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

Rachel A. Valentino

Sean F. Reardon

*Stanford University Graduate School of Education*

December, 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 Kenji Hakuta, Ilana Umansky, Sandy Nader, Camille Whitney, Christopher Candelaria, and Lindsay Fox for their invaluable feedback on earlier versions of this research and paper. Please direct questions to rachel.valentino@stanford.edu.

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

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

Abstract

This paper investigates the differences in academic achievement trajectories from elementary through middle school among English Learner students in four different instructional programs: English Immersion, Transitional Bilingual, Developmental Bilingual, and Dual Immersion programs. Comparing students with the same parental preferences but who attend different programs, we find that the ELA test scores of ELs in all bilingual programs grow at least as fast as, if not faster than those in English immersion. The same is generally true of math, with the exception of developmental bilingual programs, where average student scores grow more slowly than those of students in English immersion. Further, Latino ELs perform better longitudinally in both subjects when in bilingual programs than their Chinese EL counterparts. We find no differences in program effectiveness by ELs' initial English proficiency.

Over 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 more than doubled (NCES, 2011; Census, 2012). 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 long-running 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 bilingual education altogether (Rolstad, Mahoney, & Glass, 2005).

On one hand, 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 (Rossell & Baker, 1996; Porter, 1990; Baker 1998). On the other hand, 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; See also Genesee, Geva, Dressler, & Kamil, 2008.). This perspective also stresses that academic content in subjects like math and science may be lost in translation when instruction is not in students' first language.

There is slightly more empirical support for the latter argument, suggesting that bilingual education is superior to English-only instruction for ELs (Rolstad et al., 2005; Goldenberg, 2008; Greene, 1998; Willig, 1985), or at a minimum not detrimental (see Slavin & Cheung, 2005). However, much of the research on the issue is not very rigorous (see Rossell & Baker, 1996). 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., 2011); and much of it is based on studies conducted on French Immersion programs in Canada (Cummins, 1999) 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 ELs in classrooms 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 longitudinal student-level data from a large school district and more rigorous methods to address 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 achievement trajectories in English Language Arts (ELA) and math through middle school? and (2) Do these growth effects by program vary by the ethnicity or initial English proficiency of the EL student?

## Review of the Literature

### *Theoretical Perspective*

Two theoretical perspectives frame the debate about the benefits and drawbacks of bilingual education. One perspective argues 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). The contrary perspective argues that spending more time-on-task with maximum exposure to English language instruction results in quicker acquisition of and better performance in English (Rossell & Baker, 1996; Porter, 1990; Baker 1998).

The first perspective—that bilingual education is preferable to English immersion instruction—is based in two arguments. First, 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.

Second, 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; Genesee et al., 2008). 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 school in a new country. These findings are consistent with the idea that children should learn to read in their home language first, rather than learning to read in general and read in a new language simultaneously (Cummins, 1999).

The second perspective—that English Immersion classrooms are better for English Learners than bilingual classrooms—is based in the argument that spending instructional time in a language other than English necessarily detracts from students' exposure to English. Given that the primary language of instruction in U.S. schools is English, the argument goes, delaying students' development of English skills delays their opportunity to learn academic material. To date, research has not yet consistently supported either hypothesis. In part this is due to the fact that much of the research relies on research designs that do not provide a strong causal warrant. Moreover, bilingual education is implemented in different ways in different studies. These factors make it difficult to draw any firm conclusions about the relative benefits of bilingual instruction and English immersion instruction. In the section that follows, we attempt to highlight the most rigorous studies to date.

*Effectiveness of bilingual instruction*

Bilingual education has been shown to influence a number of student outcomes. These include both oral and written language development, rate of reclassification as fluent English proficient, and academic course-taking patterns (Jepsen, 2010; Saunders & O'Brien, 2006; Riches & Genesee, 2006; Umansky & Reardon, 2014; Umansky, 2014). Since this paper considers academic outcomes in ELA and math, we focus here on reviewing literature that considers effects on academic outcomes.

There is a sizable body of literature documenting the effects of bilingual education compared to English immersion instruction on ELs' academic performance (Lindholm-Leary & Borsato, 2006). A handful of reviews and meta-analyses have tried to summarize the literature, but the conclusions of these meta-analyses vary depending on the study inclusion criteria they use. In a review of studies comparing bilingual programs with English immersion ones, Rossell

and Baker (1996) find that about 30% of the studies show that bilingual education is worse than English immersion for reading outcomes, 20% show that it is better than English immersion, and the remaining 50% find that there is no difference between the two. Findings for math are similar. While comprehensive and effective at highlighting the mixed nature of research on bilingual education, this review relies heavily on the effectiveness of French immersion programs in Canada, the results of which may not generalize to bilingual programs in the U.S. Further, although studies were restricted to those including a comparison group, most did not rely on experimental or quasi-experimental research designs.

The two meta-analyses that used the most stringent study inclusion criteria generally 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 findings from these meta-analyses tend to suggest that bilingual instruction leads to equal or better academic outcomes than English-only instruction, with the exception of a few studies, many of these studies relied on small locally specific samples leading to limited generalizability, and most only tracked student outcomes for a few years at best. Further, much of this literature does little to tease apart the differential effectiveness of specific bilingual instructional models (e.g. transitional vs. developmental bilingual), making it difficult to disentangle which components make bilingual programs work. 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. We review the evidence on the differential effectiveness of these models below.

*Transitional bilingual.* 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, with the intention of transitioning ELs into English immersion programs quickly – usually by grade two or three. Transitional programs use ELs' home languages to support learning, but do not have a goal of promoting bilingualism prior to transitioning to English immersion.

In a longitudinal quasi-experimental study, Matsudaira (2005) uses a regression discontinuity design to estimate the effect of enrolling in a transitional bilingual[1] education class. The analysis finds negligible effects of bilingual education in ELA and math across grades three through eight. However, because the estimates are based on a regression-discontinuity design, the findings apply only to ELs with relatively high levels of English proficiency (that is, those scoring just below the cut score of EL classification), making it difficult to know whether the findings would generalize to ELs with lower initial English proficiency. Further, in the district studied, there was considerable movement of students in and out of bilingual programs; only 30% of ELs remained in a bilingual program for two or more years. It is possible that if there was higher compliance of students attending programs for more years, the effects would be different.

[1] Although the specific approach to bilingual instruction used in this district was not specified in the paper, given that the majority of ELs transition out of bilingual education by 4th grade (only 15 percent remain), we assume that this is a study of a transitional bilingual program.

Slavin et al (2011) randomly assigned students to a transitional bilingual or an English immersion program and tracked Spanish-speaking ELs' reading and vocabulary achievement from kindergarten through fourth grade. They found that in the early grades, ELs enrolled in an English immersion program outperformed their peers who attended transitional bilingual programs in academic outcomes in English, but by fourth grade, no significant differences on these assessments emerged (Slavin, Madden, Calderon, Chamberlain, & Hennessy, 2011). These findings suggest 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."

*Developmental bilingual*. Other research has compared the effects of transitional bilingual programs with those of developmental bilingual programs. 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.

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 (2011), the authors found that in early grades, students attending transitional and developmental bilingual programs performed worse in ELA than their peers enrolled in English immersion classrooms, but by second grade this significant difference

disappeared. They also found that by sixth grade, ELs in English immersion actually appeared to fall further behind their peers in bilingual programs. The findings from this study should, however, be interpreted with caution, as the authors' matching algorithm did not account for students' pre-test scores.

*Dual Immersion*. The above studies do little, however, to shed light on the potential benefits of dual immersion instructional programs, which to date have not been as extensively researched. Dual immersion programs are more similar to developmental than transitional bilingual instruction because they hold a goal of facilitating biliteracy through longer term program, but they differ in that they enroll both native English speakers and ELs in the same classroom. In some ways, dual immersion programs can be thought of as a hybrid approach of English immersion and developmental bilingual instruction, as they are based on the notion that the integration of native speakers of both languages into a single classroom offers students the opportunity to learn with students who model high quality language in the language they are not yet proficient in (Valdés, 1998). In some dual immersion models, regardless of grade, approximately 50% of instructional time is spent in English and the other 50% is spent in the ELs' native language (often referred to as the target language). In others, and in the case of the dual immersion model in this paper, the majority of instruction occurs in the target language in the early elementary grades. This gradually becomes more balanced across each grade until late elementary school, at which point about half of the instructional time is spent in the target language and the other half is spent in English (Lindholm-Leary, 2005; Christian, 1998).

Two noteworthy studies consider the effects of such programs on students' outcomes. Thomas and Collier (2002) found that across five large school districts, ELs attending dual immersion programs almost always performed higher academically in English, Spanish, and

math than 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. This study provides good descriptive evidence of differences in EL students' performance across programs, but only controlled for a very limited set of student-level 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 related to their later academic outcomes.

The second study randomly assigned preschool students to either dual immersion or English-only preschool classrooms and found that by the end of first grade, dual immersion instruction led to significant gains in the Spanish language development of both language minority students and native English speaking children without loss to their development of academic skills in English (Barnett, Yarosz, Thomas, Jung, & Blanco, 2007). It is unclear whether the results of this study generalize to elementary school dual immersion programs, however, because the randomized treatment assignment was maintained only though the preschool year. Moreover, the study focused on language minority children in general, only some of whom might have classified as ELs once they enrolled in kindergarten.

Taken together, these studies yield quite mixed results, but suggest that at the very least, bilingual education (generally defined) does not hinder academic performance in English in the medium term.

*Motivation for the current study*

*Long-term effects by subject*. 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., 2011) 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 both English language arts (ELA) and math. We hypothesize that the two-language instructional programs will lead to slower initial growth, but faster later growth in ELA than will English immersion instruction because more exposure to English will lead to quicker acquisition of English language skills initially, but the transfer of skills across languages will allow students in bilingual programs to catch up after a few years. For math, however, several competing hypotheses seem plausible. On the one hand, we expect that two-language programs should enable faster acquisition of math skills than English-only programs because instruction in EL students' home languages will allow access to academic content. On the other hand, two-language programs may spend more instructional time in ELA than English immersion classrooms, and less time on math instruction, particularly if two-language programs enroll students with lower levels of English proficiency than English immersion programs. Finally, if performance on math tests is partly mediated by language skills,

and if ELs in two-language programs initially develop English language skills more slowly than those in English immersion programs (as we hypothesize above), ELs' test scores may not reflect their math skills in early elementary school as well for those in two-language programs as those in English immersion programs (because math tests are administered in English). This would make it appear that two-language programs lead to lower initial math skills than do English immersion programs. Because it is not clear which of these different mechanisms might dominate, we have no clear hypotheses about the effect of EL instructional programs on math.

*Effects by subgroup*. Most research that has been conducted on EL instruction in the U.S. focuses exclusively on the effectiveness of different instructional programs for Spanish-speaking ELs. Worse still, some studies treat all ELs as one undifferentiated category, without considering differences in students' home language and initial English proficiency. 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 et al., 2008; Lado, 1964). One potential reason might be that visual processes are 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 (and thus transfer) across languages (a reality that is more likely between typologically similar languages).

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 over several years 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 alphabets, we hypothesize that Chinese ELs in English immersion programs will perform better than their Chinese peers in bilingual programs. To our knowledge only one study to date has specifically estimated the differential effectiveness of bilingual instruction for Latino and Chinese ELs. Conger (2010) finds that bilingual instruction has a negative effect on English proficiency for both Latino and Chinese ELs. She argues, however, that the apparent similarity in program effects may be driven by differential selection processes, rather than by true similarities in the effects of bilingual education. We build on Conger's work by estimating program effects by ethnicity on academic trajectories (rather than English proficiency) separately for Latino and Chinese ELs.

In addition to estimating our models separately by ethnicity, we also test whether the effects of EL instructional programs differ by students' initial English proficiency. To our knowledge there is little research to date on this question, with the exception of a study by Jepsen (2010), which found that bilingual programs had positive effects on English proficiency among those students with high prior English listening/speaking proficiency, and negative effects among those with low prior proficiency. Jepsen (2010) did not examine academic outcomes, however. Because of the limited prior research in this area, we have no clear hypotheses about whether and how EL instructional program effects may differ in relation to ELs' initial English proficiency.

*Rigorous methods*. One challenge in the study of EL instruction is potential selection bias. Many of the studies reviewed here include only a small set of control variables in regression models to reduce selection bias, but because the selection process is generally unknown, it is not clear whether these variables provide sufficient controls. In our analyses we us random coefficients growth models with a relatively robust set of controls. Importantly, we are able to include a set of variables that directly control for parental preferences regarding the type of EL program they would like their child enrolled in. The school district where our research is based uses a complicated student assignment algorithm to assign EL students to schools and, within schools, to instructional programs. The algorithm takes parental preferences into account, but when schools and programs are oversubscribed, it relies on random assignment. Our models use this feature of the assignment process to estimate the effects of different programs, comparing the academic outcomes of ELs whose parents preferred the same school and program but who attended different programs. Because we can control explicitly for the parental preferences used in the algorithm, our results arguably have a somewhat stronger causal warrant than if we could control only for observable student characteristics.

Taken together, this study adds to the literature on the effects of EL instructional programs in several ways: 1) it estimates effects of four different EL programs; 2) it examines long-term program impacts on academic trajectories; 3) it examines differences in program effects by student ethnicity/home language and initial English proficiency; and 4) it uses a set of models that provide a stronger causal warrant than much of the research to date.

## Data and Methods

### *Data*

The data used in the current study come 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 between the 2001-2002 and 2009-2010 academic years. Approximately 1,500 ELs enter our sample each year. Our outcome data come from the state standardized tests in English language arts (ELA) and math that students took each year from second through up to eighth grade. We standardize these ELA and math scores relative to the state distribution within each grade and year, so all outcome test scores are reported in terms of standard deviations from the statewide mean. While we use ELA scores through eighth grade, we only analyze math scores through sixth grade. We do so because, starting in seventh grade, students may take a subject-specific math test (e.g. general math vs. Algebra). Because not all students enroll in the same level of math class in $7^{th}$ and $8^{th}$ grade, math scores in these grades are not comparable across students. All ELs in our analysis are observed through at least $3^{rd}$ grade, but we do not observe all students in our sample through $6^{th}$ or $8^{th}$ grade, because the later cohorts of kindergarteners had not yet reached the later grades by 2012, the last year for which we have outcome data.

*Program preferences*. Prior to the start of kindergarten (but after they have been assigned to a school and EL program), students are assessed to determine their English proficiency. The district of study implements a choice model for school selection, where families rank program (i.e. 191 instructional programs located within schools) preferences. Students are then assigned to schools by a complex algorithm that attempts to assign students to the school/program combination requested by their parents, subject to a set of school diversity constraints and a set of priority rules. 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). Among students with the same priority rankings, ties are

broken using random assignment. The tie-breaker process adds some randomness. Importantly, teachers and administrators—who might have knowledge of students' skills or needs—do not play a role in assigning students to schools or instructional programs within schools. As a result, there are students whose parents requested the same school/program combinations, but who were assigned to different EL programs through the priority rules or random assignment. By controlling for program preference fixed effects in our models, we can compare students who had the same school-by-program preferences, but attended different programs and/or schools due to the use of tie-breakers[2].

One concern related to our strategy is that families may be able to tamper with the lottery and/or may differentially leave the district if they aren't assigned to one of their top school/program preferences[3]. In our district of study there is little concern about tampering, as all school/program assignments are made by the algorithm, which is administered in the district's central office. Families can, however, submit a formal appeal of extenuating circumstances (e.g. medical issues) to be granted a new assignment. The district reported to us that such appeals affect a negligible portion (less than 1%) of students assigned to schools/programs each year. In addition to the primary assignment process, there is also a second much smaller lottery (involving roughly 10% of students) that occurs after the initial assignment process to accommodate (a) late district entrants, and (b) families who wish to enter a lottery of remaining slots because other individuals who entered the lottery neglected to enroll. Through this additional lottery process, approximately 5% of all students receive a higher choice than they

[2] Note that we are not controlling for random assignment to different programs, but rather are comparing students with the same preferences but who attend different programs. Only some of the difference in the programs students attend is due to random assignment, but the use of preference fixed effects controls for a substantial source of potential bias.

[3] In fact, researchers evaluating dual immersion instruction have found some evidence of potential tampering and differential attrition by ethnicity in Portland (Steele, Slater, Miller, Zamarro, & Li, 2014).

were initially assigned. Finally, although another study in the district found evidence that families whose child did not receive their first choice school are less likely to enroll than those who did receive their first choice, this differential attrition pattern is largely driven by white (non-EL) students, and so has little effect on the students in our sample (Kasman, 2014). English Learner students enroll in the district a high rate, regardless of whether there are assigned to their first choice school and program. These patterns suggest that manipulation of the assignment process and differential enrollment/attrition patterns likely have little impact on our estimates.

[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 classify students according to the initial EL instructional program they attended, and interpret our findings as the effect of one's initial EL program. Nonetheless, the majority of our sample attend the same program for at least three (99.5%) or four (95.2%) years, from kindergarten through third grade, indicating that there is little movement in and out of programs once ELs enroll in a particular program during their kindergarten year. A student's initial program is, in most cases, the program he or she attends for at least four years. 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 program. For instance, TB programs are designed to reclassify students as fluent English proficient and transition them into EI programs more quickly than the DB and DI programs. The proportion of ELs who were initially enrolled in TB and are still enrolled in TB drops by 32 percentage points (from 90% to 58%)

from grade three to four, compared to a 15 and 3 percentage point drop between these grades for DB and DI, respectively. This difference is simply an artifact of the program design rather than reflecting a lack of compliance. Across programs, by middle school students are generally transitioned into EI programs.

*Sample Descriptives*. As can be seen in Table 2, 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. 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 each of the two-language instructional programs have lower initial English proficiency in the fall of kindergarten than those in English immersion. This may in part be because in kindergarten, the two-language programs spend much instructional time in the target language. Parents may choose these programs for their children partly because of their incoming level of proficiency. Further, in 2$^{nd}$ grade ELs in EI and TB score above their peers in DB and DI in both ELA and math. Those in DI score substantially below their peers in all of the other programs in both subjects in 2$^{nd}$ grade. This pattern remains in middle school grades, but is

slightly less pronounced. Also noteworthy, relative to the state average in those grades, the average ELA and math scores of those in all programs increase from 2nd through 6th/7th grade.

[Insert Table 2 about here]

*Methods*

*Research question 1*. In order to answer the first research question regarding the differential effect of each instructional program on ELs' academic growth through middle school, we estimate four separate random coefficient student growth models (a special case of what are sometimes called mixed models, multilevel models, or hierarchical linear models): the first without student controls, the second with added student controls, the third with added student controls and school fixed effects, and the fourth including student controls, school fixed effects, and fixed effects for parent preferences. While Model 3 adjusts for a set of observable student and school characteristics that are undoubtedly related to students' academic growth trajectories and students' choice of programs to attend, alone they may not fully account for student selection into programs. The fourth models—those with pre-treatment controls for parental preferences of the type and location of the EL instructional program—are our preferred models for identifying the effect of programs on students' outcomes. These models identify the effects of the instructional models by comparing students whose parents requested the same school-by-program combination but who were assigned to different programs by the algorithm. Allowing $p$ to index 191 school-by-instructional program combinations[4], $i$ to index students, and $t$ to index grades, we fit random coefficients models of the form:

$$Y_{tip} = \gamma_{0ip} + \mathbf{X}_{ip}\mathbf{B}_0 + \mathbf{P}_i\mathbf{\Gamma}_0 + \gamma_{1ip}GRD_{tip} + \left[\mathbf{X}_{ip} \cdot GRD_{tip}\right]\mathbf{B}_1 + \left[\mathbf{P}_i \cdot GRD_{tip}\right]\mathbf{\Gamma}_1 + e_{tip}$$

[4] 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.

where both the intercept ($\gamma_{0ip}$) and the coefficient on grade ($\gamma_{1ip}$) vary randomly across school-by-instructional program combinations[5] and among individuals within these programs:

$$\gamma_{0ip} = \alpha_0 + u_{0p} + v_{0ip}$$

$$\gamma_{1ip} = \alpha_1 + u_{1p} + v_{1ip}$$

The random effects are assumed to be mean zero and multivariate normal among students and among programs. Likelihood ratio tests of the null hypotheses that the variance for each random effect is equal to zero indicated that, in all of our models, each of the random effects improves the model fit ($p < 0.001$ in all cases).

In the above model, $Y_{tip}$ represents the ELA or math score for student *i* in grade *t* in initial program *p*. The variable $GRD_{tip}$ indicates a student's grade, centered at grade 2, so that $\gamma_{0ip}$ and $\gamma_{1ip}$ indicate students' average test scores in grade two and average rates of change of their test scores from grades 2 to 8 (or to grade 6, in the case of math), respectively. The intercepts can be thought of as estimates of the cumulative effects of the programs through second grade (since students enroll in the EL programs at the start of kindergarten), and the slopes can be thought of estimates of the effects of the programs on the rate of learning in grades two through six or eight. These estimates, however, will be subject to selection bias if the fall kindergarten control variables included in the models are not sufficient. In all models, the slopes are constrained to be linear. We tested other less parametric model specifications that included grade fixed effects, but found that these nonparametric models did not significantly improve models fit.

[5] Note, these program random effects represent which of the 191 instructional programs ELs were enrolled in. This is not to be confused with the program preference fixed effects captured in $X_{ip}$, which uses dummy variables to indicate which of the 191 programs students listed as their first preference prior to enrollment, but not necessarily where students enrolled.

In this model, $\mathbf{X}_{ip}$ is a vector of student characteristics and parental program preferences. All of these variables are centered around their sample mean so that the intercepts and grade slopes apply to the average student in the sample. $\mathbf{P}_i$ represents a vector of dummy variables indicating the student's initial program type (TB, DB, or DI, with EI the omitted category). The coefficients of interest 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. To the extent that the models contain sufficient control variables to eliminate selection bias, $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ can be interpreted as the effects of the TB, DB, and DI instructional programs on ELs' test scores by second grade and their rates of growth following second grade.

We fit several versions of this model. Model 1 does not include any student-level covariates (no vector $\mathbf{X}_{ip}$) to provide a baseline descriptive model. Model 2 includes a vector of stable student/family control variables, $\mathbf{X}_{ip}$, which includes the students' gender, ethnicity, special education status, and initial English proficiency score. Due to the Family Education Rights and Privacy Act (FERPA) we were unable obtain data on students' free and reduced price lunch (FRPL) status from the school district. However, the district did provide aggregate percents of ELs who are eligible for free or reduced price lunch in each instructional program-by-ethnicity cell. These figures are presented at the bottom of Table 2. The percents do not vary sizably across programs or ethnicities, with one exception, Chinese DI; but the number of Chinese ELs in DI is quite low relative to the other programs. The limited variation suggests that controlling for this variable is unlikely to change our results net of the other variables we already control for (such as race and initial English proficiency)[6].

[6] Although we were not authorized to provide FRPL status-adjusted coefficients, we also note that an analyst at the school district internally ran our models controlling for FRPL to validate our findings, and confirmed that the pathway coefficients did not change more than 0.001 SD after making the adjustment.

In Model 3 we add initial school of attendance fixed effects to Model 2. This allows us to adjust for any school-specific factors that might account for observed differences across programs. The program coefficients in this model are identified off of within-school variation in program enrollment. Finally, in Model 4 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 preference fixed effects to our existing vector of student/family controls, $\mathbf{X}_{ip}$ to obtain within-program preference estimates. 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 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.

Because school-choice data are only available for students who entered the district in kindergarten starting in 2004, we only analyze academic outcomes through 7$^{th}$ 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 those students for whom we have preferences data, the sample in Model 4 is roughly half the size of the sample in Models 1 and 2. To ensure that any differences between Models 3 and 4 are not due to the difference in samples, we also fit Model 3 using the smaller sample used in Model 4. These are presented as "Model 3: Restricted Sample" in our results tables.

*Research question 2.* In order to test whether program effects vary by ethnicity, we fit the same models for Latino and Chinese students separately. All control variables in each of these models are centered around their ethnicity-specific sample means. To test whether program

effects vary by initial English proficiency, we add a set of two-way interactions between program type dummies and standardized initial English language proficiency score, and a set of three-way interactions between program dummies, initial proficiency, and grade.[7] A full set of model estimates are available in the online appendix.

*Interpretation of coefficients.* Recall that the coefficients of interest in our models are the vectors $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$, which represent the differences among instructional programs in ELs' test scores by second grade and their rates of growth following second grade. If program assignment were ignorably assigned, conditional on the covariates in the models, these can be interpreted as describing the effects of the programs on test score trajectories by second grade ($\mathbf{\Gamma}_0$) and from second grade through 6th or 8th grade ($\mathbf{\Gamma}_1$). Although we cannot be sure that assignment is ignorable, the fullest model specifications include a number of control variables, including (1) students' initial English proficiency scores, which are strong predictors of later academic scores; (2) school fixed effects, which control for any school-specific factors correlated with selection processes and students' potential outcomes; and (3) parental first-choice preference fixed effects, which will capture differences among students in factors related to parents' desire for their students to be in different programs.

Although these control variables might account for much of the selection bias one might worry about, they may not fully capture any differences among programs in EL students' initial academic skill. Despite that fact that we control for initial English ability, which is associated with later academic performance, we may not fully capture important variation between programs in academic skill; some ELs may be low in English language proficiency but otherwise perform high academically in their home language. If this initial academic skill were correlated

[7] We also fit models including the initial English proficiency interactions in the same models used to test for differential effects by ethnicity; these yielded the same conclusions as the models shown here.

with program enrollment, net of the other variables in our models, our estimates may be biased. Although pre-kindergarten or kindergarten measures of ELA and math skill are not available (because state tests are first administered in 2nd grade, not kindergarten), the school district did administer a general early childhood developmental inventory (ECDI) in the fall of kindergarten in the last three years. We cannot include this variable as a control in our models due to the limited years of availability, but Table 1a of the online appendix shows that average ECDI scores in the fall of kindergarten do not differ significantly among the EL programs, and that the inclusion of ECDI as a control variable does not significantly change the 2nd grade ELA and math coefficients for this sample after adjusting for our existing set of controls. This analysis suggests that our results do not suffer from omitted variable bias due to the omission of an unobserved measure of pre-kindergarten academic skill.

## Results

### *Differences in academic trajectories among EL instructional programs*

Results for our first research question, regarding the differential effect of each instructional program on ELs' academic growth through middle school, are presented in Table 3. The table includes estimates from the five models described above (Models 1-4, plus a second version of Model 3 based on the Model 4 sample). For each model, we tested the null hypotheses that the program-specific intercepts are equal and that the program-specific grade slopes are equal; *p*-values for these joint tests are at the bottom of Table 3. In general, the coefficient estimates are relatively similar across the specifications. For the sake of parsimony, and because it includes the most extensive set of control variables, we focus primarily Model 4 in our discussion of the results below.

[Table 3 about here]

*ELA*. The estimated intercepts indicate the differences in average ELA scores in second grade among the programs. By second grade, students in EI classrooms have average ELA scores that are not statistically distinguishable from the performance of the average student in the state. Relative to students in EI classrooms, and net of the covariates and fixed effects in the model, students in TB score significantly higher (by 0.08 SD) on the ELA test in second grade, while those in DB score no different, and those in DI score significantly lower (about 0.19 SD lower).

The estimated differences between programs in rates of growth in ELA scores from second through 7$^{th}$ grade show a somewhat different pattern. In general, the test scores of ELs in EI increase at a rate that is significantly slower than the rate of the average student in the state (recall that, because test scores are standardized relative to the state distribution in each grade, the average student in the state has a growth rate of exactly 0). Further, the rate at which the ELA test scores of ELs in TB increase is significantly faster than those of EI, while the rate for DB is not significantly distinguishable from those of students in EI, conditional on the covariates in the model. Finally, although ELs in DI classrooms have ELA scores well below those of their peers in EI classrooms in second grade, from second through seventh grade the ELA test scores of ELs in DI increase at a rate that is 0.064 standard deviations faster per grade than those in EI. This rate is sufficiently faster than EI students that by sixth grade the average ELA scores of DI-enrolled students match the state average, and surpass those of observationally similar ELs in EI and DB (see Figure 1). 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.

[Insert Figure 1 about here]

One thing to note in Table 3 is that the estimates are generally consistent in the models with and without controls for parental program preferences (i.e., in Model 3: Restricted and Model 4). This suggests that differences in parental preferences are not highly confounded with EL's potential academic trajectories. Although it is possible that there are still other factors that we did not observe that affect selection into programs and that are correlated with academic trajectories, this pattern of results, in conjunction with the ECDI results presenting in Appendix A, suggests that the coefficients might be interpreted as largely unbiased estimates of the effects of the different EL instructional programs in this district.

*Math*. In math, Models 3 and 4 likewise yield similar results to one another. By second grade, the math scores of EL students enrolled in EI classrooms are significantly higher than the state average (about 0.15 SD), while the scores of observationally similar ELs in TB and DB classrooms are even higher (by about 0.21 and 0.12 SD, respectively). The scores of those in 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 slope estimates in Table 3 indicate that the math test scores of students receiving EI instruction grow significantly more slowly than the state average. The math scores of EL students in DB classrooms grow significantly more slowly than those in EI, by about 0.04 standard deviations per grade; the growth rates of the scores of those in TB and DI programs are not statistically distinguishable from those of similar students in EI classrooms (See Figure1).

*Differences in program effects by ethnicity and initial English proficiency*

Estimates from models designed to determine whether program effects vary by EL students' ethnicity or initial level of English proficiency are presented in Table 4. Here we report results from only Model 4 (estimates from the other models are available in the online appendix

B). Table 4 clearly shows that the academic trajectories differ sharply between Latino and Chinese ELs; among those enrolled in EI, for example, the typical Latino EL has ELA and math scores about 0.8 to 1.0 SD, respectively, below those of the typical Chinese EL student in 2nd grade. This large achievement gap is evident in Figures 2 and 3.

[Insert Table 4 about here]

In addition to these large between-group differences in average scores, the effects of all three bilingual programs, relative to EI, appear to also vary between the two groups. For Latino ELs, the 2nd grade ELA scores of those in DB and DI are significantly lower than those attending EI, while the scores of those attending TB are not significantly different from the scores of those attending EI. However, the estimated growth rates in Table 4 indicate that although Latino ELs in all three of the bilingual programs score significantly lower than (or at best no different than) those in EI in 2nd grade, their rates of growth in ELA are significantly faster than the rate of growth of their Latino peers in EI. As can be seen in the left panel of Figure 2, this means that although in 2nd grade Latino ELs in two-language instructional programs score below or the same as their Latino peers in EI, by 7th grade, Latino ELs in all of these programs are scoring above those in EI on average (see Figure 2). The growth rate for Latino students in DI classrooms is roughly twice the growth rate of Latino ELs in the TB and DB programs.

[Insert Figure 2 about here]

The pattern of differences among programs in ELA trajectories for Chinese ELs is very different. In 2nd grade, Chinese ELs in TB have scores that are significantly higher than Chinese ELs in EI, and the ELA scores of those in DB and DI are not significantly different from the scores of those in EI. However, the average growth rates of ELA scores of Chinese ELs in TB and DI classrooms do not significantly differ from that of observationally similar students in EI

classrooms, and the average growth rate for Chinese ELs in DB classrooms is significantly slower than that of their Chinese EL peers in EI. This indicates that in general, the ELA score trajectories of Chinese ELs are most positive for those in DI, followed by EI. Best seen in Figure 2, it is noteworthy that regardless of program, the test scores of Chinese ELs are almost always above the ELA scores of the average student in the state.

In math, the coefficients on the grade-by-program interaction variables (but not the program intercept differences) for Latino ELs are somewhat similar to those in the ELA models. Latino ELs in EI score significantly below the state average in math in 2nd grade and the rate at which their math scores grow over time is significantly slower than the average rate of math score growth in the state. The 2nd grade scores of Latino ELs in TB and DI are not significantly different from those in EI, while the 2nd grade math scores of those in DB are significantly higher than those in EI. However, the rate of test score growth of Latino ELs in DI are significantly faster than the rate at which the math scores of those in EI increase. The slopes for Latino ELs in TB and DB do not differ from (or at best are marginally significantly better than) the average slope of their Latino peers in EI. By sixth grade, Latino ELs in each of the two-language programs have higher average math scores than their observationally similar peers in EI classrooms, a pattern similar to the patterns in ELA scores (see Figure 3).

[Insert Figure 3 about here]

Chinese ELs show almost exactly the same pattern of results in math as they do in ELA, with the exception of one finding, that the 2nd grade math test scores of Chinese ELs in DB are significantly higher than the 2nd grade math scores of their Chinese peers in EI. The by-ethnicity results suggest that Latino ELs perform the best in both ELA and math in the long term when they are enrolled in any of the bilingual programs, but especially have the most optimal long-

term outcomes in DI. While Chinese ELs also do best longitudinally in ELA and math when enrolled in DI, they also do very well in EI – the program that uses no home-language instruction. They perform worst longitudinally in DB in both subjects, but especially in math.

For both Latino and Chinese ELs and both math and ELA, separate significance tests of the null hypotheses that program grade two intercepts are jointly equal to zero and that program slopes are jointly equal to zero can be found at the bottom of Table 4. All tests indicate significant between-program differences in intercepts and slopes. Finally, we note that tests of whether the patterns of program effects differ between Chinese and Latino students (estimated by fitting a fully interacted model on the full sample of Latino and Chinese students; results not shown) indicate that program-specific intercepts and rates of test score growth among Chinese ELs differ significantly from those of Latino ELs.

The right panel of Table 4 reports the estimated differences in program effects by ELs' initial level of English proficiency (EP). Note that in this table, results from a single model are presented across two columns (main effects and by initial EP side-by-side). There is little evidence of significant differences in program effects by initial proficiency, as evidenced by the large *p*-values (at the bottom of Table 4) from the tests of the null hypotheses that the grade two program effects are equal and that the program effects on growth rates are equal.

## Discussion

In this paper, we estimate the associations among 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 arguably provide more plausible estimates of program effects than much of the existing literature, as we are able to eliminate two key potential sources of selection bias: the confounding of program enrollment with parental preferences (a common unobservable characteristic in other similar studies) and the confounding of program enrollment with differences in academic preparation prior to kindergarten.

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 instructional programs in kindergarten. By second grade ELs in dual immersion classrooms 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 well above those of ELs in English immersion in both ELA and math, and those in developmental bilingual have math test scores that are significantly higher than their peers in English immersion.

Second, the effects of EL instructional programs on longer-term academic trajectories (into middle school) differ from the apparent effects on shorter-term academic outcomes. For example, in the short term (through second grade), ELs in dual immersion score substantially below their EL counterparts attending other instructional programs in ELA. By seventh grade, however, students in dual immersion and transitional bilingual programs have much higher ELA scores than those in English immersion classrooms. This pattern of a reversal in the relative effects of EL programs is consistent with other research that, for Latino ELs, both the development of English proficiency and reclassification patterns are slower in early elementary school for those in bilingual EL programs than for those in English immersion programs, but that

ELs in two-language programs catch up or surpass their English immersion-enrolled peers by middle school (Umansky & Reardon, 2014).

In some ways these patterns are not particularly surprising; indeed, they are likely at least partly an artifact of the programs' designs. 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). As a result, they develop English proficiency more slowly in the early grades. This may partly explain their lower ELA performance. Moreover, because the ELA and math tests are administered solely in English, students in dual immersion classrooms may not be able to fully demonstrate their knowledge, particularly in math. Thus, although ELs in dual immersion score poorly on tests administered in English in the early grades, this is not necessarily an indicator that they are not developing important content knowledge and literacy skills that in the long term will ultimately transfer to English language and other academic development.

Further, the test score growth rates of ELs in dual immersion far out-pace those of ELs in the other programs. It is possible that dual immersion programs have this effect because they combine 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, 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 understanding, 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 (Genesee, 2008; Cummins, 1979; 2000; Goldenberg, 1996). Given this argument, however, one might be surprised that the ELA test scores of ELs in developmental bilingual increase more slowly than those in English immersion, as this seems inconsistent with theory and research on transfer across languages. However, as is evident in Table 4, this negative effect is driven by the effects among Chinese ELs, which we discuss in further detail below.

One implication of the comparison of short- and long-term effects is that EL programs should be evaluated using both short- and long-term outcomes. Measuring EL programs' "effectiveness" by looking at only short-term outcomes might lead one to conclude that dual immersion programs are the least effective of the four models, and that programs that emphasize more English instruction earlier (transitional bilingual and English immersion) are superior. An examination of longer-term findings yields a different conclusion, however, which highlights the need to include longer-term outcomes in evaluations of EL programs.

A third notable finding is that the effects of the different EL instructional programs appear to differ for Latino and Chinese ELs. For instance, we generally find that, compared to Latino ELs in English immersion classrooms, Latino ELs in bilingual programs initially score lower on ELA tests in 2$^{nd}$ grade and improve their ELA scores faster following second grade. The reverse pattern was observed for Chinese ELs in transitional and developmental bilingual programs, though not for those in dual immersion programs. Indeed the one commonality

between the Latino and Chinese patterns is that for both groups, in both math and ELA, EL students in dual immersion programs had the fastest growth rates from 2nd grade into middle school (though in the case of the Chinese ELs, growth rates in dual immersion classrooms were not significantly faster than those of children in English immersion).

The significant negative effects of both types of bilingual instruction relative to English immersion instruction on Chinese ELs' test score growth has two 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, the outcomes of interest are measured by tests administered in English). This might explain why Chinese students in dual immersion classrooms have ELA and math trajectories that are not statistically distinguishable from those in English immersion, given that dual immersion classrooms include native English speakers and some instructional time 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 apparent positive early effects of transitional bilingual education for Chinese students. Moreover, 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 potential explanation is that the Chinese and Spanish language bilingual programs are implemented differently in this district. We were not able to directly observe EL classrooms as part of this study, but it may be that finding well-qualified teachers for Chinese bilingual programs is harder than for Spanish language programs (a difficulty some district officials have described to us); as a result the Chinese programs may not be implemented with the same fidelity as the Spanish programs, leading to different patterns of effects.

A fourth notable result is that the effects of the EL instructional models appears to be similar for ELs at all levels of initial English proficiency. This is in contrast to Jepsen's (2010) findings that bilingual instruction had a positive effect on English proficiency among ELs with high prior proficiency, and negative effects among those with low prior proficiency. However, Jepsen considered differences in program effectiveness for English proficiency rather than academic outcomes which could be one explanation for the divergence in results. Further, his measure of prior English proficiency was defined as proficiency in the year prior, while our measure considered initial English proficiency.

*Concluding remarks & Study Limitations*

Although this study provides some suggestive evidence about the effects of different EL instructional program models, it has a number of limitations. First, our estimated program effects are not based on a randomized experiment to draw full causal conclusions. Our estimates are interpretable as "effects" of the programs only to the extent that the models include sufficient control variables to render program enrollment ignorably assigned. We are able to include not only a standard set of demographic controls, but controls for initial English proficiency, school

fixed effects, and a rich set of parental preference control variables. In addition, our supplemental analyses based on the subsample of students with ECDI scores suggests that our main estimates are not biased by the exclusion of measures of pre-kindergarten academic skills. These features of the analysis suggest that we might think of our estimates as largely, but not completely, unbiased. They provide a useful piece of evidence in what should surely be a more extensive and ongoing research agenda.

Second, the data we use come from a single school district, one which is somewhat unique in terms of its ethnic and linguistic diversity and its historical commitment to providing multiple different types of EL instructional models. It is not clear whether the patterns we observe here generalize to other settings, particularly given the heterogeneity of the EL population and of the design and delivery of two-language instructional models across the U.S. 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) weighted toward instruction in the EL students' home language (Collier & Thomas, 2004). Our study speaks to the effectiveness of four distinct and very specific program models that primarily serve Latino and Chinese EL students in one large school district.

Third, our interpretation of "program effectiveness" is limited to outcomes measured by tests administered in English. We cannot estimate the effects of the programs on other important outcomes that matter for EL students' development. For example, we find that the test scores of Chinese ELs in developmental bilingual programs grow at a rate that is statistically slower than that of their peers in English immersion classrooms. However, ELs enrolled in bilingual programs for six years or more may reap the added benefits of bilingualism and biliteracy, potentially important skills for both personal development and future labor market success

(Gándara & Rumberger, 2009). Because we have no measure of home language proficiency or literacy, we cannot estimate the programs' effects on these outcomes.

A fourth limitation of the study is that we are blind to differences among the programs in the quality of instruction and classroom environments. Our inclusion of school fixed effects in the models does adjust for differences in classroom and instructional quality across schools, but it does not eliminate any bias due to systematic differences within schools. To the extent that there are systematic differences in classroom quality across programs within schools, or to the extent to which teacher qualifications and skills differ among the programs, we may be capturing differences in teaching quality rather than what the differences in the effectiveness of the four instructional models would be if each were well-implemented and staffed.

In sum, the results here suggest, in broad strokes, that there are meaningful differences in the effects of different models of EL instruction. These effects are not simple to characterize, as they vary as children progress through school; they differ for Latino and Chinese EL students; and they differ somewhat between math and reading outcomes. In particular, the findings here suggest that, for Latino students in particular, two-language programs lead to better academic outcomes than English immersion programs in the long-term. Nonetheless, we do not think these findings, by themselves, should lead all districts to exclusively adopt two-language programs. Our estimates are not sharply enough identified; the sample is not generalizable enough; and the mechanisms driving these patterns are not clear enough to warrant strong policy recommendations. Instead, we hope they contribute to a robust, empirically-grounded discussion about how best to educate our EL students.

*References*


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.

Census Bureau (2012). Fact Finder Table S161: Language Spoken at Home: 2008-2012 American Community Survey 5-Year Estimates. Available at: http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_S1601

Christian, D. (1998). Summary: Two-way immersion programs In C. Klee, A. Lynch, & E. Tarone (Eds). *Research and practice in immersion education: Looking back and looking ahead.* (39-44). University of Minnesota: Center for Advanced Research on Language Acquisition.

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

Conger, D. (2010). Does bilingual education interfere with English-language acquisition? Social Science Quarterly, 91, 1103-1122.

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.

District Program Guide (2014). English Learner Program Guide: 2014-2015 School Year. Multilingual Department.

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 (2009). *Immigration, Language, and Education: How Does Language Policy Structure Opportunity?* Teachers College Record.

Genesee, F., Geva, E., Dressler, C., Kamil, M.L. (2008). 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*. (61-93). 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.
Jepsen, C. (2010). Bilingual education and proficiency. *Education Finance and Policy, 5*, 200-227.
Kasman, M. (2014). *School selection, student assignment, and enrollment in a school district with open enrollment and mandatory choice policies*. (Unpublished doctoral dissertation). Stanford University, Stanford, CA.
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.
Lindholm-Leary, K., & Borsato, G. Academic achievement. In F. Genesee, K. Lindholm-Leary, & D. Christian. *Educating English language learners*. (176-211). New York, NY: Cambridge University Press.
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.
Matsudaira, J.D. (2005). Sinking or Swimming: Evaluating the Impact of English Immersion versus Bilingual Education. *Unpublished Manuscript.*
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, 2011).The condition of education 2011: Section 1 Participation in Education. .Available at: http://nces.ed.gov/pubs2011/2011033_2.pdf
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.
Riches, C., & Genesee, F. (2006). Literacy: Crosslinguistic and crossmodal issues. In F. Genesee, K. Lindholm-Leary, & D. Christian. *Educating English language learners*. (64-175). New York, NY: Cambridge University Press.
Rossell, C.H. (1990). The effectiveness of educational alternatives for limited-English-proficiency children. In G. Imhoff (Ed.), Learning in two languages. New Brunswick, NJ: Transaction Publishers.
Rossell, C.H., & Baker, K. (1996). The effectiveness of bilingual education. *Research in the Teaching of English, 30*, 7-74.
Saunders, W.M., & O'Brien, G. (2006). Oral language. In F. Genesee, K. Lindholm-Leary, & D. Christian. *Educating English language learners*. (14-63). New York, NY: Cambridge University Press.
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. (2011). Reading and language outcomes of a multi-year randomized evaluation of transitional bilingual education. *Educational Evaluation and Policy Analysis, 33*, 47-58.

Steele, J.L., Slater, R.O., Miller, T., Zamarro, G., & Li, J. (2014, April 6). *The effect of dual-language immersion on student performance in the Portland public schools: Evidence from the first study year*. Paper presented at the American Educational Research Association Annual meeting, Philadelphia, PA.

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.usc.edu/dept/education/CMMR/CollierThomasComplete.pdf

Umansky, I.M., & Reardon, S.F. (2014). Reclassification patterns among latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal*, 1-34.

Umansky, I.M. (2014, August). *Tracking by English proficiency: English learner students' access to core content in middle and high school*. Paper presented at the annual American Sociological Association conference, San Francisco, CA.

Valdés, G. (1998). Dual-language immersion programs: A cautionary note concerning the education of language minority students. In C. Klee, A. Lynch, & E. Tarone (Eds). *Research and practice in immersion education: Looking back and looking ahead*. (15-38). University of Minnesota: Center for Advanced Research on Language Acquisition.

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 EL 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 EL 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 | EL students served in classrooms with only English instruction | 100% EL or language minority. Students typically begin to transition out by 3rd grade, even if not yet reclassified as English proficient. | 100% EL 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. ELs 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. |

Source: District Program Guide (2014)

Table 2. Proportions of ELs of each ethnicity and of total ELs initially attending each program, average pre-treatment variables, by program; and proportion of ELs with each initial preference, by program

| | English Immersion | Transitional Bilingual | Developmental Bilingual | Dual Immersion | All Programs |
|---|---|---|---|---|---|
| | Proportion of ELs of each ethnicity initially in each program (Column proportions sum to one) | | | | Proportion of total ELs of each Ethnicity |
| 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 |
| All Ethnicities | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Proportion of total ELs, by program | 0.187 | 0.165 | 0.081 | 0.567 | 1.000 |
| | Additional pre-treatment covariates, by program | | | | |
| Average Initial English Proficiency | 0.129 | -0.203 | -0.147 | -0.136 | 0.000 |
| Female | 0.475 | 0.508 | 0.506 | 0.509 | 0.489 |
| Ever classified as Special Ed | 0.117 | 0.109 | 0.139 | 0.152 | 0.122 |
| N (ELs, full sample) | 7,793 | 2,573 | 2,271 | 1,113 | 13,750 |
| | Average ELA and math test scores, by program | | | | |
| 2nd grade ELA | 0.136 | 0.168 | -0.153 | -0.456 | 0.047 |
| 7th grade ELA | 0.279 | 0.205 | 0.059 | -0.133 | 0.200 |
| 2nd grade math | 0.263 | 0.408 | 0.073 | -0.234 | 0.219 |
| 6th grade math | 0.346 | 0.276 | 0.073 | -0.157 | 0.252 |
| | Proportion of ELs whose first choice is each program, by ethnicity (Row Proportions sum to one) | | | | Proportion of ELs with no preference |
| Latino | 0.290 | 0.124 | 0.113 | 0.227 | 0.243 |
| Chinese | 0.563 | 0.165 | 0.136 | 0.063 | 0.072 |
| Proportion of total ELs | 0.477 | 0.126 | 0.104 | 0.127 | 0.165 |
| N (ELs by program of attendance, preferences sample) | 4,469 | 1,411 | 1,046 | 803 | -- |
| | Proportion of ELs eligible for free/reduced-price lunch in each program, by ethnicity | | | | |
| Latino | 0.76 | 0.80 | 0.80 | 0.79 | 0.79 |
| Chinese | 0.72 | 0.85 | 0.88 | 0.51 | 0.76 |
| All ELs | 0.67 | 0.83 | 0.83 | 0.71 | 0.83 |

Note: Initial English proficiency is standardized around the sample average.

Table 3. Estimated parameters of average English language arts (ELA) and Math 2nd grade scores and growth trajectories, by initial (or predicted initial) program attended.

| | ELA | | | | | Math | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Model 1: Descriptive | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences | Model 1: Descriptive | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| | | | | | Intercepts | | | | | |
| Intercept (Average for English Immersion) | -0.018 | -0.02 | 0.031* | 0.135 | 0.018 | 0.095+ | 0.115*** | 0.160*** | 0.138*** | 0.151*** |
| | (0.049) | (0.028) | (0.016) | (0.017) | (0.015) | (0.056) | (0.030) | (0.018) | (0.017) | (0.016) |
| Transitional Bilingual (TB) | 0.034 | 0.133* | 0.061* | 0.071+ | 0.076* | 0.184+ | 0.270*** | 0.209*** | 0.219*** | 0.209*** |
| | (0.091) | (0.055) | (0.032) | (0.038) | (0.036) | (0.105) | (0.059) | (0.036) | (0.038) | (0.039) |
| Developmental Bilingual (DB) | -0.230* | -0.068 | -0.062+ | -0.038 | -0.017 | -0.148 | 0.047 | 0.091* | 0.145** | 0.116* |
| | (0.093) | (0.056) | (0.034) | (0.045) | (0.044) | (0.107) | (0.060) | (0.038) | (0.045) | (0.047) |
| Dual Immersion (DI) | -0.393*** | -0.144* | -0.098* | -0.122** | -0.191** | -0.258+ | 0.039 | 0.043 | 0.107+ | 0.032 |
| | (0.117) | (0.072) | (0.050) | (0.059) | (0.059) | (0.134) | (0.077) | (0.056) | (0.060) | (0.064) |
| | | | | | Slopes | | | | | |
| Grade (Average for English Immersion) | 0.003 | 0.002 | -0.004 | -0.023* | -0.023*** | -0.008 | -0.013 | -0.015 | -0.039*** | -0.039*** |
| | (0.006) | (0.006) | (0.004) | (0.005) | (0.004) | (0.010) | (0.010) | (0.007) | (0.007) | (0.006) |
| Transitional Bilingual (TB) X Grade | -0.016 | -0.017+ | 0.001 | 0.018 | 0.023* | -0.044* | -0.041* | -0.027* | -0.029+ | -0.020 |
| | (0.011) | (0.011) | (0.007) | (0.012) | (0.010) | (0.019) | (0.019) | (0.014) | (0.015) | (0.013) |
| Developmental Bilingual (DB) X Grade | 0.007 | 0.01 | 0.015+ | 0.004 | -0.022+ | -0.024 | -0.014 | -0.023 | -0.029 | -0.042* |
| | (0.011) | (0.011) | (0.008) | (0.014) | (0.013) | (0.019) | (0.019) | (0.014) | (0.018) | (0.016) |
| Dual Immersion (DI) X Grade | 0.051*** | 0.054*** | 0.043*** | 0.046** | 0.064*** | 0.008 | 0.013 | -0.011 | -0.018 | -0.010 |
| | (0.015) | (0.015) | (0.012) | (0.019) | (0.018) | (0.026) | (0.025) | (0.021) | (0.024) | (0.023) |
| Student random intercepts & slopes | X | X | X | X | X | X | X | X | X | X |
| L2 Stable Student Controls | | X | X | X | X | | X | X | X | X |
| L2 School Fixed Effects | | | X | X | X | | | X | X | X |
| L2 School-Program Preference Controls | | | | | X | | | | | X |
| L3 School * EL Instructional Program RE | X | X | X | X | X | X | X | X | X | X |
| Joint test of program intercepts (p-value) | 0.000 | 0.002 | 0.006 | 0.030 | 0.001 | 0.011 | 0.000 | 0.000 | 0.000 | 0.000 |
| Joint test of program slopes (p-value) | 0.001 | 0.000 | 0.004 | 0.043 | 0.000 | 0.083 | 0.112 | 0.150 | 0.134 | 0.049 |
| N (Observations - Level 1) | 65,912 | 65,912 | 65,912 | 28,428 | 28,428 | 55,499 | 55,499 | 55,499 | 27,386 | 27,386 |
| N (Students - Level 2) | 13,750 | 13,750 | 13,750 | 7,729 | 7,729 | 13,750 | 13,750 | 13,750 | 7,729 | 7,729 |
| N (School * EL Program - Level 3) | 191 | 191 | 191 | 150 | 150 | 191 | 191 | 191 | 150 | 150 |

Notes: Stable student controls include gender, ethnicity, special education status, 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 2-7 for Models 2 restricted and 3. School-program random effects represent the initial program (e.g. Dual Immersion program in school A) that students were enrolled in.

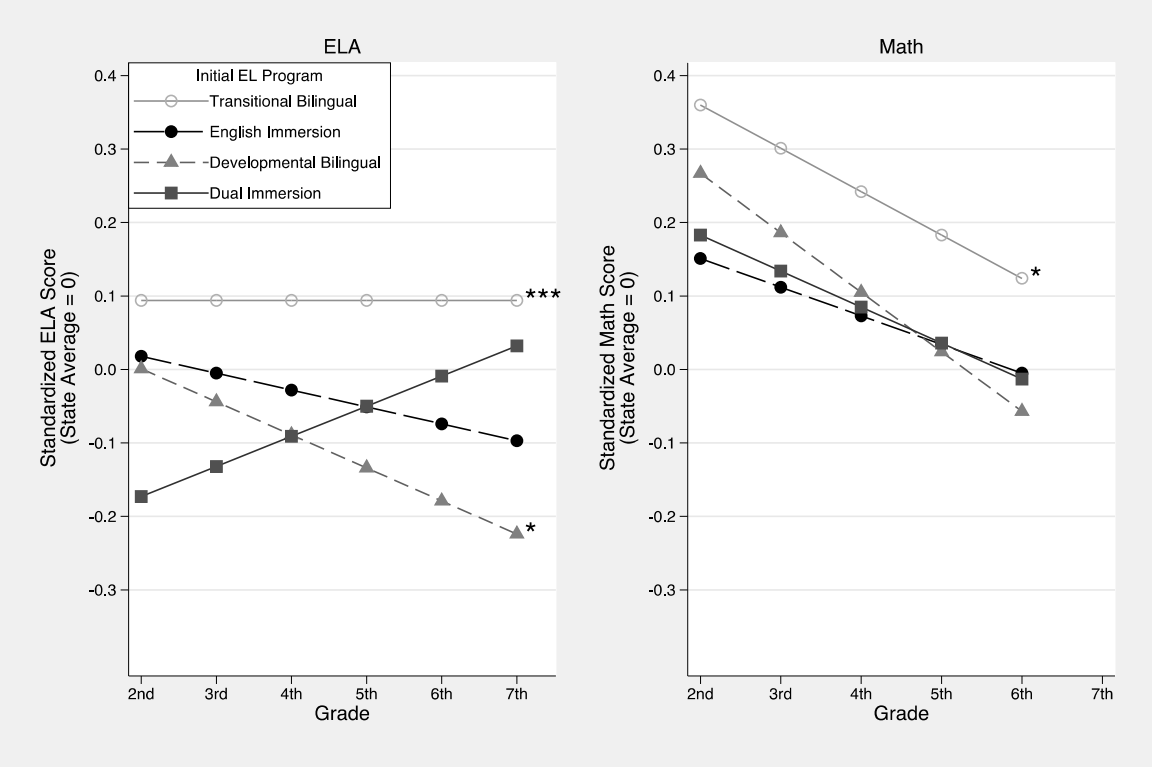
Table 4. Estimated parameters of average English language arts (ELA) and math growth trajectories, by initial program attended and ethnicity (left panel 1) and initial English proficiency (right panel 2)

| | (1) Separate models by ethnicity | | | | (2) Models by Initial English Proficiency (side by side columns are one model) | | | |
|---|---|---|---|---|---|---|---|---|
| | Latino | | Chinese | | Main Effects (e.g. TB x Grade) | Program X Initial English Proficiency Effects | Main Effects (e.g. TB x Grade) | Program X Initial English Proficiency Effects |
| | (1) ELA | (2) Math | (3) ELA | (4) Math | (1) ELA | | (2) Math | |
| | Intercepts | | | | | | | |
| Intercept (Average for English Immersion) (initial EP) | -0.451*** | -0.452*** | 0.354*** | 0.586*** | 0.020 | 0.191*** | 0.152*** | 0.160*** |
| | (0.026) | (0.027) | (0.032) | (0.032) | (0.015) | (0.015) | (0.016) | (0.015) |
| Transitional Bilingual (TB) (x initial EP) | -0.041 | 0.070 | 0.222** | 0.365*** | 0.072+ | 0.014 | 0.208*** | 0.010 |
| | (0.057) | (0.059) | (0.076) | (0.076) | (0.037) | (0.028) | (0.039) | (0.028) |
| Developmental Bilingual (DB) (x initial EP) | -0.156** | 0.130* | 0.072 | 0.227* | -0.012 | 0.006 | 0.119* | 0.004 |
| | (0.061) | (0.063) | (0.101) | (0.101) | (0.044) | (0.033) | (0.047) | (0.032) |
| Dual Immersion (DI) (x initial EP) | -0.366*** | -0.077 | -0.199 | 0.035 | -0.194**** | 0.053 | 0.032 | 0.027 |
| | (0.058) | (0.060) | (0.124) | (0.124) | (0.060) | (0.036) | (0.064) | (0.036) |
| | Slopes | | | | | | | |
| Grade (Average for English Immersion) (x initial EP) | -0.043*** | -0.081*** | -0.024** | -0.023* | -0.022*** | -0.013*** | -0.038*** | -0.014** |
| | (0.009) | (0.010) | (0.008) | (0.010) | (0.005) | (0.004) | (0.006) | (0.005) |
| Transitional Bilingual (TB) X Grade (x initial EP) | 0.065** | 0.042+ | -0.009 | -0.046+ | 0.021* | -0.001 | -0.021 | -0.002 |
| | (0.020) | (0.023) | (0.018) | (0.024) | (0.010) | (0.006) | (0.013) | (0.008) |
| Developmental Bilingual (DB) X Grade (x initial EP) | 0.056** | 0.026 | -0.068** | -0.148*** | -0.021 | 0.009 | -0.039* | 0.004 |
| | (0.021) | (0.023) | (0.026) | (0.035) | (0.013) | (0.007) | (0.016) | (0.009) |
| Dual Immersion (DI) X Grade (x initial EP) | 0.113*** | 0.061** | 0.049 | 0.005 | 0.061*** | 0.017* | -0.010 | 0.001 |
| | (0.021) | (0.022) | (0.032) | (0.044) | (0.018) | (0.009) | (0.023) | (0.011) |
| Student random intercepts & slopes | X | X | X | X | X | | X | |
| L2 Stable Student Controls | X | X | X | X | X | | X | |
| L2 School Fixed Effects | X | X | X | X | X | | X | |
| L2 School-Program Preference Controls | X | X | X | X | X | | X | |
| L3 School * EL Instructional Program RE | X | X | X | X | X | | X | |
| Joint test of program intercepts (p-value) (x initial EP) | 0.000 | 0.045 | 0.010 | 0.000 | >0.50 | | >0.50 | |
| Joint test of program slopes (p-value) (x initial EP) | 0.000 | 0.013 | 0.028 | 0.001 | 0.136 | | >0.50 | |
| N (observations - Level 1) | 9,335 | 9,043 | 12,918 | 12,382 | 28,428 | | 27,386 | |
| N (students - Level 2) | 2,646 | 2,646 | 3,362 | 3,362 | 7,729 | | 7,729 | |
| N (School * EL Program - Level 3) | 127 | 127 | 92 | 92 | 150 | | 150 | |

Notes: All coefficients estimated from Model 4 (controls, school and preference fixed effects). Stable student controls include gender, ethnicity, special education status, initial English proficiency score, and initial program preferences. All covariates, including the fixed effects, are group-mean centered within the sample used in each model (i.e. in the Latino models, initial English proficiency is centered around the mean initial English proficiency for Latinos, while in the Chinese models it is centered around the mean initial English proficiency for Chinese ELs). 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 2-7 and in math grades 2-6. School-program random effects represent the initial program (e.g. Dual Immersion program in school A) that students were enrolled in.

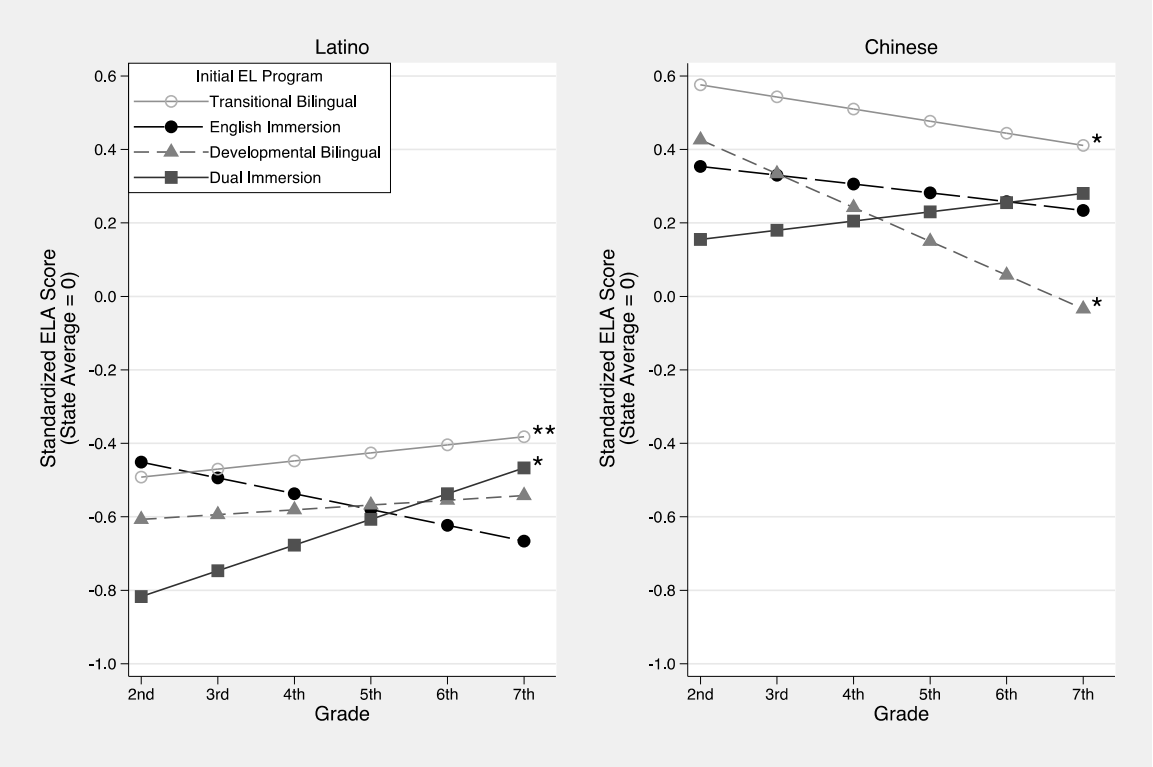
Figures

Figure 1. Estimated average ELA and math achievement trajectories, relative to state average: EL kindergarten entrants, by instructional program



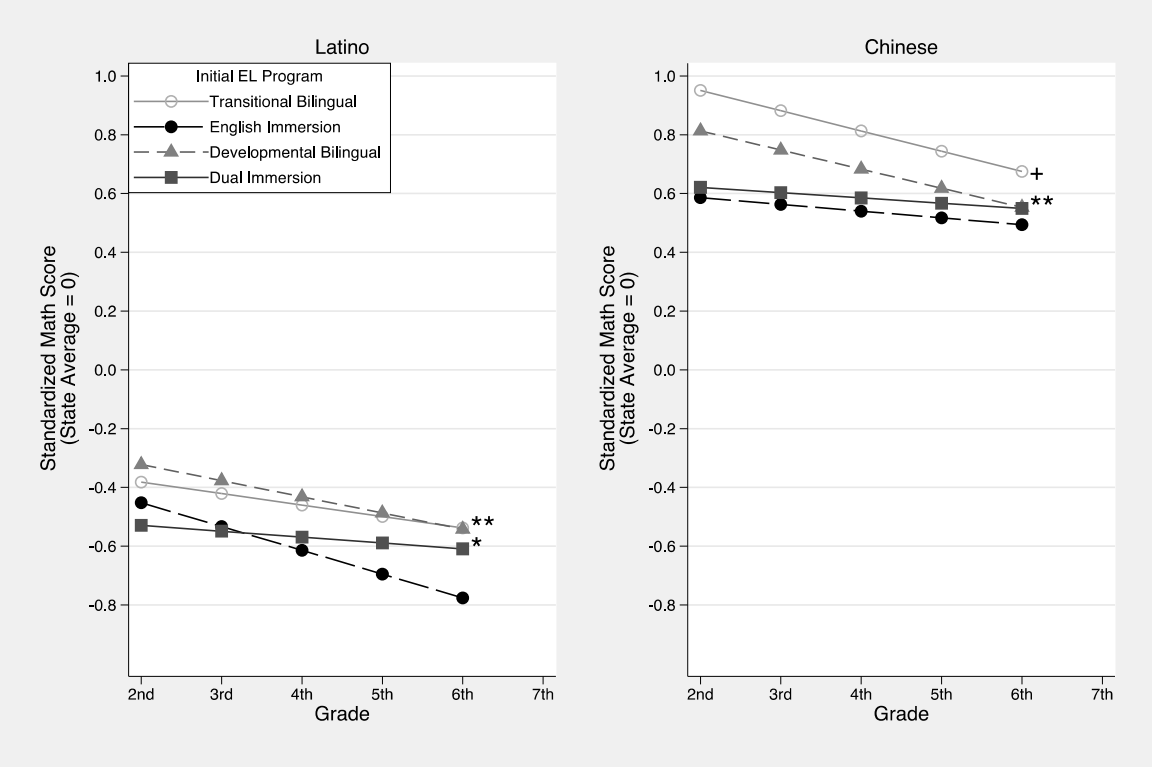

Note: Figure 1 is based on the estimates from Model 4 in Table 3. We did tests of significant differences across programs in the final grade of evaluation. Stars indicate that the average 6th (math) or 7th (ELA) grade outcome is significantly different from that among those whose initial EL program was English immersion.

Figure 2. Estimated average ELA achievement trajectory relative to the state average: EL kindergarten entrants, by instructional program and ethnicity




Note: Figure 2 is based on the estimates from the ELA models in the left panel of Table 4. We did tests of significant differences across programs in the final grade of evaluation. Stars indicate that the average 6th (math) or 7th (ELA) grade outcome is significantly different from that among those whose initial EL program was English immersion.

Figure 3. Estimated average math achievement trajectory relative to the state average: EL kindergarten entrants, by instructional program and ethnicity




Note: Figure 3 is based on the estimates from the math models in the left panel of Table 4. We did tests of significant differences across programs in the final grade of evaluation. Stars indicate that the average 6th (math) or 7th (ELA) grade outcome is significantly different from that among those whose initial EL program was English immersion.

## Online Appendix

### Appendix A: Early Childhood Development Inventory Analysis

Although our set of control variables might account for much of the selection bias one might worry about, they may not fully capture any differences among programs in EL students' initial academic skill. If this initial academic skill were correlated with program enrollment, net of the other variables in our models, our estimates may be biased. Although pre-kindergarten or kindergarten measures of ELA and math skill are not available (because state tests are first administered in 2nd grade, not kindergarten), the school district did administer a general early childhood developmental inventory (ECDI) in the fall of kindergarten in several recent years. This inventory is an assessment of phonological awareness, literacy, math and science, and—importantly—is administered in the child's home language if necessary. Because this measure was only administered in the last few years of the study, we have it only for the most recent cohorts in the study, meaning that we cannot use it as a control in the growth models. However, we are able to use it to assess whether its inclusion would alter our estimates.

In Table 1a we present two sets of results. The left panel shows estimates from models that predict initial ECDI scores as a function of initial EL program and other control variables. The estimates indicate that average ECDI scores in the fall of kindergarten do not differ significantly among the EL programs (as noted by the joint $F$-tests reported at the bottom of the table). On the right panel of Table 1a we present estimates of program differences in 2nd grade ELA and math scores from models that do not and do include the ECDI score as a covariate. The coefficients across the two models do not change significantly with the inclusion of the ECDI. Taken together, these results suggest that our results do not suffer from omitted variable bias due to the omission of an unobserved measure of pre-kindergarten academic skill.

Table 1a. Models predicting initial early childhood development inventory (ECDI) score in kindergarten.

| | Models predicting initial ECDI in Kindergarten* | | | | Models with and without initial ECDI predicting 2nd grade state test scores | | | |
|---|---|---|---|---|---|---|---|---|
| | Covariates | Covariates + School FE | Covariates + Pref FE | Covariates + School FE + Pref FE | ELA (without fall K ECDI) | ELA (with fall K ECDI) | Math (without fall K ECDI) | Math (with fall K ECDI) |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| Initial ECDI | -- | -- | -- | -- | -- | 0.018*** | -- | 0.020*** |
| | | | | | | (0.001) | | (0.001) |
| Transitional Bilingual | 1.247 | 0.822 | 2.666 | 1.607 | 0.103+ | 0.074 | 0.248** | 0.215** |
| | (1.594) | (1.048) | (2.091) | (1.449) | (0.059) | (0.050) | (0.085) | (0.080) |
| Developmental Bilingual | -0.131 | 0.869 | 0.406 | 1.742 | 0.067 | 0.035 | 0.108 | 0.072 |
| | (1.770) | (1.859) | (2.067) | (2.208) | (0.084) | (0.077) | (0.108) | (0.101) |
| Dual Immersion | -0.686 | 2.534+ | -1.818 | 0.793 | -0.097 | -0.111 | 0.054 | 0.038 |
| | (1.940) | (1.400) | (1.440) | (1.296) | (0.069) | (0.069) | (0.087) | (0.089) |
| Basic Controls | X | X | X | X | X | X | X | X |
| School FE | | X | | X | X | X | X | X |
| Preference FE | | | X | X | X | X | X | X |
| Joint F test of programs | 0.42 | 0.56 | 1.98 | 0.16 | 2.61 | 2.38 | 1.39 | 1.17 |
| p-value of joint F | 0.66 | 0.57 | 0.14 | 0.85 | 0.08 | 0.1 | 0.25 | 0.31 |
| N | 3,184 | 3,184 | 3,184 | 3,184 | 3,184 | 3,184 | 3,184 | 3,184 |

Note: Mean(sd) of initial ECDI is 66.83(19.67). Standard errors for all models are clustered at initial program (ie. School-program)

## Appendix B. Additional tables with complete set of models

Table 3a expanded to include all models. Estimated parameters of average English language arts (ELA) and math growth trajectories, by initial program attended and ethnicity.

| | **Latino** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ELA | | | | Math | | | |
| | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| | Intercepts | | | | | | | |
| Intercept (Average for English Immersion) | -0.426*** | -0.408*** | -0.468*** | -0.451*** | -0.408*** | -0.400*** | -0.465*** | -0.452*** |
| | (0.026) | (0.021) | (0.025) | (0.026) | (0.027) | (0.021) | (0.025) | (0.027) |
| Transitional Bilingual (TB) | -0.034 | -0.103* | -0.072 | -0.041 | 0.07 | 0.011 | 0.051 | 0.070 |
| | (0.053) | (0.044) | (0.053) | (0.057) | (0.055) | (0.045) | (0.054) | (0.059) |
| Developmental Bilingual (DB) | -0.233*** | -0.211*** | -0.111+ | -0.156** | -0.091+ | -0.028 | 0.159** | 0.130* |
| | (0.050) | (0.040) | (0.057) | (0.061) | (0.052) | (0.040) | (0.057) | (0.063) |
| Dual Immersion (DI) | -0.263*** | -0.266*** | -0.301*** | -0.366*** | -0.102 | -0.113* | -0.028 | -0.077 |
| | (0.060) | (0.045) | (0.052) | (0.058) | (0.062) | (0.044) | (0.052) | (0.060) |
| | Slopes | | | | | | | |
| Grade (Average for English Immersion) | -0.025*** | -0.034*** | -0.043*** | -0.043*** | -0.050*** | -0.064*** | -0.080*** | -0.081*** |
| | (0.007) | (0.007) | (0.010) | (0.009) | (0.011) | (0.009) | (0.012) | (0.010) |
| Transitional Bilingual (TB) X Grade | 0.028* | 0.056*** | 0.074*** | 0.065** | 0.012 | 0.037* | 0.058* | 0.042+ |
| | (0.014) | (0.014) | (0.021) | (0.020) | (0.022) | (0.018) | (0.025) | (0.023) |
| Developmental Bilingual (DB) X Grade | 0.053*** | 0.061*** | 0.050* | 0.056** | 0.038+ | 0.045** | 0.012 | 0.026 |
| | (0.013) | (0.012) | (0.023) | (0.021) | (0.020) | (0.016) | (0.027) | (0.023) |
| Dual Immersion (DI) X Grade | 0.092*** | 0.090*** | 0.112*** | 0.113*** | 0.063* | 0.068*** | 0.058* | 0.061** |
| | (0.016) | (0.015) | (0.022) | (0.021) | (0.026) | (0.020) | (0.026) | (0.022) |
| Joint test of program intercepts (p-value) | 0.000 | 0.000 | 0.000 | 0.000 | 0.031 | 0.079 | 0.026 | 0.045 |
| Joint test of program slopes (p-value) | 0.000 | 0.000 | 0.000 | 0.000 | 0.047 | 0.001 | 0.019 | 0.013 |
| N (observations - Level 1) | 20,304 | 20,304 | 9,335 | 9,335 | 17,555 | 17,555 | 9,043 | 9,043 |
| N (students - Level 2) | 4,554 | 4,554 | 2,646 | 2,646 | 4,554 | 4,554 | 2,646 | 2,646 |
| N (School * EL Program - Level 3) | 156 | 156 | 127 | 127 | 156 | 156 | 127 | 127 |

Table 3a continued.

| | Chinese | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ELA | | | | Math | | | |
| | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| | Intercepts | | | | | | | |
| Intercept (Average for English Immersion) | 0.299*** | 0.355*** | 0.354*** | 0.354*** | 0.569*** | 0.611*** | 0.585*** | 0.586*** |
| | (0.033) | (0.025) | (0.031) | (0.032) | (0.037) | (0.032) | (0.032) | (0.032) |
| Transitional Bilingual (TB) | 0.234*** | 0.137** | 0.229** | 0.222** | 0.366*** | 0.263*** | 0.383*** | 0.365*** |
| | (0.067) | (0.053) | (0.070) | (0.076) | (0.076) | (0.066) | (0.072) | (0.076) |
| Developmental Bilingual (DB) | 0.089 | 0.08 | 0.020 | 0.072 | 0.183* | 0.173* | 0.218* | 0.227* |
| | (0.080) | (0.066) | (0.098) | (0.101) | (0.090) | (0.083) | (0.101) | (0.101) |
| Dual Immersion (DI) | -0.088 | -0.119 | -0.057 | -0.199 | 0.106 | 0.161 | 0.124 | 0.035 |
| | (0.141) | (0.140) | (0.116) | (0.124) | (0.157) | (0.166) | (0.118) | (0.124) |
| | Slopes | | | | | | | |
| Grade (Average for English Immersion) | 0.015* | 0.003 | -0.020** | -0.024** | 0.000 | -0.004 | -0.028* | -0.023* |
| | (0.006) | (0.005) | (0.008) | (0.008) | (0.011) | (0.009) | (0.012) | (0.010) |
| Transitional Bilingual (TB) X Grade | -0.052*** | -0.026** | -0.022 | -0.009 | -0.071** | -0.051** | -0.055* | -0.046+ |
| | (0.012) | (0.010) | (0.017) | (0.018) | (0.023) | (0.019) | (0.026) | (0.024) |
| Developmental Bilingual (DB) X Grade | -0.027+ | -0.015 | -0.061* | -0.068** | -0.056* | -0.069** | -0.139*** | -0.148*** |
| | (0.014) | (0.013) | (0.025) | (0.026) | (0.028) | (0.024) | (0.040) | (0.035) |
| Dual Immersion (DI) X Grade | 0.026 | 0.020 | 0.011 | 0.049 | -0.025 | -0.052 | 0.006 | 0.005 |
| | (0.026) | (0.028) | (0.030) | (0.032) | (0.050) | (0.050) | (0.045) | (0.044) |
| Student random intercepts & slopes | X | X | X | X | X | X | X | X |
| L2 Stable Student Controls | X | X | X | X | X | X | X | X |
| L2 School Fixed Effects | | X | X | X | | X | X | X |
| L2 School-Program Preference Controls | | | | X | | | | X |
| L3 School * EL Instructional Program RE | X | X | X | X | X | X | X | X |
| Joint test of program intercepts (p-value) | 0.004 | 0.51 | 0.011 | 0.010 | 0.000 | 0.001 | 0.000 | 0.000 |
| Joint test of program slopes (p-value) | 0.000 | 0.045 | 0.077 | 0.028 | 0.010 | 0.003 | 0.002 | 0.001 |
| N (observations - Level 1) | 31,858 | 31,858 | 12,918 | 12,918 | 26,254 | 26,254 | 12,382 | 12,382 |
| N (students - Level 2) | 6,237 | 6,237 | 3,362 | 3,362 | 6,237 | 6,237 | 3,362 | 3,362 |
| N (School * EL Program - Level 3) | 122 | 122 | 92 | 92 | 122 | 122 | 92 | 92 |

Table 3b expanded to include all models. Estimated parameters of average English language arts (ELA) and math 2nd grade scores and growth trajectories, by initial program attended & initial English Proficiency (EP).

| | ELA | | | | Math | | | |
|---|---|---|---|---|---|---|---|---|
| | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| | Intercepts | | | | | | | |
| Intercept (Average for English Immersion) | -0.018 | 0.033* | 0.013 | 0.020 | 0.118*** | 0.159*** | 0.142*** | 0.152*** |
| | (0.028) | (0.016) | (0.017) | (0.015) | (0.030) | (0.018) | (0.017) | (0.016) |
| Transitional Bilingual (TB) | 0.127* | 0.052 | 0.069+ | 0.072+ | 0.268*** | 0.206*** | 0.218*** | 0.208*** |
| | (0.055) | (0.032) | (0.038) | (0.037) | (0.058) | (0.036) | (0.038) | (0.039) |
| Developmental Bilingual (DB) | -0.072 | -0.063+ | -0.039 | -0.012 | 0.043 | 0.086* | 0.147*** | 0.119* |
| | (0.056) | (0.035) | (0.045) | (0.044) | (0.060) | (0.039) | (0.045) | (0.047) |
| Dual Immersion (DI) | -0.137+ | -0.093+ | -0.138* | -0.194**** | 0.045 | 0.037 | 0.108+ | 0.032 |
| | (0.072) | (0.051) | (0.059) | (0.060) | (0.077) | (0.057) | (0.059) | (0.064) |
| Initial English Proficiency | 0.158*** | 0.163*** | 0.192*** | 0.191*** | 0.118*** | 0.123*** | 0.157*** | 0.160*** |
| | (0.012) | (0.012) | (0.016) | (0.015) | (0.013) | (0.013) | (0.016) | (0.015) |
| TB X Initial English Proficiency (EP) | -0.003 | 0.001 | 0.019 | 0.014 | 0.012 | 0.008 | 0.023 | 0.010 |
| | (0.021) | (0.021) | (0.029) | (0.028) | (0.024) | (0.024) | (0.029) | (0.028) |
| DB X Initial EP | 0.026 | 0.018 | 0.000 | 0.006 | 0.014 | -0.000 | -0.001 | 0.004 |
| | (0.023) | (0.024) | (0.034) | (0.033) | (0.026) | (0.026) | (0.033) | (0.032) |
| DI X Initial EP | 0.053+ | 0.042 | 0.070+ | 0.053 | 0.030 | 0.017 | 0.046 | 0.027 |
| | (0.030) | (0.030) | (0.037) | (0.036) | (0.034) | (0.034) | (0.037) | (0.036) |

Table 3b continued.

| | ELA | | | | Math | | | |
|---|---|---|---|---|---|---|---|---|
| | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences | Model 2: Student Controls | Model 3: Student Controls + School FE | Model 3: Restricted Sample | Model 4: Controls + School FE + Preferences |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| | Slopes | | | | | | | |
| Grade (Average for English Immersion) | 0.001 | -0.004 | -0.021*** | -0.022*** | -0.016 | -0.014* | -0.040*** | -0.038*** |
| | (0.006) | (0.005) | (0.006) | (0.005) | (0.010) | (0.007) | (0.007) | (0.006) |
| Transitional Bilingual (TB) X Grade | -0.015 | 0.002 | 0.014 | 0.021* | -0.038* | -0.024+ | -0.029+ | -0.021 |
| | (0.011) | (0.016) | (0.012) | (0.010) | (0.019) | (0.014) | (0.015) | (0.013) |
| Developmental Bilingual (DB) X Grade | 0.01 | 0.016+ | 0.005 | -0.021 | -0.012 | -0.018 | -0.028 | -0.039* |
| | (0.011) | (0.009) | (0.015) | (0.013) | (0.019) | (0.015) | (0.019) | (0.016) |
| Dual Immersion (DI) X Grade | 0.054*** | 0.041** | 0.043* | 0.061*** | 0.014 | -0.007 | -0.016 | -0.010 |
| | (0.015) | (0.014) | (0.019) | (0.018) | (0.026) | (0.022) | (0.024) | (0.023) |
| Grade X Initial EP | 0.002 | 0.002 | -0.014*** | -0.013*** | 0.011** | 0.011** | -0.013** | -0.014** |
| | (0.002) | (0.003) | (0.004) | (0.004) | (0.004) | (0.004) | (0.005) | (0.005) |
| TB X Grade X initial EP | -0.007+ | -0.007 | -0.002 | -0.001 | -0.015* | -0.015* | -0.005 | -0.002 |
| | (0.004) | (0.005) | (0.007) | (0.006) | (0.006) | (0.006) | (0.008) | (0.008) |
| DB X Grade X initial EP | -0.004 | -0.005 | 0.007 | 0.009 | -0.011 | -0.010 | 0.000 | 0.004 |
| | (0.005) | (0.005) | (0.007) | (0.007) | (0.007) | (0.007) | (0.010) | (0.009) |
| DI X Grade X initial EP | -0.004 | -0.005 | 0.018* | 0.017* | -0.018+ | -0.021* | -0.003 | 0.001 |
| | (0.007) | (0.007) | (0.009) | (0.009) | (0.010) | (0.010) | (0.011) | (0.011) |
| Student random intercepts & slopes | X | X | X | X | X | X | X | X |
| L2 Stable Student Controls | X | X | X | X | X | X | X | X |
| L2 School Fixed Effects | | X | X | X | | X | X | X |
| L2 School-Program Preference Controls | | | | X | | | | X |
| L3 School * EL Instructional Program RE | X | X | X | X | X | X | X | X |
| Joint test of program x EP intercepts (p-value) | 0.243 | >0.50 | 0.293 | >0.50 | >0.50 | >0.50 | >0.50 | >0.50 |
| Joint test of program x EP slopes (p-value) | >0.50 | >0.50 | 0.144 | 0.136 | 0.040 | 0.040 | >0.50 | >0.50 |
| 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 * EL Program - Level 3) | 191 | 191 | 150 | 150 | 191 | 191 | 150 | 150 |