that there are a handful of two-language instructional programs for ELs of other ethnic backgrounds (e.g. Japanese), but the samples of students and the number of such programs available are too small to analyze their ethnicity-specific program effects with adequate power. Thus, while these students and similar interactions by ethnicity and program are included for all student groups in our models, we do not report of specific program effects for these additional subgroups.

Results

Research Question 1.

Results for our first research question of interest, regarding the differential effect of each instructional program on ELs' academic growth through middle school, are presented in Table 5. Model 1 presents our base growth models without controls, model 2 presents our growth models including our stable student controls, model 3 presents a version of model 2 that adds controls for program preferences, and model 2: restricted sample presents results running model 2 on the same sample as model 3. The results are generally similar across these specifications.

We start by interpreting the intercepts. Because our data set considers ELs who enter the district of study in kindergarten, but we do not begin observing academic outcomes until second grade, our intercepts may reflect two factors: (1) differences in effects across programs that operate by second grade (relative to attending EI), and (2) differences in selection across programs. To the extent that the latter is true, any causal interpretation of our estimates could be biased by selection. However, selection bias is likely to be relatively small because we control for (1) students' initial English proficiency scores, which are strong predictors of later academic scores, and (2) because we control for parental preference fixed effects, which likely capture a large portion of any residual selection we might be concerned about. We thus interpret our estimates (for both intercepts and slopes) as largely the result of differences in program effectiveness.

[Insert Table 5 about here]

For the sake of parsimony, we primarily rely on Model 2 (with the full sample) and Model 3 (controlling for student preferences) in discussing our estimates; the estimates, however, are very consistent across specifications.

ELA. We first discuss program intercepts.The effect through second grade in ELA of attending EI is not statistically distinguishable from the performance of the average student in the state. The effect through second grade in ELA is significantly higher for those in TB than for EI (about 0.15 SD higher), not significantly different for those in DB, and significantly lower for those in DI than those in EI (about 0.16-0.24 SD lower).

The of program attendance on rates of growth in ELA through 8^{th} grade tell a slightly different story than our observed effects through second grade. These results indicate that in general, the test scores of ELs in EI increase at a rate that similar to those of the average student

in the state (growth of which is 0 standard deviations per grade). Further, the rate at which the ELA test scores of ELs in TB and DB increase is not significantly distinguishable from those of students in EI. Finally, although ELs attending DI do not initially have ELA scores well above their peers in EI in second grade, from second through eighth grade the ELA test scores of ELs in DI increase at a rate that is as much as 0.092 standard deviations faster per grade than those in EI (or up to 0.091 standard deviations per grade in general). This rate is so fast, that by fifth grade their test scores in ELA catch up to the state average, and on average by seventh grade ELs in DI are scoring above their EL counterparts in in all of the other programs. These findings suggest that while in the early years of attendance DI programs may have a negative effect on performance in ELA, in the long term, the short term negative effects are more than overturned by the positive effects on test score growth.

These trends can best be seen in Figure 1 which presents results from Models 2 and 3 from Table 5, respectively. The comparison of these two figures also demonstrates the similarity of results before and after controlling for program preferences. The consistency of results across model specifications with and without program preferences means that differences in preferences are not confounding our estimates in model 2. Although it is possible that there are still other factors that we did not observe that affect selection into programs and that are correleated with academic trajectories, we have no strong reason to believe that there are factors that would not be captured by our existing set of controls.

[Insert Figures 1 about here]

Math. Much like ELA, in math models 2 and 3 yield similar results. According to the intercepts in Table 5, the effect of attending EI on math scores by second grade is large and yields significantly higher scores than the state average (about 0.14 SD). Similar to ELA, by

second grade the math scores of those attending TB are significantly higher (by about 0.27 sd) than those attending EI. The scores of those in both DB and DI do not significantly differ from those in EI in second grade, which indicates that students in these programs, like those in EI, score above the state average in math in second grade.

The slopes in Table 5 indicate that the math test scores of students receiving EI instruction either grow significantly more slowly than the state average (according to model 3, which controls for preferences), or at best do not differ significantly from the state average (according to model 2 which controls for stable student characteristics but not preferences). Further, the math test scores of EL students in TB grow significantly even more slowly than those in EI, by about 0.04 standard deviations per grade according to both models 2 and 3. The math test scores of ELs receving DB instruction grow at a rate that is not statistically distinguishable from those receving EI instruction, while, the test scores of ELs in DI grow faster than ELs receiving EI instruction, although the statistical significance of these results are marginal at best ($p<0.10$ in model 3 only). Best seen in Figure 2, it is also note-worthy that ELs in DI are the only ones whose test score trajectories are not slower than the state average, but rather mirror that of the average student in the state. And again, comparing the patterns of rates of growth by instructional program across models 2 and 3 in Figure 2, it can be seen that trends are generally consistent. For this reason, and because we more strongly believe the causal claim of the models that control for parent preference fixed effects, we rely on results from model 3 for figures depicting results by subgroup.

[Insert Figure 2 about here]

Research Question 2.

We turn now to question 2, which considers whether program effects vary by EL students' ethnicity or initial level of English proficiency. Results are presented in Table 6, and report on only models 2 and 3. Furthermore, in Table 6 we only report point estimates for slopes for the sake of parsimony. The intercepts can easily be observed in Figures 3 and 4, and are available from the authors upon request.

Second grade intercepts by ethnicity tell a very similar story to that of all ELs. For both Latino and Chinese ELs, by second grade the scores of those in DB are not significantly distinguishable from those in EI. The scores of ELs in TB are significantly higher than those in EI, and the scores of those in DI are significantly lower than those in EI. This pattern of program effectiveness relative to EI is largely consistently across ethnicities and subjects. However, as is evident in both Figures 3 and 4, Chinese ELs score well above their Latino EL peers (and the state average) in both ELA and math, not only in second grade, but across all grades.

According to the slopes found in Table 6, the ELA test scores of both Latino and Chinese ELs in DI grow faster than the those of ELs in any of the other three programs. The same is generally true in math, or at the very least, ELs of both ethnicities who receive DI instruction do not perform significantly different in math than those in EI.

[Insert Table 6 about here]

Differences by ethnicity emerge among TB and DB programs. The ELA and math test scores of Chinese ELs in both bilingual programs grow significantly more slowly than their Chinese peers in EI. On the other hand, the test scores of Latino ELs in TB and DB increase at a rate that is not statistically distinguishable from ELs in EI, at least in ELA. The significantly slower rates of growth of Chinese ELs' test scores in both the TB and DB programs are not only significant, but are also large; about 0.05 to 0.08 standard deviations less per grade in ELA and

about 0.09 to 0.14 standard deviations less per grade in math than their Chinese EL peers in EI. This means that although Chinese ELs in TB or DB score higher in both ELA and math than their Chinese EL counterparts in EI and DI in second grade, on average Chinese ELs in the latter two programs catch up to or surprass Chinese ELs in the former two programs by middle school. This is true of both subjects. Latino ELs in DB do not experience these same slow rates of growth in ELA, but rather experience the same average rates of test score growth as their peers in EI, and although the math test scores of Latino ELs in TB grow more slowly than their Latino counterparts receiving EI instruction, the difference (β = -0.04 to -0.05) is still over 50 percent smaller than that of Chinese ELs. Finally, a test that the Chinese program slopes (e.g. TBE x Grade x Chinese) are jointly equal to zero was statistically significant for both ELA ($\chi^2(3)$= 60.79, p<0.001) and math ($\chi^2(3)$= 46.42, p<0.001), indicating that the program-specific rates of test score growth among Chinese ELs significantly differ from those of Latino ELs. *P*-values for these joint tests across all model specifications for both intercepts and slopes can be found at the bottom of Table 6.

There were generally no differences in program effects by initial level of English proficiency. There was one exception in our results.The restricted samples considering growth for more recent years indicate that initial level of English proficiency is more positively associated with test score growth in ELA for students in DI programs than for those in EI programs. These results are not dramatically large and are only significant at the $p < 0.05$ level in one of the three model specifications, so we don't place heavy weight on this finding.

[Insert Figures 3 & 4 about here]

Discussion

In this paper, we provide a descriptive and quasi-experimental analysis of the association between elementary school EL instructional programs and EL students' longitudinal academic outcomes in ELA and math. We build on prior research on the topic by focusing on academic outcomes in two subjects through middle school, by comparing the effectiveness of four different two-language instructional models, and by evaluating whether these EL programs are differentially effective for students of different ethnicities or language backgrounds. In addition, our models controlling for prior program preferences arguably provide better estimates of program effects than does much of the existing literature, as we are able to eliminate at least a portion of selection bias that is due to individual preferences (a common unobservable characteristic in other similar studies).

Four key findings are worth noting in this study. First, we find that in the short run (by second grade), there are substantial differences in the academic performance in ELA and math among EL students who start in different EL instructional programs in kindergarten. Second, we find that despite these early observed differences, in the long run the story is quite different. For example while in the short term (through second grade), ELs in dual immersion score substantially below their EL counterparts attending other instructional programs in ELA, in the long term, they score substantially above their peers in other programs. Third, we find that there are differences in program effects by ethnicity, particularly in terms of rates of growth. For instance, we find that in ELA, while the test scores of both Latino and Chinese ELs grow the fastest when they are enrolled in dual immersion programs, the ELA scores of Chinese ELs in English immersion also growth significantly faster than their Latino peers in English immersion. Finally, we find no evidence that programs are differentially effective for ELs who enter the district with high versus low initial English proficiency in Kindergarten.

For our first finding, that in the short run there are sizable differences in program effects on EL outcomes in ELA and math, we specifically find that by second grade ELs in dual immersion have ELA test scores that are well below those of their peers in English immersion. At the same time, ELs in transitional bilingual have test scores that are well above those of ELs in English immersion in both ELA and math. These findings highlight that if we measure "effectiveness" as performance in English in second grade (i.e. in the short term), we might conclude that dual immersion programs are the least effective, and in fact that programs that emphasize more English instruction earlier (transitional bilingual and English immersion) result in better effects for ELs. However, provided our longer-term findings, these short-term results highlight the potential problems with relying on short term outcomes (as much of the existing research does) to determine program effectiveness.

Our second finding, that in the long term there are differences across programs in EL student's rates of growth in ELA and math, do not necessarily mirror our short-term findings. These long term results are especially pronounced in the case of dual immersion. Here, although students score below their peers in English immersion in the short term, by middle school they score at least as well as, if not better than their peers in the other programs in both ELA and math. These findings are likely an artifact of program design. ELs in dual immersion spend more time early on in the target language (e.g. Spanish, Cantonese, etc) than any of the other programs do (about 80-90% of their instructional time in kindergarten through first grade; see Table 1). The early scores are in part due to the fact that tests (both those in ELA and math) are in English. Although ELs in dual immersion score poorly on tests in English in early grades, it is not necessarily an indicator that they aren't developing important content knowledge and literacy

skills that in the long term will ultimately transfer to English language and other academic development. These skills may simply not be captured in second grade by the English tests used.

Further, the rates at which the test scores of ELs in dual immersion increased, particularly in ELA, far out-paced those of the ELs in the other programs. It is possible that dual immersion programs have this effect because they combine the best of both English immersion and bilingual instructional models into one program. Specifically, dual immersion instruction (a) exposes ELs to native English-speaking peers, while still (b) providing instruction in ELs' home language to support continued development of that language. The first piece is important because having classmates, at least one third of whom are native English speakers, may prove useful for modeling English language use. The second piece is important for two key reasons: first, because use of ELs' home languages will help to ensure that they do not fall behind in core academic subjects due to a lack of understarnding, and second because ELs might benefit from transfer of language skills from one language to the other if continued development of literacy in their home language is supported. More specifically, there is evidence that languages share core underlying structures that require similar proficiency skills, and that children who are just beginning to learn to read and write can benefit from continued support or their home language development because such underlying proficiency skills ultimately transfer across languages (Cummins, 1979; 2000; Goldenberg, 1996).

Our third main finding is that there are sizable differences in program effects across ethnicity. First, we see that regardless of program, Chinese ELs' test scores are far above their Latino peers in both ELA and math across grades. This is an achievement gap that is not unique the district studied in this paper, and worthy of attention in and of itself for considering how to promote equal outcomes of ELs across ethnic groups. In addition, our findings by programs and

ethnicity indicate that in ELA, while the test scores of both Latino and Chinese ELs enrolled in dual immersion programs grow the fastest relative to the other instructional programs, the story is different for English immersion and both bilingual programs. The test scores of Chinese ELs in English immersion grow significantly faster than those of Latino ELs in English immersion. This is true in both ELA and math. Further, the test scores Chinese ELs in the transitional and developmental bilingual programs grow significantly more slowly than their Latino peers in these programs.

These findings of the significant negative effects of both types of bilingual instruction and positive effects of English immersion instruction on Chinese ELs' test score growth has a three plausible explanations. The first comes from evidence suggesting that the extent to which home language use in the classroom transfers to second language acquisition depends on the structural similarity of the two languages (Lado, 1964; Genesee et al., 2006; Echman, 1977). Transfer is more likely if the first and second languages are typologically similar (e.g. Spanish or French and English), but less likely if the languages are typologically distant (e.g. Japanese or Chinese and English). In the latter case, because alphabets, phonemes, and overall language structures are mis-matched, bilingual education may be less effective at promoting English language development. This could in turn mean that more time spent "on task" in English may be a more effective means of academic instruction for Chinese ELs than it is for Latino ELs (if of course one's outcomes of interest are tests in English). Although our results seem consistent with this explanation, it is not clear that typological similarity entirely accounts for the difference, especially given the positive effects we see for Chinese ELs in dual immersion. Indeed, some researchers have argued that even if transfer is less likely among some languages than others, there may still be benefits of bilingual education across language types because there are

underlying proficiencies that are common across all languages such as language processing and reading comprehension (Goldenberg, 2008).

Another plausible explanation is therefore differential fidelity of implementation across Chinese and Latino bilingual programs within the district of study. If more instructional time is spent in English within Chinese bilingual programs than the program designs intend, and since Chinese ELs in bilingual programs are exclusively enrolled in classrooms with other Chinese ELs (and thus isolated from English-speaking peers to model effective English use), this could lead to diminished results compared to those that we would have expected had the program been implemented entirely as intended.

Finally, our fourth major finding indicates that no particular instructional model appears to be better at serving ELs at one level of initial English proficiency over another. Although we have no strong inclinations for why this may be, one potential explanation is that such an analysis of differential effectiveness by initial level of English proficiency is less relevant for 5-year-old ELs who, despite variation in English speaking ability, vary less in any form of literacy (i.e. they have not yet learned to read or write in *any* language). Rather, it is possible that such an analysis is more relevant for older ELs who enter the U.S. school system in later grades and who have already developed a strong literacy base in their home language. In this case, those low in initial English proficiency may need more home language support to access core academic content and may thus benefit more from two-language instructional models, while those with higher initial English proficiency may benefit more from direct immersion in English-only classrooms with peers who are fluent in English. It is difficult to know for certain without such an analysis of ELs who enter the American school system in later grades.

Concluding remarks & Study Limitations

The interpretation of these findings is subject to the study's limitations. The degree to which our specific program findings generalize to other settings is not entirely clear. Our data come from a single school district, which is somewhat unique in terms of its ethnic and linguistic diversity and its commitment to providing multiple different types of EL instructional models. There is a lot of heterogeneity across the U.S. in the delivery of two-language instructional models. For instance, some bilingual programs begin in kindergarten providing instruction half of the time in each language, while others start heavily (about 90% of instructional time) in children's home language in kindergarten (Collier & Thomas, 2004). Our study speaks to the effectiveness of four distinct and very specific program models, primarily for two groups of ELs: Spanish and Chinese.

Further, our interpretation of "program effectiveness" is limited to outcomes in English. It is possible that through these outcomes we are unable to capture other important outcomes that matter for EL students' development. For example, we find that the test scores of Latino ELs in developmental bilingual programs grow at a rate that is not statistically distinguishable from the rate at which the scores of ELs in English immersion grow. On the one hand, this highlights that, at the very least, long-lasting bilingual education will not hinder Latino ELs' academic development. However, it is also important to note that ELs enrolled in bilingual programs for six years or more may also reap the added benefit of bilingualism and biliteracy – a potentially important skill for future labor market potential (Gándara & Rumberger, forthcoming). Because we do not observe student outcomes in their home languages, we cannot capture this potential added benefit in our measures of effectiveness.

A third limitation of the study is that we do not have explicit controls to adjust for classroom quality. To the extent that there are systematic differences in classroom quality across programs, our program effects could in part be biased. However, in one iteration of our analyses, we included school fixed-effects to compute within school program effects. While this approach does not adjust for classroom quality perse, it does adjust for differences in quality across schools. These results (available from the authors upon request) did not deviate from those of our final model specification.

Finally, in this paper we attempt to parse out the effect of selection through a rich set of preference controls. However, we recognize that our study is not a randomized experiment, which limits our ability to infer causality with full certainty. Still, we believe that taken together, these approaches provide a much stronger causal warrant than much of the existing literature. Perhaps most importantly, we believe that our results provide a platform through which continued research about the effectiveness of different instructional models used to serve ELs can be initiated.

References

Angrist, J.D., & Krueger, A.B. (1999). Chapter 23 Empirical strategies in labor economics. *Handbook of Labor Economics, 3,* 1277-1366.

Baker, K. (1998). Structured English immersion: Breakthrough in teaching limited-English-proficient students.

Barnett, W.S., Yarosz, D.J., Thomas, J., & Blanco, D. (2007). Two-way and monolingual English immersion in preschool education: An experimental comparison, *Early Childhood Research Quarterly, 22,* 277-293.

Card, D. (1999). Chapter 30 The causal effect of education on earnings. *Handbook of Labor Economics, 3,* 1801-1863.

Collier, V.P., & Thomas, W.P. (2004) The astounding effectiveness of dual language education for all. George Masson University.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49,* 221-251.

Cummins, J. (1999). Alternative paradigms in bilingual education research: Does theory have a place? *Educational Researcher, 28,* 26-32.

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire.* Clevedon, UK: Multilingual Matters.

Dee, T.S. (2004). Are there civic returns to education? *Journal of Public Economics, 88,* 1697-1720.

Fuligni, A.J. (1997). The academic achievement of adolescents from immigrant families: The roles of background, attitudes, and behavior. *Child Development, 68,* 351-363.

Gándara, P. & R. Rumberger (forthcoming). *Immigration, Language, and Education: How Does Language Policy Structure Opportunity?* Teachers College Record.

Genesee, F., Geva, E., Dressler, C., Kamil, M.L. (2006). Cross-linguistic relationships in second-language learners. In. D. August & T. Shanahan (Eds). Developing reading and writing in second-language learners: Lessons from the report of the national literacy panel on language-minority children and youth. Mahwah, NJ: Lawrence Earlbaum Associates.

Geva, E. (2006). Learning to read in a second language: Research, implications, and recommendations for services. In: Tremblay, R.E., Barr, R.G., & Peters, R.D.V., eds. *Encyclopedia on Early Childhood Development,* 1-12. Available at: http://www.child-encyclopedia.com/documents/GevaANGxp.pdf . Accessed January 20, 2014.

Goldbenberg, C. (2008). Teaching English language learners. What the research does and does not say. *American Educator. Summer 2008.* 8-44.

Goldenberg, C. (1996). The education of language-minority students: Where are we, and where do we need to go? *The Elementary School Journal, 96,* 353-361.

Greene, J.P. (1998). A meta-analysis of the effectiveness of bilingual education. Claremont, CA: Thomas Rivera Policy Institute.

Hakuta, K. Butler, Y.G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy Report 2000-1). Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Kieffer, M.J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher, 39,* 484-486.

Lado, R. (1964). *Language teaching: A scientific approach.* New York: McGraw-Hill.

Maldonado, J.R. (1977). *The effect of the ESEA Title VII program on the cognitive development of Mexican American American students.* Unpublished doctoral dissertation, University of Houston, Houston, TX.

National Center for Education Statistics (2011). NAEP Data Explorer. Composite Scale Status of English Language Learner, 2 categories. Average scale scores & standard deviations. Personal Report. Available at: http://nces.ed.gov/nationsreportcard/naepdata/

National Center for Education Statistics (NCES, 2009). Table A-6-1. Number and percentage of children ages 5-17 who spoke only English at home, who spoke a language other than English at home and who spoke English with difficulty, and percent enrolled in school: Selected years, 1980-2009. Available: http://nces.ed.gov/programs/coe/tables/table-lsm-1.asp

Porter, R.P. (1990). *Forked tongue: The politics of bilingual education*. New York Basic Books.

Ramirez, J. D., Yuen, S., Ramey, D., & Pasta, D. (1991). Longitudinal study of structured English immersion strategy, early-exit and late-exit bilingual education programs for language-minority children (Final Report, Vols. 1 & 2). San Mateo, CA: Aguirre International. (ED 330 216)

Reardon, S.F., & Galindo, C. (2009). The Hispanic-White Achievement Gap in Math & Reading in Elementary Grades. *American Educational Research Journal, 46,* 853-891.

Rossell, C.H., & Baker, K. (1996). The effectiveness of bilingual education. *Research in the Teaching of English, 30,* 7-74.

Slavin, R.E., & Cheung, A. (2005). A synthess of research on language of reading instruction.

Review of Educational Research, 75, 247-284.

Slavin, R.E., Madden, N., Calderon, M., Chamberlain, A., & Hennessy, M. (2010). *Reading and language outcomes of a five-year randomized evaluation of transitional bilingual education.* Baltimore, MD: Johns Hopkins University.

Thomas, W., & Collier, V. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement.* Santa Cruz, CA and Washington, DC: Center for Research on Education, Diversity & Excellence. Available: http://www.crede.ucsc.edu/research/llaa/1.1_final.html

Willig, A.C. (1985). A meta-analysis of selected students on the effectiveness of bilingual education. *Review of Educational Research, 55*, 267-317.

Tables

Table 1. Description of the four ELL academic programs offered in the district of study.

Program	English Immersion	Transitional Bilingual	Developmental Bilingual	Dual Immersion
Program Intention	To support language & academic development with only English instruction for low-incidence ELL groups or for students whose parents want their children to be in English Immersion.	To develop English proficiency and academic mastery with primary language support to access the core curriculum as needed.	To develop competency in English while maintaining native language proficiency (i.e. bilingualism) and academic competency.	To help native speakers, bilingual students, and English-only students become fluent in both languages.
Population Served	ELL students served in classrooms with only English instruction	100% ELL or language minority. Students typically begin to transition out by 3rd grade, even if not yet reclassified as English proficient.	100% ELL or language minority. Students may transition out of this program upon reclassification (commonly 5th grade)	1/3 – 1/2 not proficient in the target language 2/3 – 1/2 proficient in the target language.
Instructional Time	100% in English. ELLs receive at least 30 minutes a day of English Language Development coursework.	K: 50-90% target depending on students' proficiency. The proportion of time spent in English increases at quick pace.	K: 50-90% target depending on students' proficiency. - Proportion English increases each year depending on the students.	K-1st: 80-90% in target language By 5th: 50% in English & 50% in target language.

Table 2. Proportion of students in the same pathway as their initial pathway, by grade.

	Grade 2	Grade 3	Grade 4	Grade 5
English Immersion	0.995	0.987	0.971	0.965
Transitional Bilingual	0.997	0.896	0.584	0.559
Developmental Bilingual	0.99	0.891	0.711	0.646
Dual Immersion	0.999	0.966	0.931	0.917
All Programs	0.995	0.952	0.852	0.83

Table 3. Proportions of ELs of each ethnicity initially attending each program; proportions of total ELs initially enrolled in each program, and average initial English proficiency, by program.

	English Immersion	Transitional Bilingual	Developmental Bilingual	Dual Immersion	Overall Proportion of ELs of each Ethnicity
	Proportion of students of each ethnicity initially in each program				
Latino	0.214	0.369	0.504	0.716	0.331
Chinese	0.468	0.562	0.434	0.139	0.454
Japanese	0.015	0.000	0.000	0.002	0.009
Korean	0.008	0.000	0.000	0.023	0.007
Filipino	0.052	0.002	0.017	0.005	0.033
Other Ethnicity	0.242	0.066	0.045	0.114	0.166
	Proportion of total ELs with each initial program				ALL ELs
Proportion of total ELs, by program	0.187	0.165	0.081	0.567	1.000
Average Initial English Proficiency, by program	0.486	0.682	0.332	0.354	0.547
N (students)	7,793	2,573	2,271	1,113	13,750

Note: Initial English proficiency is standardized relative to the state average for that year and grade, which is why none appear in negative standard deviation units.

Table 4. Proportions of students initially enrolling in each program, by year of kindergarten entry into the district.

Entry year	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total
English Immersion	0.521	0.547	0.567	0.572	0.592	0.561	0.579	0.581	0.578	0.567
Transitional Bilingual	0.217	0.177	0.186	0.194	0.18	0.196	0.188	0.181	0.168	0.187
Developmental Bilingual	0.231	0.237	0.188	0.155	0.13	0.148	0.14	0.124	0.136	0.165
Dual Immersion	0.031	0.038	0.059	0.079	0.098	0.095	0.092	0.114	0.117	0.081
N	1452	1558	1519	1484	1484	1555	1403	1566	1728	13750

Table 5. Estimated parameters of average English language arts (ELA) and Math growth trajectories, by initial (or predicted initial) program attended.

	ELA				Math			
	Model 1: Descriptive	Model 2: Student Controls	Model 2: Restricted Sample	Model 3: Student Controls & Preferences	Model 1: Descriptive	Model 2: Student Controls	Model 2: Restricted Sample	Model 3: Student Controls & Preferences
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
				Intercepts				
Intercept (Average for English Immersion)	-0.021	-0.023	-0.03	-0.011	0.095+	0.116**	0.176***	0.144***
	(0.049)	(0.028)	(0.034)	(0.027)	(0.056)	(0.036)	(0.036)	(0.028)
Transitional Bilingual (TB)	0.034	0.140**	0.164*	0.154**	0.184+	0.290***	0.289***	0.271***
	(0.092)	(0.054)	(0.067)	(0.055)	(0.105)	(0.071)	(0.071)	(0.057)
Developmental Bilingual (DB)	-0.229*	-0.047	-0.109	-0.065	-0.148	0.059	0.058	0.064
	(0.093)	(0.055)	(0.077)	(0.064)	(0.107)	(0.081)	(0.082)	(0.067)
Dual Immersion (DI)	-0.394***	-0.163*	-0.179*	-0.244***	-0.258+	0.041	0.043	-0.064
	(0.117)	(0.071)	(0.082)	(0.071)	(0.134)	(0.087)	(0.087)	(0.072)
				Slopes				
Grade (Average for Engilsh Immersion)	0.010+	0.011*	0.003	0.001+	-0.008	-0.009	-0.039**	-0.037***
	(0.005)	(0.005)	(0.009)	(0.007)	(0.010)	(0.010)	(0.012)	(0.010)
Transitional Bilingual (TB) X Grade	-0.017	-0.020+	-0.02	-0.008	-0.044*	-0.044*	-0.048+	-0.042*
	(0.010)	(0.010)	(0.018)	(0.014)	(0.019)	(0.019)	(0.025)	(0.020)
Developmental Bilingual (DB) X Grade	0.005	0.006	0.003	-0.003	-0.024	-0.016	-0.025	-0.023
	(0.011)	(0.011)	(0.020)	(0.017)	(0.019)	(0.019)	(0.029)	(0.024)
Dual Immersion (DI) X Grade	0.055***	0.057***	0.083***	0.092***	0.008	0.013	0.036	0.044+
	(0.015)	(0.015)	(0.022)	(0.019)	(0.026)	(0.025)	(0.031)	(0.027)
Student random intercepts & slopes	X	X	X	X	X	X	X	X
L2 Stable Student Controls		X	X	X		X	X	X
L2 School-Program Preference Controls				X				X
L3 School * EL Instructional Program		X	X	X		X	X	X
N (observations - Level 1)	65,912	65,912	28,428	28,428	55,499	55,499	27,386	27,386
N (students - Level 2)	13,750	13,750	7,729	7,729	13,750	13,750	7,729	7,729
N (School * ELL Program - Level 3)	191	191	150	150	191	191	150	150

Notes: Stable student controls include gender, ethnicity, and initial English proficiency score. All models allow students' individual intercepts and slopes to vary. The reference category is English Immersion, and as such the intercept and grade terms represent the average starting point and trend for those initially attending this program. Grade slopes for ELA represent an effect from grades 2-8 for models 1 & 2 and grades 2-5 for models 2 restricted and 3. Neighborhood random effects represent the closest school to the child in the year of kindergarten attendance. These are similar to attendance zones, but do not necessarily reflect the school the student actually attended.

Table 6. Estimated parameters of average academic growth rates, by initial program attended, ethnicity, and initial English proficiency.

	ELA			Math		
	Model 2: Student Controls b/se	Model 2: Restricted Sample b/se	Model 3: Student Controls & Pref-erences b/se	Model 2: Student Controls b/se	Model 2: Restricted Sample b/se	Model 3: Student Controls & Pref-erences b/se
	Slopes					
Grade	0.009+	0.003	0.001	-0.012	-0.040***	-0.037***
	(0.005)	(0.008)	(0.006)	(0.009)	(0.012)	(0.009)
Grade X Chinese	0.034***	0.023*	0.030**	0.055***	0.065***	0.066***
	(0.005)	(0.010)	(0.010)	(0.008)	(0.013)	(0.013)
Grade X Initial English Proficiency	-0.006**	-0.014***	-0.013**	0	-0.014**	-0.014***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
TBE X Grade	-0.018+	-0.027	-0.017	-0.042*	-0.054*	-0.050**
	(0.010)	(0.017)	(0.014)	(0.018)	(0.024)	(0.020)
DBE X Grade	-0.001	-0.002	-0.010	-0.032+	-0.038	-0.040+
	(0.010)	(0.020)	(0.016)	(0.019)	(0.028)	(0.028)
DI X Grade	0.065***	0.072**	0.079***	0.022	0.047	0.042
	(0.015)	(0.023)	(0.020)	(0.027)	(0.033)	(0.028)
TBE X Grade X Chinese	-0.069***	-0.081***	-0.060**	-0.090***	-0.129***	-0.114***
	(0.011)	(0.021)	(0.021)	(0.017)	(0.027)	(0.028)
DBE X Grade X Chinese	-0.067***	-0.051*	-0.064**	-0.099***	-0.130***	-0.137***
	(0.012)	(0.026)	(0.026)	(0.019)	(0.034)	(0.036)
DI X Grade X Chinese	-0.023	-0.062+	-0.063*	-0.055	-0.07	-0.070
	(0.024)	(0.036)	(0.033)	(0.038)	(0.048)	(0.045)
TBE X Grade X Initial English Proficiency	-0.002	-0.001	-0.001	-0.011*	-0.002	-0.002
	(0.004)	(0.006)	(0.006)	(0.005)	(0.008)	(0.008)
DBE X Grade X Initial English Proficiency	-0.001	0.008	0.006	-0.006	0.002	0.001
	(0.004)	(0.007)	(0.007)	(0.006)	(0.009)	(0.009)
DI X Grade X Initial English Proficiency	-0.001	0.017*	0.015+	-0.012	-0.005	-0.003
	(0.006)	(0.008)	(0.008)	(0.008)	(0.010)	(0.010)
L2 Student random intercepts & slopes	X	X	X	X	X	X
L2 Stable Student Controls	X	X	X	X	X	X
L2 School-Program Preference Controls			X			X
L3 School * EL Instructional Program	X	X	X	X	X	X
Joint test of Chinese Intercepts (p-value)	0.001	0.052	0.044	0.000	0.017	0.016
Joint test of Chinese Slopes (p-value)	0.000	0.000	0.000	0.000	0.000	0.000
Joint test of English Prof intercepts (p-value)	0.025	0.111	0.139	0.662	0.351	0.414
Joint test of English Prof slopes (p-value)	0.663	0.133	0.114	0.142	0.937	0.921
N (observations - Level 1)	65,912	28,428	28,428	55,499	27,386	27,386
N (students - Level 2)	13,750	7,729	7,729	13,750	7,729	7,729
N (School * ELL Program - Level 3)	191	150	150	191	150	150

* Notes: Stable student controls included gender, ethnicity, and initial English proficiency score. All models allow students' individual intercepts and slopes to vary. The reference category is English Immersion such that the slopes represent the average trend for those initially attending this program. Grade slopes for ELA represent an effect from grades 2-8 for model 2 and grades 2-5 for models 2 restricted and 3. Grade slopes for math represent an effect from grades 2-6 for model 2 and grades 2-5 for models 2 restricted and 3.

Figures

Figure 1.

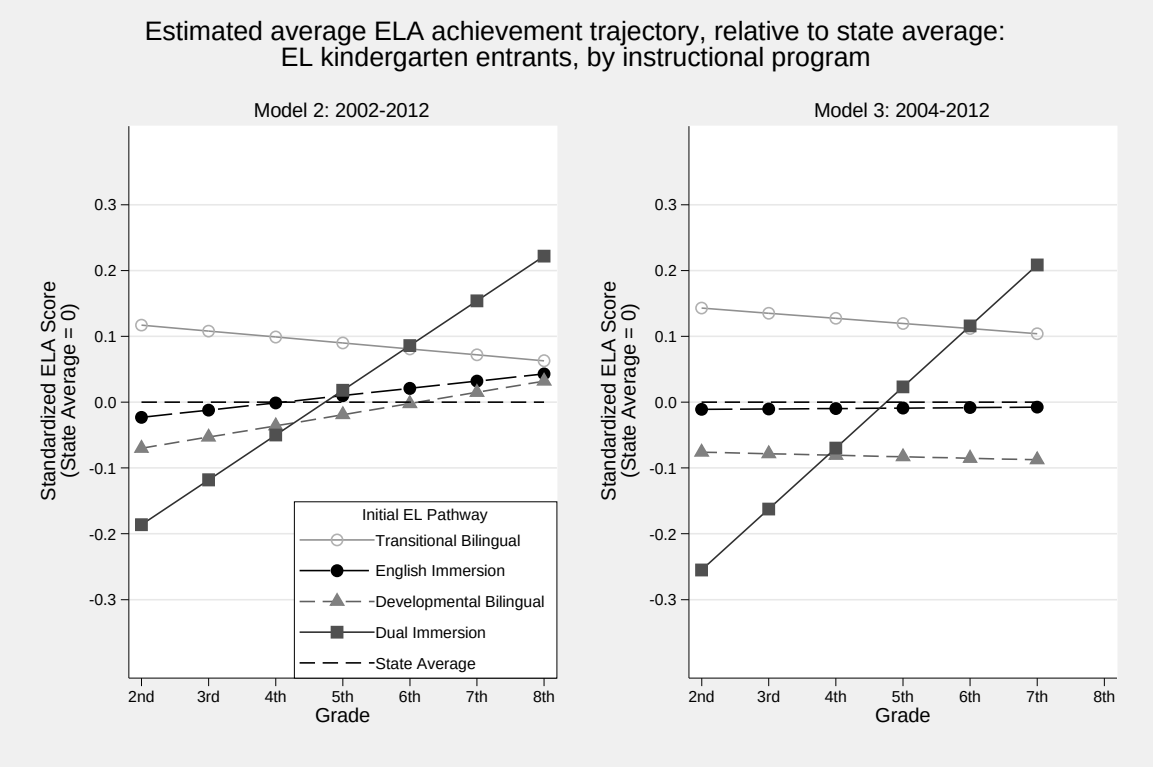


Figure 2.

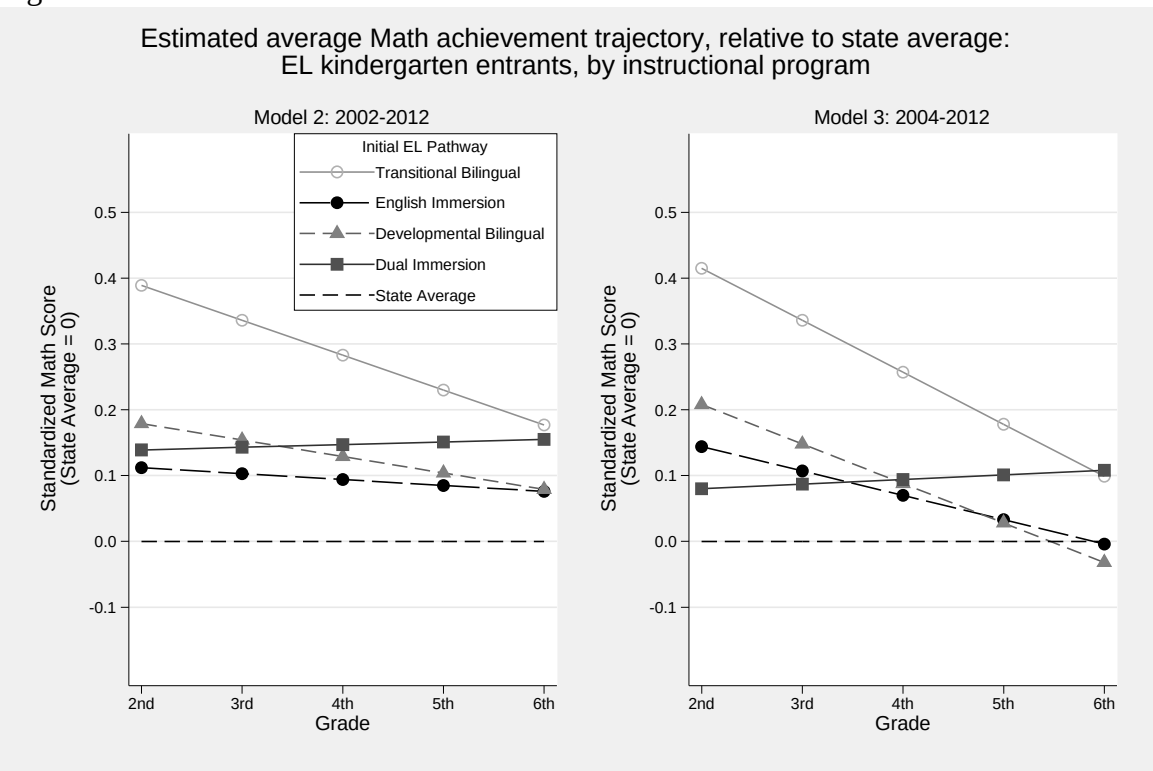


Figure 3.

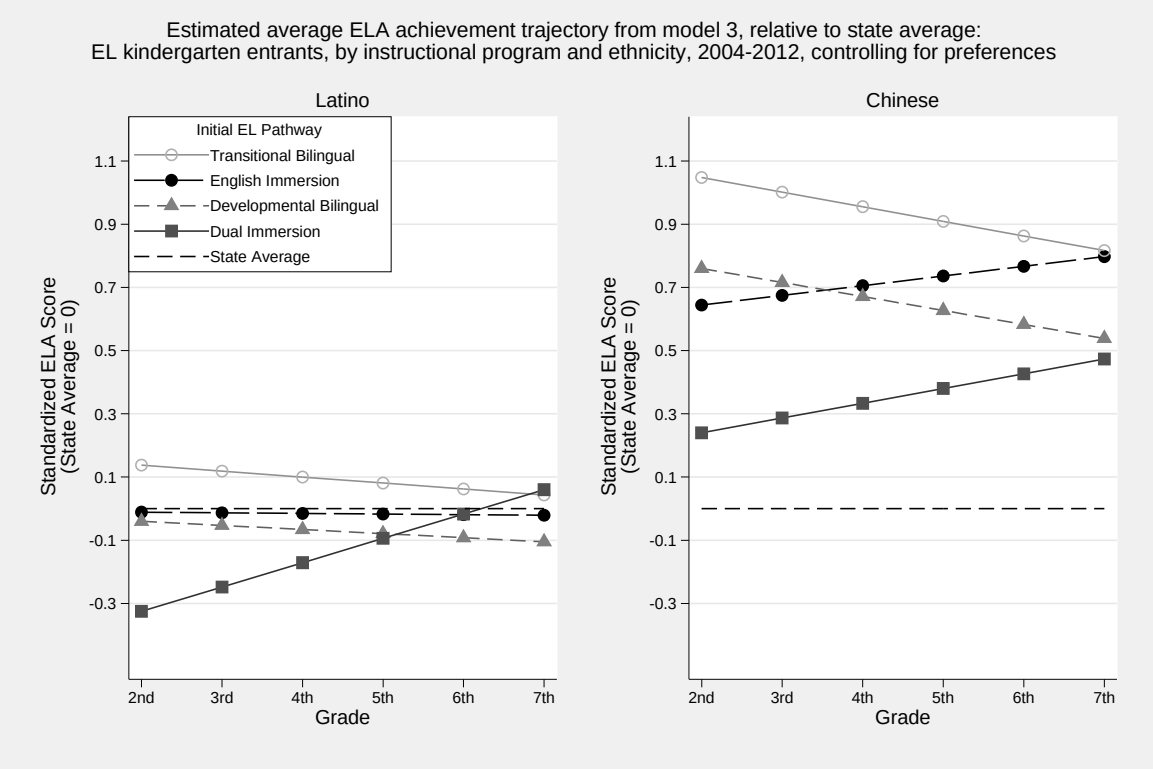
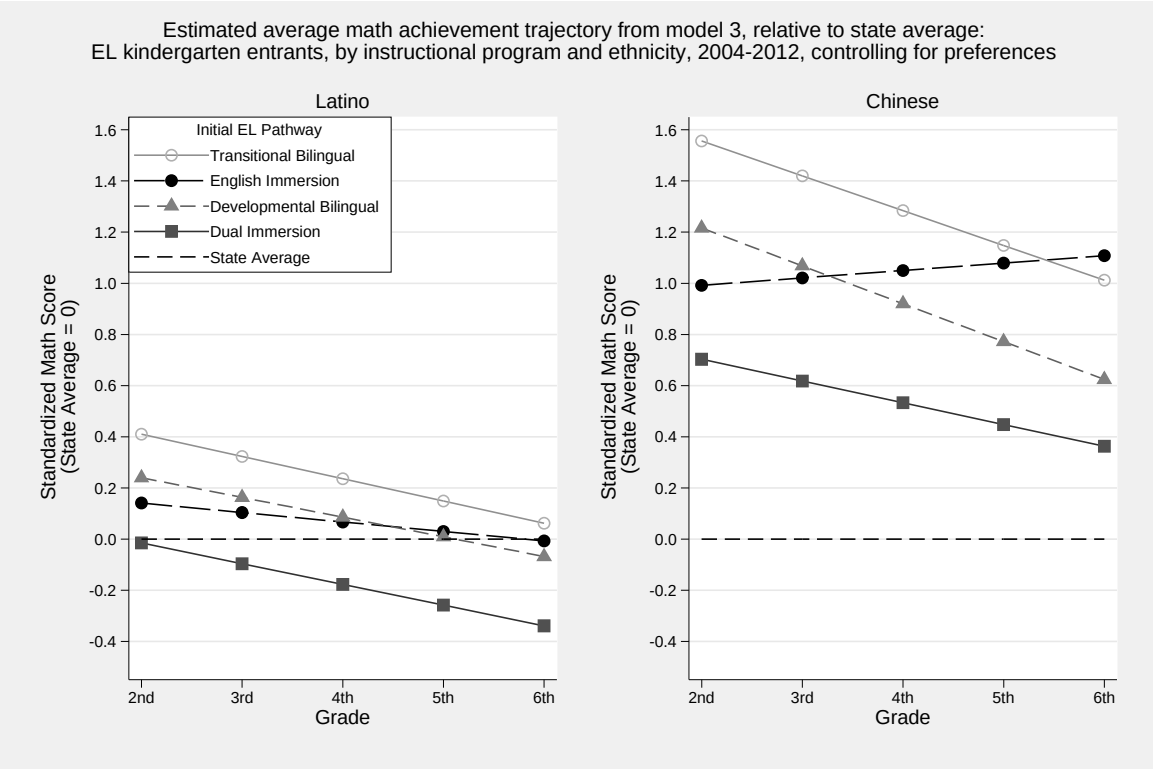


Figure 4.



Appendix A. Instrumental Variables Approach to Research Question One

IV Methods

To answer question one, in addition to our models controlling for preferences, we also conduct an additional exploratory quasi-experimental analysis, using an instrumental variable (IV) estimation approach in a two-stage least squares (2SLS) framework. We consider this approach exploratory because we do not have a means of testing our exclusion restriction and the standard errors in our models sizably increase, and because our sample size is too small, we don't have enough power to estimate these models separately by ethnicity.

The use of IV to infer causality relies heavily on the selection of an instrument that generates exogenous variation in the initial program students attend but that is not otherwise a direct predictor of students' academic outcomes. We use the initial addresses of students to construct three instrumental variables of the geographic proximity of students to each instructional program type (TB, DB, and DI, with EI selected as the omitted category). We define proximity as "crow-flies" distance, or the shortest distance between two points (i.e. the shortest distance between where a student lives and a given program). This particular distance IV creates exogenous variation in the kinds of programs students attend based on where they live. These types of instrumental variables are commonly used in other research that estimates returns to higher education (e.g. Dee, 2004; Angrist & Krueger, 1999; Card, 1999). For the exclusion restriction to be met in this model, the distance that students live to a particular program would have to be highly correlated with the program they ultimately attend, but not a direct predictor of their academic outcomes after conditioning on our set of controls. This is certainly plausible.

In the first stage equations of the IV models, we regress indicators of students' initial program on a set of variables indicating students' proximity to each program in their

kindergarten year and a vector of the other control variables included in the second stage equation. Note that the proximity variables indicate the distance to the nearest program of a specific type that enrolls EL students of the same target language. Because there are few EL instructional programs using languages other than Spanish, Cantonese, or Mandarin, we include only Latino and Chinese students in these models. We have three (rather than four) first stage equations because the English immersion program is our omitted category. These first stage equations take the following form:

First Stage equations:

$$\begin{aligned} TB_i &= \gamma_{10} + \gamma_{11}TBprox_i + \gamma_{21}DBprox_i + \gamma_{31}DIprox_i + \gamma_{14}EIprox_i + \mathbf{X}_i\boldsymbol{\Delta}_1 + e_{1i} \\ DB_i &= \gamma_{20} + \gamma_{21}TBprox_i + \gamma_{22}DBprox_i + \gamma_{23}DIprox_i + \gamma_{24}EIprox_i + \mathbf{X}_i\boldsymbol{\Delta}_2 + e_{2i} \\ DI_i &= \gamma_{30} + \gamma_{31}TBprox_i + \gamma_{32}DBprox_i + \gamma_{33}DIprox_i + \gamma_{34}EIprox_i + \mathbf{X}_i\boldsymbol{\Delta}_3 + e_{3i} \end{aligned} \quad (2)$$

Table 7 presents the results from our first stage equations, which are consistent with the assumption that proximity to a particular program may provide a valid source of identification. The distance one lives to each program is negatively correlated with whether or not they actually attended that program. In other words, as the distance a student lives from the closest TB program increases, his/her likelihood of having attended that program decreases. This is true for all instructional models. Further, the distance one initially lived to other programs is often also a good predictor of the program students actually attended. In all cases, the set of proximity variables are jointly strong instruments for each program (i.e., in each case the F-statistic of the null hypothesis that the coefficients on all four of the proximity variables are zero is well above 10, the threshold value for determining whether a set of instruments is jointly strong enough to avoid finite sample bias).

[Insert Table 7 about here]

To consider results from the second stage, we estimate three-level HLM models of the following form:

$$
\begin{aligned}
&Level\ 1: && Y_{tin} = \alpha_{0in} + \alpha_{1in} G_{itn} + e_{tin} \\
&Level\ 2: && \alpha_{0in} = B_{00n} + \widehat{\boldsymbol{P}}_{\boldsymbol{in}}\boldsymbol{\Gamma_0} + \boldsymbol{X_{in}}\boldsymbol{B_0} + r_{0in} \\
& && \alpha_{1in} = B_{10n} + \widehat{\boldsymbol{P}}_{\boldsymbol{in}}\boldsymbol{\Gamma_1} + \boldsymbol{X_{in}}\boldsymbol{B_1} + r_{1in} \\
&Level\ 3: && B_{00n} = \gamma_{000} + u_{00n} \\
& && B_{10n} = \gamma_{100} + u_{10n}
\end{aligned}
\qquad (3)
$$

Here, we model the intercept (α_{0in}) and slope (α_{1in}) of the test score trajectory of student i in neighborhood n as linear functions of the student's vector of predicted probabilities of enrolling in each instructional program ($\widehat{\mathbf{P}}_{in}$) and a vector of student characteristics ($\mathbf{X}_{in}$), plus student- and neighborhood-specific random effects.

We consider this IV approach exploratory, as we do not have adequate pre-test scores to test the validity of the model by testing to see whether distance one lives significantly predicts academic scores *before* program attendance. Still, the instruments are strong and so we believe that the results warrant consideration. We also fit Model 2 from Table 5 using the IV sample, and label it "Model 2: IV sample" in Table 8.

IV Second Stage Results

Our IV models are generally much more imprecisely estimated than the non-IV models. When comparing "Model 2: IV Sample" to "Model 4: IV Second Stage" in Table 8, we see that for intercepts, the standard errors in Model 4 are at least three times as large as those in Model 2, both for ELA and math. In some cases, they are as much as ten times as large. Slopes are a bit more precisely estimated, but still the standard errors at least tripled from Model 2 to Model 4 in most cases. The IV estimates in Model 4 have such wide confidence intervals that even when they look different than the estimates from Model 2 in Table 8, the confidence intervals for the estimates in Model 4 include the Model 2 coefficients. It is therefore difficult to know whether

the IV models are mostly imprecise and not very informative, or whether they point to actual null results. The one exception to this seems to be our findings for the TB program. Here, the IV models provide dramatically different results than the non-IV models, both for intercepts in ELA and math and for slopes in math. If we were to believe these results as causal, they would suggest that in ELA, ELs initially enrolled in TB score significantly lower than those in EI in second grade, and have scores that do not grow at a rate that is statistically distinguishable from 0 through grade 8. In math, our IV results indicate that the second grade scores of ELs in TB are not significantly different from those in EI (in contrast to our finding in Model 2 showing that they are significantly higher than those of their peers in EI). We also find that their math scores do not grow at a rate that is different from that of ELs in EI (in contrast to our finding in model 2 that their scores grow significantly more slowly than their peers in EI). Because of our inability to accept these IV results with strong confidence, we choose to generally rely on those presented as Model 2 of the main paper for the purpose of interpretation.

[Insert Table 8 about here]

Table 8. First Stage Results for Instrumental Variables Models

	Initial Program Attended		
	Transitional Bilingual (TB) b/se	Developmental Bilingual b/se	Dual Immersion b/se
Proximity to TB	-0.094***	0.020***	-0.002
	(0.004)	(0.004)	(0.003)
Proximity to DB	0.020***	-0.085***	0
	(0.003)	(0.003)	(0.002)
Proximity to DI	0.028***	0.017***	-0.021***
	(0.002)	(0.002)	(0.002)
Proximity to EI	0.030***	0.015+	0.015*
	(0.009)	(0.008)	(0.006)
Student Controls	X	X	X
N (Students)	11,803	11,803	11,803
First Stage F-Statistic	170.13	210.33	42.83

Table 9. Second Stage Instrumental Variable Results

	ELA		Math	
	Model 2: IV Sample b/se	Model 4: IV Second Stage b/se	Model 2: IV Sample b/se	Model 4: IV Second Stage b/se
	Intercepts			
Intercept (Average for English Immersion, EI)	-0.218***	-0.130*	-0.110***	-0.01
	(0.015)	(0.054)	(0.016)	(0.058)
Transitional Bilingual (TB)	0.200***	-0.153+	0.316***	-0.042
	(0.019)	(0.093)	(0.020)	(0.106)
Developmental Bilingual (DB)	-0.040*	-0.234**	0.092***	-0.085
	(0.020)	(0.091)	(0.021)	(0.102)
Dual Immersion (DI)	-0.120***	0.091	0.03	0.076
	(0.027)	(0.297)	(0.029)	(0.312)
	Slopes			
Grade (Average trends in EI)	0.018***	0.002	-0.006	-0.021
	(0.003)	(0.013)	(0.005)	(0.020)
Transitional Bilingual (TB) X Grade	-0.036***	-0.025	-0.058***	0.032
	(0.004)	(0.022)	(0.006)	(0.037)
Developmental Bilingual (DB) X Grade	0	0.029	-0.023***	0.027
	(0.004)	(0.020)	(0.006)	(0.032)
Dual Immersion (DI) X Grade	0.046***	0.115+	0.015	-0.108
	(0.006)	(0.065)	(0.010)	(0.102)
L2 Stable Student Controls	X	X	X	X
L2 Student random intercepts & slopes	X	X	X	X
L3 Neighborhood	X	X	X	X
Uses proximity IV to predict initial path		X		X
N (observations - Level 1)	53,545	53,545	45,094	45,094
N (students - Level 2)	11803	11803	11803	11803
N (neighborhoods - Level 3)	81	81	81	81

* Notes: Stable student controls included gender, ethnicity, initial English proficiency score, and free and reduced price lunch status. All models allow students' individual intercepts and slopes to vary. The reference category is English Immersion such that the intercept and grade terms represent the average trend for those initially attending this program. Neighborhood random effects represents the closest school to the child in the year of kindergarten attendance. These are similar to attendance zones, but do not necessarily reflect the school the student actually attended. Grade slopes for ELA represent an affect by grade from grade 2-8. For math this is estimated for grades 2-6.