Research Article

Jennifer L. Steele*, Johanna Watzinger-Tharp, Robert O. Slater, Gregg Roberts-Aguirre and Karl Bowman

Achievement Effects of Dual Language Immersion in One-Way and Two-Way Programs: Evidence from a Statewide Expansion

https://doi.org/10.1515/bejeap-2022-0241
Received July 2, 2022; accepted July 7, 2024; published online July 31, 2024

Abstract: The rising demand for dual-language immersion (DLI) programs, which offer core instruction in two languages from early grades onward, has raised questions about program design and access. We leverage the rapid expansion of DLI schools across the U.S. state of Utah to estimate effects of DLI program availability on the academic achievement of primary English speakers and English learners (ELs) in programs that serve mainly the former (one-way) or at least a third of the latter (two-way). Using within-school variation in first graders' access to DLI programs, we find no overall effects on English, math, or science scores from grades 3 to 6. However, ELs whose primary languages match the schools' partner languages in two-way schools show notable outperformance in math and higher English-language proficiency at grade 5. Benefits of DLI access are driven by schools with a larger share of primary speakers of the partner language.

Keywords: bilingual education; English learners; education policy; culturally relevant instruction; human capital acquisition

***Corresponding author: Jennifer L. Steele**, American University, School of Education, 4400 Massachusetts Ave., NW, Washington, DC, 20016-8030, USA, E-mail: steele@american.edu. https://orcid.org/0000-0002-5014-6345
Johanna Watzinger-Tharp, Linguistics, University of Utah, Salt Lake City, UT, USA, E-mail: j.tharp@utah.edu. https://orcid.org/0000-0001-6300-1754
Robert O. Slater, American Councils for International Education, Washington, DC, USA, E-mail: rslater1@gmail.com
Gregg Roberts-Aguirre, DLI Alliance, Salt Lake City, UT, USA, E-mail: gregg.roberts-aguirre@dli-alliance.org
Karl Bowman, Utah State Board of Education, Salt Lake City, UT, USA, E-mail: Karl.Bowman@schools.utah.gov

1 Introduction

Policy considerations around language education are often fraught. The United Nations recommends that young children in linguistically diverse societies have access to education in their primary language (UNESCO 2016a), but fulfillment of this recommendation has proven difficult both logistically and politically in many nations (UNESCO 2016b). In the United States, advocacy for "English only" policies in schools and other public domains, linked closely to anti-immigrant ideology (Padilla et al. 1991), yielded voter-initiated bans on bilingual education for English learners (ELs) in California, Arizona, and Massachusetts from the late 1990s through the mid-2010s (Lam and Richards 2020; Mora 2009), at which point they were overturned in California and Massachusetts (Commission on Language Learning 2017; Kamenetz 2016). In the past decade, however, the U.S. has seen a surge of public interest in bilingual and dual-language education programs as a means not only of supporting the roughly 5 million English learners (ELs) in U.S. public schools, but also of promoting multilingualism in the U.S. In particular, dual language immersion (DLI) – an instructional model that delivers core content instruction in two languages to primary English speakers and ELs from early grades onward – has gained prominence as the public has become aware of the cognitive and economic advantages of bilingualism (Fabián Romero 2017; Maxwell 2014).[1]

DLI programs offer general academic instruction in two languages beginning in early grades and often extending into middle or high school. They include both *two-way programs*, in which at least a third of classroom students are primary speakers of each of the two classroom languages (in the U.S., typically English and a non-English "partner" language), and *one-way programs*, in which most students in the classroom share a common primary language and are immersed in a non-primary partner language. Both types of programs are designed to move students toward bilingualism and biliteracy, regardless of their primary or home languages (Fortune 2012). But by design, two-way programs can facilitate rapid access to content and communication for ELs alongside their primary English-speaking counterparts. In contrast, one-way programs operate under the assumption that students share a common primary language (in our study, English) and are new to the partner

1 We use the term "primary English speaker" to refer to students who enter school proficient in English, even though some may be primary speakers of other languages as well. Throughout the paper, we use the term "primary language" to describe what parents report as the child's first or home language. We use the term "EL" to refer to students who enter school without English as a home or primary language and who score between 1 and 4.9 (i.e., below 5) on the WIDA English screening test at school entry.

language, though in practice, they may still serve some language minority students, including primary speakers of the partner language.[2]

In 2008, aiming to prepare its young people for a competitive global economy, Utah became the first U.S. state to invest in dual-language education statewide (Utah Senate 2016). It established a common DLI curriculum and teacher professional development program and provided schools with $10,000 for each new grade level in which they offered DLI. By the 2019–2020 academic year, the state featured 244 DLI schools dispersed across 22 of its 41 districts and enrolling about 57,900 DLI students, including 75 programs in Mandarin Chinese, 32 in French, 1 in German, 13 in Portuguese, 1 in Russian, and 113 in Spanish. Thirty-one of the Spanish programs were classified as two-way, meaning that at least one-third and no more than two-thirds of students had reported Spanish as their home or primary language at the time of enrollment.

This study estimates plausibly causal effects of that scale-up effort on the academic achievement of students in schools that launched DLI programs by comparing before-and-after academic performance within the same schools, net of observed school-by-year attributes. We estimate the effects of DLI access expansion on core academic performance in grades 3–6 across the 22 Utah school districts that eventually adopted DLI. We also estimate DLI-access effects on the reclassification of ELs as English proficient.

Our study contributes to international research on DLI in several important ways. First, because we observe up to 16 cohorts of students who entered first grade in the years before and after their schools implemented DLI programs, ours is the first study we know of to estimate plausibly causal effects of a statewide DLI program at scale. Second, we can estimate differential effects not only for one-way versus two-way programs, but within each program type, for primary English speakers versus ELs whose home language matches the school partner language. This is important when school systems face questions about where to situate DLI programs or how to allocate slots. Finally, the consistency of Utah's DLI instructional model across the state, with common curricula, teacher professional development, and instructional schedules, allows us to examine student demographic composition as a possible moderator of program effects.

---

2 These terms can have different meanings depending on context. In some places, one-way immersion connotes "English-only" immersion for English learners, or what we refer to in this paper as monolingual English instruction. The commonality in terms is that in one-way programs, a large majority of students in the classroom share a primary language and are working to learn the same partner language. Of course, whether the primary language they share is the socially dominant language matters to their experience and to their position within the social context. In this paper, most students in one-way programs are primary English speakers.

Recent papers have demonstrated identification challenges for staggered difference-in-difference designs, showing that two-way fixed effects are poorly identified when treatment varies at the level of unit-by-time period (Callaway and Sant'Anna 2021; de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021; Imai and Kim 2021; Kropko and Kubinec 2020). Kropko and Kubinec (2020) show that reasonable identification can be leveraged from unit fixed effects or time fixed effects and appropriate parametric specification of the corresponding longitudinal or cross-sectional component, depending on the question of interest. In this paper, we use student-by-grade analyses with base school fixed effects and linear time trends to model observable statewide achievement trends. We include Utah schools that never launched DLI programs in our comparison groups to guard against confounding of the time trend with treatment-school capacity. In addition, we test our descriptive within-school, within-cohort estimates for sensitivity to controls for observables, which we find to be minimal. We test for between-school selection on observables corresponding to DLI launches, and we control for these time-varying observable attributes at the level of grade-by-cohort-by-base school. We use first-cohort sample restrictions, early-versus-late adopting school restrictions, and placebo tests for sensitivity to unobserved selection of families and schools into treatment status. Our results are mostly insensitive to these tests. We test for heterogeneity of treatment effects over time by examining effects in early-treated cohorts and by fitting our main models at the student-by-grade level.

Descriptively, within cohorts and schools, DLI students outperform their peers in English, math, and science by about 18 %–25 % of a standard deviation (SD), even with individual and school-by-grade-by-year controls. However, subsample instrumental variable estimates and full-sample intent-to-treat (ITT) estimates are null for one-way programs and for primary English speakers in two-way programs. For ELs whose primary language (Spanish) matches the school partner language, we find benefits in mathematics of 0.13–0.15 SD In schools in which at least 40 % of students are primary speakers of the DLI partner language (in this case, Spanish), ITT effects of DLI access range from 0.06 SD in ELA to 0.09 SD in math to 0.095 SD in science. A key finding of our paper is that the primary language composition of the school strongly moderates DLI effects in Utah.

In this article, we briefly summarize the existing literature on DLI program effects and our contribution. We then describe Utah's DLI program and our dataset. Next, we present our analytic approach, followed by descriptive, IV, and ITT achievement results, robustness tests and moderation analyses, and EL reclassification analyses. We conclude with a discussion of implications for policymakers.

2 Extant Evidence

Recent estimates place the number of public dual-language immersion schools at about 3,000, implying that they account for about 2 % of public schools in the United States (Lam and Richards 2020). Though this figure remains modest, it represents a five-fold increase from nine years prior, when the leading estimate was 600 programs nationally (Center for Applied Linguistics 2011a, 2011b). In addition, public demand for these programs is strong in many cities across the U.S., yielding long wait lists and raising concerns about equitable access (Lam and Richards 2020; Williams 2017). The growing embrace of dual language education may be driven, at least in part, by economic concerns. Rigorous estimates of the earnings returns to bilingualism in North America range from 2-3% for non-English languages in the U.S. (Saiz and Zoido 2005), to 4–6% percent for French in Anglophone Canada (Christofides and Swidinsky 2010), and demand for bilingual workers in many sectors of the U.S. economy appears to be growing (Committee for Economic Development 2006). Meanwhile, European nations have increased dual language education offerings to better prepare young people for the global marketplace (Anghel, Cabrales, and Carro 2016). Families' demand for DLI programs in Europe seems also to depend on local economic returns to bilingualism, including proficiency in regional languages (Cappellari and Di Paolo 2018; Vega-Bayo and Mariel 2022; Yuki 2022).

Some evidence suggests that bilingualism carries cognitive advantages. In the lab, bilinguals outperform monolinguals on some types of cognition tests, including working memory, attention control, and task switching (e.g. Bialystok 2011; Bialystok and Craik 2010), though these laboratory studies are generally descriptive and not causal. Bilingualism has been linked to metalinguistic awareness (Cenoz 2003; Keshavarz and Astaneh 2004) and to children's social perceptive-taking (Fan et al. 2016; Greenberg, Bellana, and Bialystok 2013), and it may help connect young people with their heritage languages and cultures (Potowski 2004).

The potential advantages of bilingualism have raised questions about whether schools should cultivate it more broadly and at younger ages (Yuki 2022). Studies of French DLI programs serving primary English speakers in Canada and the U.S. have shown that immersion students perform as well as or better than their peers in English-tested content by about fifth grade (Barik and Swain 1978; Caldas and Boudreaux 1999; Lapkin, Hart, and Turnbull 2003; Marian, Shook, and Schroeder 2013), and some studies used baseline matching on pre-interevention characteristics (Lambert et al. 1993; Lambert, Tucker, and d'Anglejan 1973). More recently, Watzinger-Tharp, Swenson, and Mayne (2016) matched DLI students from 26 elementary schools to similar non-DLI students in matched non-DLI schools in Utah,

finding no significant differences in math performance in grade 3, but three additional percentile points of math growth from grades 3 to 4 among DLI students.

Most studies of DLI in the U.S. have focused on the academic performance of ELs, comparing DLI programs to other types of language support programs. Researchers have often examined differences in outcomes between ELs taught in English-only or transitional bilingual programs, which focus on English language development, versus those taught in developmental bilingual or DLI programs, which promote maintenance of the students' non-English home language. These studies have sometimes shown vastly better performance by ELs enrolled in two-way immersion programs than in transitional bilingual or English-only programs. But they have typically failed to adjust for the selection of families into programs (Collier and Thomas 2004; De Jong 2004; Lindholm-Leary and Block 2010).

Newer work has attempted to provide plausibly causal estimates of DLI effects on ELs and primary English speakers using econometric methods. Employing data from a large urban district and using extensive statistical controls, Umansky and Reardon (2014) examined EL reclassification rates of about 5400 Spanish-speaking ELs assigned to DLI, transitional or developmental bilingual programs, or monolingual English programs. They found that cumulative reclassification rates were highest for monolingual English programs until grade 7, at which point DLI programs surpassed them, reaching a 13-point advantage by the end of high school. In the same district, focusing on about 14,000 students adding fixed effects for parent program preferences, Valentino and Reardon (2015) found that ELs placed in DLI programs grew at a faster rate in ELA than their peers placed in transitional bilingual, developmental bilingual, and monolingual English programs. Their ELA performance exceeded that of similar peers in developmental bilingual and monolingual English programs by grade 6. In contrast, Kuziemko (2014), leveraged variation in schools' compliance with the Proposition 227 bilingual education ban in California to find positive effects of the ban on immigrant children's English speaking skills in the schools' Census areas, though children's fluency was based on Census self-reports. Chin, Daysal, and Imberman (2013) leveraged a bilingual education access threshold in Texas to show that bilingual education had no effect on the academic skills of primary Spanish speakers but increased the skills of primary English speakers in the same schools, perhaps by instructionally grouping students with different English-speaking skills.

Leveraging the launches of English-Spanish DLI programs in Spain, where most students were primary Spanish speakers learning the non-dominant partner language (English), Anghel, Cabrales, and Carro (2016) examined the sixth-grade achievement of about 4000 students whose preschools were selected to begin offering DLI programs when the students reached first grade. Comparing the sixth-grade exam scores of students in treated versus untreated schools across two years of

DLI program launches, similar to the approach we adopt in the current study, the authors found no statistically significant effects on subjects taught in Spanish (math and reading), and negative effects on those taught in the partner language of English (science, history, and geography).

Other recent studies have used data from oversubscribed DLI school lotteries to identify causal program effects. Steele and colleagues (2017) focused on about 1600 students randomized through pre-K or kindergarten lotteries in Portland, Oregon, finding higher ELA achievement among DLI lottery winners of 0.13 SD in grade 5 and 0.22 SD in grade 8. They found no statistically significant differences in effects between ELs and primary English speakers or between one-way and two-way programs, but the study was not powered to detect subgroup effects. They also found that ELs randomly assigned to DLI were reclassified at higher rates than their non-DLI peers by grade 6. Employing data from 510 kindergarten lottery applicants to two two-way-DLI programs in Charlotte-Mecklenburg, North Carolina, Bibler (2020) estimated per year ITT effects of 0.037 standard deviations in reading for primary English speakers and 0.055 standard deviations in math for ELs, with Local Average Treatment Effects about 25 % larger.

An important nuance is that one-way and two-way DLI programs may provide very different student experiences, especially for English learners and other students whose home language is not English. For an EL whose primary language matches the partner language, both types of programs offer access to at least half-time instruction in the primary language, facilitating access to academic content, but two-way programs may also offer greater affirmation of the partner language among peers and teachers in the school. For primary English speakers as well, two-way programs may offer a more complete language immersion experience among peers who are already fluent in the partner language. In addition, research on culturally relevant instruction suggests that cultural alignment between the partner language and a critical mass of students in the school could influence the effects of DLI programs. For instance, in describing the practices of culturally relevant instruction used by successful teachers of African American students, Ladson–Billings (1992, p. 387) noted that "[s]tudents' real life experiences are legitimated as part of the 'official curriculum.'" Moll and González (1994) described how schools in four language-minority communities helped students draw on the "funds of knowledge" in their communities, "taking full advantage of social and cultural resources in the service of academic goals" (p. 441). Paris and Alim (2014) built on this i.e. calling for "culturally sustaining pedagogy" (p. 85) that supports students' home languages and cultures to promote democratic ideals. Still, despite a substantial body of literature discussing the facets of culturally relevant instruction, only a few studies have sought to estimate achievement effects on a large scale (Sleeter 2012). To address the question causally, Dee and Penner (2016) undertook a

regression discontinuity study of high school ethnic studies courses in San Francisco, finding large positive effects on attendance, grade point averages, and credit acquisition among ninth graders identified as academically at risk. Their work also builds on a large-scale study that linked exposure to Mexican American studies courses in Arizona high schools to higher graduation and exit examination pass rates, even after accounting for an extensive set of student background characteristics (Cabrera et al. 2014).

Because our current study examines one-way and two-way program effects separately across many schools in Utah, it contributes toward disentangling language-access effects from cultural adjacency effects for English learners, where both program types provide the former, and two-way programs may provide the latter. We cannot definitively say that any differences in effects between one-way and two-way programs in Utah are attributable to cultural adjacency of the DLI programs, because other differences may exist in how the programs are run and taught. But with 32,941 ever-ELs in the sample, including 2824 in ever-one-way schools, 6040 in ever-two-way schools, and 24,077 in never-DLI schools), we can comment on these differences in a way that prior studies have been less able to do because of design or sample size constraints.

3 Policy Context and Data

3.1 Setting and Policy Context

With the 2008 passage of Senate Bill 41, Utah became the first U.S. state to launch a DLI expansion initiative, followed by Delaware in 2011 and North Carolina in 2013 (Delaware Department of Education 2011; North Carolina Department of Public Instruction 2020). The current analysis stems from a federally funded research-practice partnership designed to identify insights from Utah's DLI scale-up, which commenced in the fall of 2009.

Because kindergarten is optional in Utah, schools typically started new DLI programs with first grade and then added a grade each year (Utah State Board of Education 2020). As noted, schools received $10,000 for each new grade they established, and an additional $5000 per year in program maintenance thereafter, though expenditures reportedly represented only an additional 1 % of per-pupil funding across DLI schools.[3] The funding was designed to incentivize DLI program launches across the state, but decisions about launching

3 The state's estimate of 1 % additional spending per DLI pupil is comparable to estimates from a DLI cost analysis in Portland, OR (Steele et al. 2018) and includes teacher curriculum and professional development, which were centralized by the state. Thus, per-pupil benefits to school were

programs and allocating DLI slots were made by districts and schools. Most DLI districts reported that they used a lottery process when DLI slots were oversubscribed, but because districts did not systematically track lottery applicants, we were unable leverage random assignment in our study design. Districts also varied in the extent to which they prioritized slots for students in a school's residential school zone.

Guided by promising practices in other localities (Lyster 2007; Met 1994), Utah employs a 50/50, two-teacher model for grades 1–6, meaning elementary school students spend 50 % of their time in each language, switching teachers and languages midday. In grades 1–3, partner-language instruction focuses on math and social studies. In grades 4–5, it focuses on science and some math, and in grade 6, it focuses on science and social studies. Language arts in the partner language is taught in all grades but is emphasized in grades 4–6. As the programs expand into middle school (grades 7–8), students take two classes per day in the partner language. In high school, Utah makes college-level coursework available in the partner language for students who pass an Advanced Placement exam in that language.

To promote high instructional consistency across DLI schools, Utah developed uniform curricula for DLI programs and provides common professional development to DLI teachers. Teachers are hired from local labor markets where possible, and through international guest worker programs as needed (Watzinger-Tharp, Swenson, and Mayne 2016). One-way programs and two-way programs operate similarly, with common curriculum and teacher professional development opportunities. From a policy perspective, the key difference between them is in the primary language composition of the students they serve. This difference is of interest because it could affect the extent to which schools organize themselves around the needs of ELs and their families and emphasize the heritages of non-primary English speakers. Moreover, understanding this difference could help policymakers prioritize communities of greatest need when opening new DLI programs.

3.2 Analytic Sample

Our study uses an administrative dataset provided by the Utah State Board of Education. The analytic sample includes all public school students in the state of Utah. For ELA and math test scores measured at the end of grades 3 through 6, we use the entering kindergarten cohorts of 2001–2002 through 2014–2015. For science test scores measured at the end of grades 4 through 6, we use the entering kindergarten cohorts of 2001–2002 through 2013–2014. For EL reclassification, which is

likely smaller than 1 % over time. Our current intent-to-treat analysis is unable to disentangle funding benefits from other plausible mechanisms.

measured in grades 1 through 6, we use the kindergarten cohorts of 2001–2002 through 2016–2017. We restrict our analysis to a balanced panel of schools that are observed in all years. We observe ELA and math test score outcomes from grades 3 through 6 in academic years 2004–2005 through 2017–2018 (14 cohorts), science test scores from grades 4 through 6 in academic years 2005–06 through 2017–18 (13 cohorts), and initial English learners' reclassification status as EL or English proficient (reclassified) in 2002–2003 through 2017–2018 (16 cohorts). We exclude charter schools from the analysis because they were not part of the state's DLI scale-up policies.

Table 1 presents descriptive statistics for students with achievement test scores (tested in grades 3 or higher) in the analytic sample. Descriptive statistics are based on time-invariant characteristics or on characteristics measured in the students' initial year in Utah. Our sample includes 35,306 unique students who attended ever-one-way DLI schools, 15,896 students who attended ever-two-way DLI schools, and 171,975 unique students from Utah public schools that never launched DLI programs, as their inclusion helps us estimate statewide time trends. We treat the students' first observed year in a Utah public school as her base year. The base year represents kindergarten for 59 % of the sample, and first grade for 9 %.

Table 1 shows that ever one-way, ever two-way, and never-DLI schools were demographically quite different, with ever one-way schools having the fewest students eligible for subsidized meals and the most white, non-Hispanic students (27 % and 85 %, respectively), versus 38 % and 76 % in never-DLI schools, and 57 % and 50 % in ever two-way schools. Schools that never launched DLI programs were generally situated in districts farther from Salt Lake City. Their zip code demographics, shown in Table 1, were somewhat similar to the ever two-way schools except in their lower share of Limited English Proficient residents (2.46 % versus 5.19 %).

Importantly for our subgroup analyses, Hispanic or Latinx students constituted a substantial share of public school students in the state, representing 16 % in never-DLI schools, 10 % in ever-one-way schools, and 38 % in ever-two-way schools. Students in ever two-way schools showed much higher rates of ever-EL status, at 36 %, than the 8 % in ever one-way schools and the 14 % in never-DLI schools.

Table 1 also shows the distribution of DLI slot availability (dichotomous and slots per first-grader in students' first grade year), as well as the distribution of DLI languages among students who attended ever-DLI schools. Only 5–6% of students had slots available because even the earliest-adopting programs did not open until the fall of 2009, eight years after the first cohorts in the sample began kindergarten. Among students who attended ever one-way schools, about 36 % attended schools that eventually offered Spanish, and 42 % attended schools that eventually offered Mandarin Chinese. Schools that eventually offered French, German, or Portuguese

Table 1: Characteristics of sample in their first observed year, by base school category.

	Ever one-way		Ever two-way		Never DLI	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
N Students	**35,306**		**15,896**		**171,975**	
Individual characteristics						
Female	0.49	0.50	0.49	0.50	0.49	0.50
Asian	0.03	0.17	0.07	0.26	0.04	0.19
Black	0.01	0.11	0.03	0.16	0.02	0.14
Hispanic	0.10	0.29	0.38	0.49	0.16	0.37
American Indian	0.01	0.09	0.01	0.12	0.02	0.12
White	0.85	0.36	0.50	0.50	0.76	0.43
Race other/missing	0.01	0.08	0.01	0.08	0.01	0.09
Base free/red. lunch	0.27	0.45	0.57	0.50	0.38	0.48
Primary lang. not English	0.09	0.28	0.38	0.49	0.15	0.36
Ever EL	0.08	0.27	0.36	0.48	0.14	0.35
Primary/partner lang. match	0.03	0.07	0.33	0.17	.	.
Base special education	0.11	0.31	0.12	0.32	0.12	0.33
Ever migrant	0.00	0.04	0.02	0.14	0.01	0.08
Residential zip code characteristics						
Pct. w/bachelor degree	34.64	12.12	27.11	14.15	29.53	12.25
Pct. w/graduate degree	11.87	5.78	8.88	6.50	9.48	5.66
Pct. limited English proficient	1.54	1.45	5.19	3.51	2.46	2.71
Pct. supp. nutrition asst. prog.	6.80	3.41	10.57	4.78	8.67	4.37
Peer attributes in base school and grade						
Fraction white	0.87	0.11	0.52	0.22	0.78	0.20
Fraction free/red. lunch	0.26	0.16	0.56	0.23	0.36	0.23
Fraction base EL	0.01	0.04	0.06	0.15	0.02	0.08
Fraction base special Ed.	0.12	0.05	0.12	0.06	0.13	0.06
DLI access across kindergarten cohorts 2001–02 through 2014–15						
Had slots offered in gr. 1	0.06	0.24	0.05	0.23	0.00	0.00
Base school DLI language						
Spanish	0.36	0.48	1.00		–	–
Chinese	0.42	0.49	0.00		–	–
French	0.12	0.33	0.00		–	–
German	0.07	0.26	0.00		–	–
Portuguese	0.03	0.18	0.00		–	–

Table 1: (continued).

	Ever one-way		Ever two-way		Never DLI	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Within-student test score average in observed years						
ELA	0.08	0.87	−0.31	0.94	−0.06	−0.91
Math	0.13	0.85	−0.26	0.91	−0.02	−0.90
Science	0.11	0.84	−0.34	0.92	−0.06	−0.89

Residential zip code characteristics refer to the percent of adults age 25+ in the student's initial residential zip code who held bachelor's degrees, graduate degrees, were Limited English Proficient, or received federal benefits under the Supplemental Nutritional Assistance Program. Peer attributes refer to the fraction of students in the student's initial school and grade level who were white, qualified for free or reduced-price meals, were English learners at baseline, or qualified for special education services at baseline. The bold values are sample sizes.

accounted for 12 %, 7 %, and 3 %, respectively, of students in ever one-way schools. Among ever two-way schools, all DLI programs were offered in Spanish.

Finally, Table 1 presents students' average test scores on state accountability tests across all observed grades. Utah administered the Utah Criterion Referenced Tests (CRTs) in ELA, math, and science through spring 2013. In 2014, it transitioned to the Student Assessment of Growth and Excellence (SAGE). To make the assessment scales consistent across years, we standardize test scores statewide to have a mean of 0 and SD of 1 within subject, grade, and year. We observe that students attending ever one-way schools performed roughly 0.1 SD above the mean, whereas those at ever two-way schools performed about 0.3 SDs below the mean, on average, pooled across grades and years. Those attending never-DLI schools performed only slightly below the state mean, on average.

4 Empirical Strategy

4.1 Overall Approach

Based on prior studies, we expect to find that students do well in DLI programs relative to non-DLI peers in the same cohorts and schools. But such performance advantages may be driven by unmeasured characteristics like parents' education values or knowledge of their children's academic affinities. Therefore, we employ quasi-random within-school variation in first-grade cohorts' access to DLI enrollment slots. The opening of a DLI program in a school offers the opportunity to

assess student achievement before and after such openings. If such openings occur independently of other forces affecting the schools' achievement – i.e., if schools' selection into DLI programs is exogenous of other factors affecting the school – then aggregating the pre-post comparisons will estimate the average effect of adopting a DLI program. We leverage these openings in two ways. First, we use the opening as a plausibly exogenous instrument predicting a student's enrollment in DLI in a given cohort and year. However, we have clean enrollment data for the instrumental variable analysis for only two years of test scores, whereas we can estimate intent-to-treat effects over a 14-year time span, including 9 years in which at least some schools were treated. The availability of slots in a students' first grade school and year is the intent-to-treat variable of interest.

Across analyses, we are concerned with several kinds of selection bias: (1) student-level within cohorts and schools, which is mitigated by comparing treated to untreated cohorts within schools; (2) student-level between-cohort within-school comparisons, since students may migrate between schools in response to DLI opening, but with less ease than they would choose a program within their cohort year. We use grade-by-school-by-year control variables to mitigate this type of selection. The other potential source of bias is (3) cross-sectional between-school comparisons, since we might expect schools that do and do not adopt DLI to differ in terms of leadership, teacher capacity, community norms, and so forth. We mitigate (3) by using never-treated schools to estimate statewide time trends.

Because timing of treatment was staggered, time trends could be a confound (de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021; Imai and Kim 2021; Kropko and Kubinec 2020). In a classic differences-in-differences analysis, in which treatment commences at only one point in time, we minimize the risk of such confounding by looking for parallel pre-intervention trends in both treatment and control groups, and we aim qualitatively to rule out other policy changes that may have coincided with treatment timing (Angrist and Pischke 2008; Dynarski 2003; Imai and Kim 2021). With staggered treatment inception, the absence of confounding time trends is harder to establish empirically – because pre-and-post treatment period lengths vary by unit – but easier to establish *logically*. If treatment inception is staggered, it is less plausible that the treatment is confounded with a specific point-in-time intervention or policy change. That is the benefit of staggered treatment timing. The *problems* of staggered treatment timing are twofold. First, the effects of a treatment may be heterogeneous over time, and this heterogeneity by time would be observable only in units with longer post-treatment time trends (Imai and Kim 2021). Second, selection into treatment timing may be endogenous, with early-treated units having different baseline capacity or needs than later-treated units. Any endogeneity in timing of treatment (its correlation with time-invariant treatment effects) may affect the time trend – which represents our best estimate

of what would have happened in the absence of treatment (Goodman-Bacon 2021; Imai and Kim 2021; Kropko and Kubinec 2020).

4.2 Model Specifications

We address these internal validity threats systematically. At the most basic level, we conduct descriptive within-school, within-cohort comparisons using statistical controls for individual characteristics and grade-by-school-by-year characteristics. Because treatment and comparison students are part of the same cohorts, these models are not sensitive to time-trend misspecifications, but they are vulnerable to unobserved differences between individual students in the same cohorts and schools who enroll in immersion versus non-immersion programs. This empirical model, which we consider descriptive, is specified as in Equation (1):

$$y_{igcs} = \alpha_1 + \beta_1 DLI_{igcs} + \lambda_1 c_c + \boldsymbol{\delta_1}' \boldsymbol{S_s} + \boldsymbol{\varphi_1}' \boldsymbol{X_{igcs}} + \boldsymbol{\eta_1}' \boldsymbol{K_{gcs}} + \varepsilon_{1igcs} \quad (1)$$

where the current-grade test score y_{igcs}, of student i in grade g in cohort c from baseline school s, is predicted as a function of the student's current-year enrollment status in DLI. The average difference in y_{igcs} between DLI and non-DLI students in the same grade, cohort, and base school is given by β_1, holding constant the linear effect of kindergarten cohort (λ_1) and a vector of fixed effects for initial school ($\boldsymbol{\delta_1}$), as well as a matrix of baseline student characteristics $\boldsymbol{X_{igcs}}$, which includes gender, race/ethnicity, subsidized meal eligibility at baseline, whether the student was ever classified as an EL, whether the student has a home language other than English (regardless of EL classification), special education status at baseline, and migrant status at baseline. Matrix $\boldsymbol{K_{gcs}}$ captures grade-by-cohort-by-school attributes of baseline school s for cohort c in grade g, including the current-year percent in the student's grade who are white, subsidized-meal eligible at baseline, ever classified as EL, and special education eligible at baseline. The dependent variable y_{igcs} is a test score in ELA, math, or science for student i in tested grade g (grades 3–6) standardized statewide by subject, grade, and year to mean 0 and SD 1. The error term is given by ε_{igcs} and is clustered at the base-school level. Because we are interested in the distinctive mechanisms of one-way and two-way programs in Utah, we fit the models separately for each program type. Note that we estimate Equation (1) only within the final two academic years in the dataset, 2016–2017 and 2017–2018, because these are the only two years in which the DLI enrollment variable is available statewide.

Second, using the final two years of test score data, we leverage between-cohort differences in schools' offering of immersion slots as an instrumental variable for students' actual enrollment in such slots. The instrumental variable models are fit simultaneously using two-stage least squares estimation:

$$\mathrm{DLI}_{igcs} = \alpha_2 + \beta_2 \mathrm{ITT}_{cs} + \lambda_2 c_c + \boldsymbol{\delta}_2{}' \boldsymbol{S}_s + \boldsymbol{\varphi}_2{}' \boldsymbol{X}_{igcs} + \boldsymbol{\eta}_2{}' \boldsymbol{K}_{\mathbf{gcs}} + \varepsilon_{2igcs} \quad (2)$$

$$y_{igcs} = \alpha_3 + \beta_3 \widehat{\mathrm{DLI}}_{igcs} + \lambda_3 c_c + \boldsymbol{\delta}_3{}' \boldsymbol{S}_s + \boldsymbol{\varphi}_3{}' \boldsymbol{X}_{igcs} + \boldsymbol{\eta}_3{}' \boldsymbol{K}_{\mathbf{gcs}} + \varepsilon_{3igcs} \quad (3)$$

in which first-stage Equation (2) uses cohort-by-school variation in whether DLI slots were available $\left(\mathrm{ITT}_{cs}\right)$ to predict a students' actual enrollment in DLI. The fitted probability of DLI enrollment $\left(\widehat{\mathrm{DLI}}_{igcs}\right)$ in Equation (3) varies as a function of the arguably exogenous variation in the student's base school and kindergarten cohort, adjusting for the other terms in the model. The instrumental variable analysis therefore removes bias due to individual selection into DLI within same-school kindergarten cohorts. As in Equation (1), Equations (2) and (3) include a linear control for cohort year (c_c).

Because Equations (1)–(3) can be estimated only in 2016–2017 and 2017–2018, our preferred estimation approach uses an intent-to-treat analysis in which we employ 14 years of test score data at the student-by-grade level (2004–2005 through 2017–2018) to estimate reduced-form effects of immersion program access, with identification at the level of base school-by-cohort. This ITT model is specified as follows:

$$y_{igcs} = \alpha_4 + \beta_4 \mathrm{ITT}_{cs} + \lambda_4 c_c + \boldsymbol{\delta}_4{}' \boldsymbol{S}_s + \boldsymbol{\varphi}_4{}' \boldsymbol{X}_{igcs} + \boldsymbol{\eta}_4{}' \boldsymbol{K}_{\mathbf{gcs}} + \varepsilon_{4icsg} \quad (4)$$

where the intent-to-treat variable $\left(\mathrm{ITT}_{cs}\right)$, access to immersion in cohort c of base school s, is defined as above. Informed by research suggesting that immersion-program benefits increase with students' years of exposure, we model heterogeneous effects within schools over time by focusing on student-level estimates for each grade level g. As before, we adjust for families' possible endogenous selection into base schools in response to immersion availability by controlling for matrix $\boldsymbol{K}_{\boldsymbol{gcs}}$, which contains time-varying grade-by-cohort attributes of students in base school s, and we descriptively examine demographic trends following DLI program launches. We test for endogeneity in schools' timing of program launches by sub-setting our analyses among early-adopting versus late-adopting schools and by conducting placebo tests using the last pre-treatment year as the launch year. In lieu of cohort fixed effects, which could confound the cohort-by-school variation needed to estimate within-school DLI access effects, we use a linear specification of the time trend $\left(\lambda_4\right)$, as shown in Equations (1)–(3), and as recommended by Kropko and Kubinec (2020). To reduce sensitivity to within-unit confounds in the time trend, the trend is estimated using achievement scores statewide, including in the 82 % of public schools that never introduced DLI programs during the years observed in the dataset.

We fit Equation (4) for all students, and then separately for students never classified as English learners versus those ever classified as English learners whose

primary language matched the (current or eventual) DLI program language of their base schools. Disaggregating by never-EL versus language-matched EL status lets us examine how the primary language of the student moderates the effect of DLI access.

Among students classified as English learners at school entry, we also fit Equation (4) as a linear probability model in which the dependent variable is a dichotomous indicator of continued EL classification for student *i* in cohort *c,* base school *s,* and grade *g,* including observed grades 1–6. Reclassification from EL to English-proficient occurs in the year in which students pass the WIDA English proficiency test, changing their time-varying EL status from 1 to 0. Thus, a negative coefficient on the ITT variable in grade *g* would mean that ever-EL students with DLI access had a higher rate of English proficiency classification in that grade.

As shown in Figure 1, time trends in ELA, math, and science standardized test scores did differ across school categories. Though scores were standardized to a mean of 0 and a standard deviation of 1 within subject, grade, and year, trends were slightly negative in all school categories and subject areas due to the mix of grade levels observed for each cohort. Schools that eventually opened one-way DLI programs ($n = 55$) were the highest achieving across years, Schools that eventually opened Spanish two-way DLI programs ($n = 27$) were the lowest achieving and showed the steepest negative slopes in achievement. Schools that never introduced DLI programs ($n = 363$) scored between ever one-way and ever two-way schools. The fact that raw time trends differ modestly between school types attests to the value of including never-treated schools as comparison units so that time trends reflect patterns in the state as a whole.

5 Results

5.1 Descriptive Within-Cohort Estimates

We begin with descriptive results from the within-cohort OLS regression models described in Equation (1), using data only from 2016 to 2017 and 2017–2018 as noted. In Table 2, we present ELA, math, and science estimates for three different model specifications. The specifications assess sensitivity to controls for grade-by-cohort-by-school demographic attributes and students' individual baseline attributes, respectively (Shadish, Clark, and Steiner 2008). Across models and content areas, immersion students outperform non-immersion peers in the same cohorts and base schools by roughly a fifth to a quarter of a standard deviation, on average. As controls are added, the DLI achievement estimates for one-way programs in the top panel of the table decline slightly. On the other hand, estimates for two-way programs in the bottom panel remain consistent or increase slightly

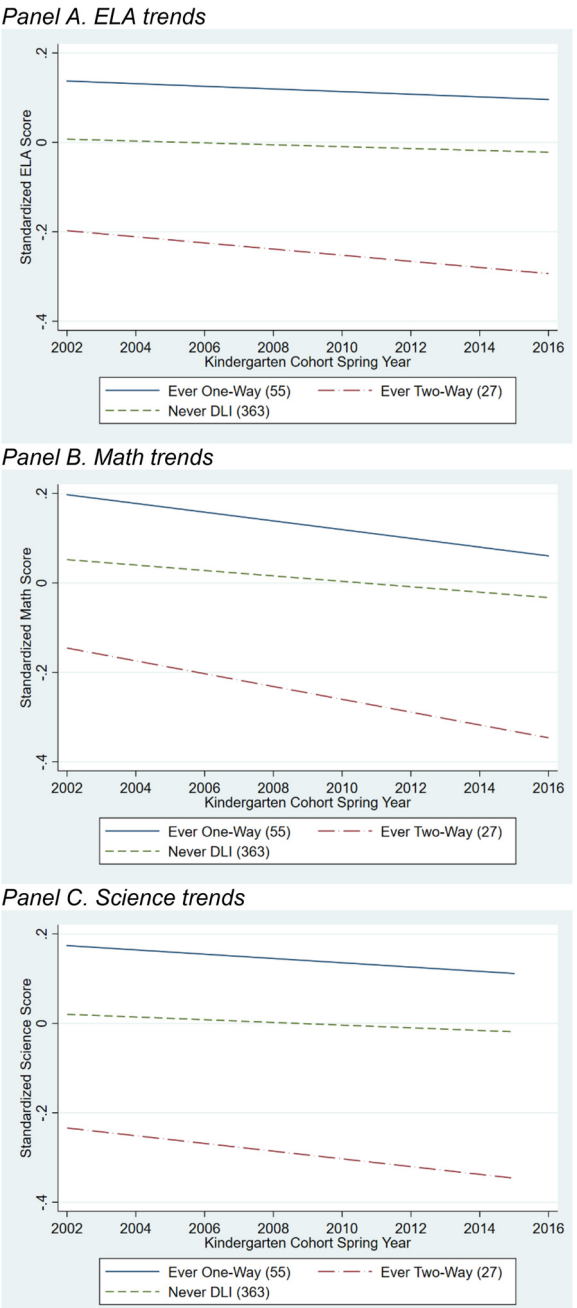

Figure 1: Standardized test score trend lines for ever-treated and never-treated schools. Graphs show lines of best fit for average standardized test scores over time, by school category, in each content area.

Table 2: Descriptive OLS estimates of DLI enrollment effects in 2016–17 and 2017–18.

	(1) ELA	(2) ELA	(3) ELA	(4) Math	(5) Math	(6) Math	(7) Science	(8) Science	(9) Science
A. One-way programs									
DLI enrolled	0.274***	0.251***	0.207***	0.243***	0.220***	0.183***	0.252***	0.229***	0.198***
	(0.017)	(0.017)	(0.016)	(0.020)	(0.020)	(0.018)	(0.018)	(0.018)	(0.017)
Obs.	514,196	514,196	514,196	514,498	514,498	514,498	461,144	461,144	461,144
R-sq.	0.003	0.010	0.101	0.002	0.012	0.080	0.002	0.012	0.079
Schools	418	418	418	418	418	418	418	418	418
B. Two-way programs									
DLI enrolled	0.243***	0.251***	0.240***	0.225***	0.233***	0.233***	0.200***	0.214***	0.223***
	(0.031)	(0.032)	(0.026)	(0.033)	(0.034)	(0.029)	(0.037)	(0.038)	(0.033)
Obs.	471,050	471,050	471,050	471,062	471,062	471,062	421,198	421,198	421,198
R-sq.	0.002	0.009	0.105	0.002	0.012	0.084	0.001	0.011	0.084
Schools	388	388	388	388	388	388	388	388	388
Grade-by-cohort-by-school controls		X	X		X	X		X	X
Individual controls			X			X			X

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models in the top and bottom panels are estimated at the student-by-grade level using the specification in Equation (1). Standard errors are clustered at the base school level. All models include base school fixed effects and a linear cohort trends, and some models, as indicated, include grade-by-cohort-by-school controls and individuals controls. Estimates pertain to the two academic years for which clean DLI enrollment data were available statewide.

as controls are added. These patterns may reflect the fact that two-way programs serve a higher concentration of students from low-income, non-white, and EL backgrounds, as shown in Table 1.

It is also consistent with selection tests we include in Appendix Table A1. For these tests, we regress individual grade-by-cohort-by-school characteristics from matrix $\boldsymbol{K}_{gcs}$ in Equation (4) into the school-by-cohort intent-to-treat variable (DLI availability), the linear cohort variable, and the set of base school fixed effects. We show in columns 1–4 that the within-school launches of one-way DLI programs are linked to a decrease of 3 percentage points in the share of same-cohort students who are eligible at baseline for subsidized meals and EL services. However, two-way DLI program launches, represented in columns 5–8, predict increases in these variables–of 5 and 11 percentage points, respectively–as well as a 6 percentage-point reduction in the fraction of white students.[4] The insight from Appendix Table A1 is therefore that some degree of systematic selection into DLI schools may occur following program launches. If selection on unobservable and observable attributes are similar, then we might expect small, positive intent-to-treat biases for one-way programs and negative biases for two-way programs.

5.2 Instrumental Variables Estimates of Local Average Treatment Effects

In Tables 3 and 4, we make further use of the plausibly endogenous regressor from Table 2, the student's DLI enrollment status in a given year. We use enrollment status as the first-stage dependent variable in a two-stage least squares model. The arguably exogenous school-by-cohort ITT indicator (DLI availability in the student's first grade year and base school) serves as an instrumental variable. This design has been widely used to estimate causal effects of sudden policy shocks on the local average treatment effects (LATEs) for those whose access to the intervention is fully regulated by the shock (e.g. Angrist 1993; Angrist and Chen 2011; Aparicio Fenoll 2018; Krueger and Zhu 2004; Steele, Murnane, and Willett 2010; Xu and Jaggars 2013). Insofar as the instrument strongly and monotonically predicts DLI enrollment and influences student achievement *only through* its effect on DLI enrollment, it can be used to estimate the causal effect of DLI enrollment on student achievement for individuals (compliers) whose DLI enrollment is modulated by the existence of DLI slots in their base school in their first-grade year (Angrist and Pischke 2008). Because an increase in DLI availability in one's base school is unlikely to *decrease* one's probability of DLI enrollment, the monotonicity assumption is logically satisfied.

4 These specifications include never-treated schools in the comparison group to stabilized time-trend estimates.

Table 3: 2SLS instrumental variables estimates of DLI enrollment effects in 2016–17 and 2017–18.

	One-way			Two-way		
	(1) ELA	(2) Math	(3) Science	(4) ELA	(5) Math	(6) Science
First stage						
DLI offered	0.316***	0.316***	0.304***	0.317***	0.317***	0.292***
	(0.012)	(0.011)	(0.012)	(0.020)	(0.019)	(0.021)
Instrument F-statistic	752.81	757.55	647.73	260.44	268.39	190.01
Second stage						
DLI enrolled	−0.000	−0.091∼	−0.028	0.046	0.040	−0.030
	(0.054)	(0.054)	(0.070)	(0.080)	(0.098)	(0.128)
Observations	514,196	514,498	461,144	471,050	471,062	421,198
Schools	418	418	418	388	388	388

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the school-by-grade level and include base school fixed effects, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level. The first- and second-stage estimates are based on simultaneously fitting Equations (2) and (3), respectively, and pooling students across grade levels. Variation in the first-stage estimates is due to sample size differences, given that science tests are not administered in grade 3 and that a few students have test scores available for ELA but not math, and vice versa. Estimates pertain only to the academic years for which clean DLI enrollment data are available statewide.

Recall that Table 2 showed that within-cohort sorting on *observable* attributes exerts only a small effect on outcomes (Oster 2019). This lends confidence that we have met the exclusion restriction of instrumental variable estimation, especially conditional on school-by-cohort and individual controls (Altonji, Elder, and Taber 2005). Still, because we have clean DLI enrollment data only for the final two years of the dataset, our power to estimate instrumental variable effects of DLI enrollment is constrained.

The top panel of Table 3 shows that a school's offer of DLI slots increases a student's probability of DLI enrollment by about 0.3 across one-way and two-way programs. The coefficients vary slightly among content areas due to different numbers of students with test scores in ELA versus math, and due to the fact that ELA and math are tested annually from grade 3, and science is tested annually from grade 4. First-stage F-statistics are 190 or higher for all models, which is well above the heuristic threshold of 10 for a single-instrument specification (Bound, Jaeger, and Baker 1995).

Yet, in the second-stage regressions in Table 3, we do not find instrumented overall benefits of DLI enrollment. Most second-stage estimates are null, and in one-way programs, we find a marginally significant LATE on math scores of −0.091

Table 4: 2SLS instrumental variables estimates of DLI enrollment effects in 2016–17 and 2017–18, for never-EL versus language-matched ever-EL students.

	One-way			Two-way		
	(1) ELA	(2) Math	(3) Science	(4) ELA	(5) Math	(6) Science
First stage						
A. Never EL						
DLI offered	0.317***	0.317***	0.306***	0.258***	0.258***	0.238***
	(0.012)	(0.011)	(0.013)	(0.021)	(0.021)	(0.021)
Instrument F-statistic	657.64	660.06	572.40	151.26	152.49	132.10
B. Language-matched EL						
DLI offered	0.370***	0.371***	0.370***	0.331***	0.332***	0.319***
	(0.055)	(0.055)	(0.052)	(0.049)	(0.048)	(0.051)
Instrument F-statistic	45.60	44.92	50.66	45.79	47.67	38.96
Second stage						
A. Never EL						
DLI enroll	−0.023	−0.103∼	−0.052	0.079	−0.061	−0.049
	(0.057)	(0.054)	(0.072)	(0.128)	(0.131)	(0.157)
Observations	445,344	445,930	400,403	391,910	392,276	351,612
Schools	418	418	418	388	388	388
B. Language-matched EL						
DLI enroll	−0.134	0.014	−0.072	0.149	0.307*	0.100
	(0.214)	(0.257)	(0.248)	(0.125)	(0.129)	(0.179)
Observations	3050	3052	2,660	14,869	14,801	12,894
Schools	50	50	50	28	28	28

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the school-by-grade level and include base school fixed effects, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level. The first- and second-stage estimates are based on simultaneously fitting Equations (2) and (3), respectively, and pooling students across grade levels. Variation in the first-stage estimates is due to sample size differences, given that science tests are not administered in grade 3 and that a few students have test scores available for ELA but not math, and vice versa. Estimates pertain only to the academic years for which clean DLI enrollment data are available statewide.

SD However, in Table 4, we disaggregate the instrumental variable estimation by whether students' primary language matches the school's partner language. The first-stage regressions remain strong for both groups, but the second-stage results are more nuanced. Among primary English speakers alone (that is, never-ELs), second-stage instrumental variable estimates are similar to those in Table 3. But for students ever classified as ELs whose primary language matches the two-way school's partner language (all of which are Spanish), we find a positive local average

treatment effect in math of 0.307 ($p < 0.05$). We consider this estimate with caution because instrumental variable estimates can be noisy (Small and Rosenbaum 2008), and it pertains only to the final two years of data for the EL subgroup. Yet it sets the stage for our remaining analyses, in which similar patterns emerge.

5.3 Intent-To-Treat Estimates of DLI Access Effects

For our intent-to-treat analyses, we have 14 years of test score data instead of two, so we have greater power for estimating ITT effects overall and by grade level. In Table 5, we show ITT estimates by grade level in one-way and two-way programs. In Table 6, we disaggregate these estimates for primary English speakers (Panel A) versus ELs whose home languages match the immersion partner language (Panel B).

In the top row of Table 5, cross-grade estimates are null for one-way and two-way programs, though the magnitudes are slightly larger for two-way programs. When we examine intent-to-treat effects by students' grade levels in subsequent

Table 5: ITT estimates overall and by grade for one-way versus two-way programs across all students.

Grade	One-way			Two-way		
	ELA (1)	Math (2)	Science (3)	ELA (4)	Math (5)	Science (6)
All grades (pooled)	−0.008	−0.009	−0.009	0.018	0.026	0.012
	(0.012)	(0.014)	(0.014)	(0.017)	(0.024)	(0.030)
3	−0.009	−0.038		0.020	0.014	
	(0.018)	(0.025)		(0.027)	(0.032)	
4	−0.028*	−0.054**	−0.043∼	0.018	−0.015	0.010
	(0.014)	(0.019)	(0.023)	(0.027)	(0.033)	(0.038)
5	−0.010	−0.023	−0.031∼	−0.029	−0.046	−0.020
	(0.019)	(0.025)	(0.019)	(0.027)	(0.031)	(0.039)
6	−0.032	−0.010	−0.043∼	0.030	0.039	0.029
	(0.021)	(0.031)	(0.024)	(0.035)	(0.036)	(0.035)
Schools	418	418	417	388	388	387
Obs. all gr.	2,884,166	2,783,637	2,282,164	2,616,152	2,526,724	2,066,698
Obs. base gr.	404,333	404,298	403,949	368,767	368,815	367,896
R-sq all gr.	0.058	0.051	0.047	0.061	0.055	0.048
R-sq base gr.	0.099	0.090	0.107	0.102	0.093	0.110

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the student-by-grade level and include never-ELs as well as ELs. Following the specification in Equation (4), models include base school fixed effects and a linear cohort trend, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level.

Table 6: ITT estimates overall and by grade for never-EL versus language-matched ever-EL students in one-way and two-way programs.

A. Never EL	**One-way**			**Two-way**		
Grade	**ELA (1)**	**Math (2)**	**Science (3)**	**ELA (4)**	**Math (5)**	**Science (6)**
All grades (pooled)	−0.007	−0.010	−0.008	0.030	0.017	0.010
	(0.013)	(0.015)	(0.014)	(0.020)	(0.022)	(0.029)
3	−0.012	−0.041		0.047∼	0.000	
	(0.019)	(0.026)		(0.028)	(0.033)	
4	−0.029*	−0.054**	−0.043∼	0.026	−0.037	−0.014
	(0.014)	(0.020)	(0.024)	(0.030)	(0.033)	(0.038)
5	−0.006	−0.022	−0.028	−0.028	−0.055∼	−0.048
	(0.020)	(0.025)	(0.020)	(0.030)	(0.032)	(0.036)
6	−0.032	−0.011	−0.048∼	0.015	0.025	0.026
	(0.022)	(0.032)	(0.025)	(0.029)	(0.032)	(0.040)
Schools base gr.	418	418	417	388	388	387
Obs. all gr.	2,525,550	2,434,829	1,999,138	2,210,517	2,131,910	1,747,257
Obs. base gr.	352,938	352,727	352,774	310,148	310,001	309,700
R-sq all gr.	0.089	0.066	0.066	0.090	0.067	0.068
R-sq base gr.	0.078	0.068	0.075	0.080	0.070	0.077
B. Language-matched EL						
All grades (pooled)	0.037	0.105	0.071	0.006	0.152***	0.025
	(0.051)	(0.064)	(0.050)	(0.028)	(0.035)	(0.047)
3	0.041	0.007		−0.019	0.130**	
	(0.072)	(0.104)		(0.040)	(0.038)	
4	−0.057	−0.109	−0.082	0.031	0.157**	0.070
	(0.088)	(0.090)	(0.070)	(0.047)	(0.049)	(0.056)
5	0.033	0.078	0.062	−0.025	0.056	0.042
	(0.093)	(0.077)	(0.076)	(0.057)	(0.069)	(0.077)
6	0.008	0.049	0.106	0.057	0.136*	0.060
	(0.084)	(0.068)	(0.081)	(0.063)	(0.063)	(0.050)
Schools base gr.	53	54	53	28	28	28
Obs. all gr.	14,971	14,592	11,695	69,004	67,388	53,704
Obs. base gr.	2322	2332	2,229	10,662	10,692	10,345
R-sq all gr.	0.071	0.044	0.043	0.056	0.030	0.027
R-sq base gr.	0.068	0.053	0.055	0.049	0.032	0.029

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the student-by-grade level. Following the specification in Equation (4), they include base school fixed effects and a linear cohort trend, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level.

rows of Table 5, we find evidence of modest negative effects of access to one-way DLI programs in fourth grade, with magnitude ranging from about −0.03 to −0.05, and marginally significant negative effects on science scores in all grades. This suggests that immersion education efforts may have detracted from or diluted science education in one-way schools relative to schools that never launched immersion programs. In two-way programs, estimates are null across grades.

We turn now to differential effects for students whose primary languages do and do not match the partner language. In our main models in Table 6, we again find null or negative effects for one-way programs for primary English speakers and null effects for one-way programs for children whose home language matches the partner language. This is important, since if the language match has a direct benefit for the learning of children who do not speak English at home, then we would expect to see these benefits in one-way programs. Instead, we see such evidence only in two-way programs. Even in our most conservative models, in Panel B of Table 6 we find that for students whose primary language of Spanish matches the partner language in a two-way school, estimates are positive, significant, and large in math, at 0.13 SD in grade 3, 0.16 SD in grade 4, 0.14 SD in grade 6, and 0.15 SD pooled across grades.[5]

5.4 Robustness Checks

Our analysis measures within-school changes in achievement following the launch of DLI programs. Thus, we are concerned about systematic sorting among families in response to DLI launches, heterogeneity in DLI program effects over time, and sorting among schools in the timing of their DLI launches. We conduct a series of robustness checks to address these concerns.

Estimates in Table 7 test the sensitivity of Table 6 results to cohort and time-trend restrictions. Panel A pertains to never-EL students, and Panel B pertains to ever-ELs whose primary language matches the partner language in ever-DLI schools. Expecting that families in the first DLI-eligible cohorts in a given school would have had less time to change schools in response to a newly launched program, we first run robustness tests in which, like Anghel, Cabrales, and Carro (2016) in Spain, we limit the ITT group to just the first DLI-eligible treatment cohort in each school. In Utah, more than half of these students were already enrolled in kindergarten in their base schools in the year before their DLI programs launched, meaning their families would have needed foreknowledge of program launches to sort into them deliberately. If estimates were substantially smaller for the first cohort than for all cohorts, this would suggest that estimates may be inflated by

5 These estimates are also robust not a quadratic instead of linear specification of the time trend.

Table 7: ITT estimate robustness tests for never-EL versus language-matched ever-EL students.

A. Never EL	**One-way**			**Two-way**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	ELA	**Math**	**Science**	**ELA**	**Math**	**Science**
1. First treated cohort	−0.005	−0.002	−0.010	0.024	0.004	0.006
	(0.019)	(0.017)	(0.018)	(0.036)	(0.032)	(0.039)
Observations	2,473,330	2,382,784	1,965,638	2,195,047	2,116,464	1,737,369
R-squared	0.089	0.066	0.067	0.090	0.067	0.068
Schools	419	419	418	389	389	388
2. First 4 treated cohorts	−0.009	−0.013	−0.013	0.020	0.002	0.006
	(0.013)	(0.014)	(0.014)	(0.022)	(0.022)	(0.028)
Observations	2,514,332	2,423,647	1,994,110	2,207,019	2,128,425	1,745,697
R-squared	0.089	0.066	0.066	0.090	0.067	0.068
Schools	419	419	418	389	389	388
3. Early-adopting schools only	−0.009	−0.009	−0.009	0.031	0.019	0.007
	(0.014)	(0.016)	(0.015)	(0.022)	(0.025)	(0.031)
Observations	2,383,211	2,297,834	1,885,696	2,153,827	2,077,183	1,702,574
R-squared	0.089	0.066	0.067	0.090	0.067	0.068
Schools	399	399	398	376	376	375
4. Late-adopting schools only	−0.006	−0.029	−0.020	0.014	−0.009	0.034
	(0.021)	(0.027)	(0.036)	(0.047)	(0.033)	(0.056)
Observations	2,219,097	2,139,911	1,755,643	2,133,448	2,057,643	1,686,884
R-squared	0.090	0.066	0.068	0.090	0.067	0.068
Schools	381	381	380	374	374	373
5. Placebo 1st cohort vs previous	−0.025	−0.022	−0.029	−0.035	−0.030	−0.043∼
	(0.016)	(0.018)	(0.018)	(0.022)	(0.031)	(0.023)
Observations	2,450,662	2,360,217	1,947,817	2,188,623	2,110,063	1,732,467
R-squared	0.089	0.066	0.067	0.090	0.067	0.068
Schools	419	419	418	389	389	388
B. Language-matched EL	**One-way**			**Two-way**		
	(7)	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
	ELA	**Math**	**Science**	**ELA**	**Math**	**Science**
1. First treated cohort	−0.023	−0.003	−0.021	0.005	0.120**	0.017
	(0.052)	(0.075)	(0.059)	(0.037)	(0.038)	(0.046)
Observations	13,018	12,632	10,447	61,767	60,133	49,100
R-squared	0.072	0.045	0.045	0.057	0.030	0.027
Schools	48	49	49	28	28	28
2. First 4 treated cohorts	0.036	0.097	0.065	0.009	0.151***	0.030
	(0.052)	(0.066)	(0.055)	(0.028)	(0.033)	(0.046)
Observations	14,526	14,143	11,495	67,646	66,024	53,167
R-squared	0.071	0.043	0.044	0.056	0.030	0.027
Schools	56	56	55	28	28	28

Table 7: (continued)

B. Language-matched EL	One-way			Two-way		
	(7) ELA	(8) Math	(9) Science	(10) ELA	(11) Math	(12) Science
3. Early-adopting schools only	0.022	0.137	0.046	−0.013	0.135*	0.014
	(0.065)	(0.093)	(0.099)	(0.035)	(0.046)	(0.038)
Observations	7714	7552	5,956	34,602	33,777	26,901
R-squared	0.079	0.051	0.055	0.053	0.028	0.027
Schools	38	38	38	15	15	15
4. Late-adopting schools only	−0.144	−0.193	−0.203	−0.036	0.157∼	−0.005
	(0.133)	(0.181)	(0.162)	(0.058)	(0.082)	(0.146)
Observations	7257	7040	5,739	34,402	33,611	26,803
R-squared	0.071	0.045	0.041	0.060	0.034	0.032
Schools	18	18	17	13	13	13
5. Placebo 1st cohort versus previous	0.074	0.143∼	0.129	0.032	0.095*	0.037
	(0.070)	(0.073)	(0.077)	(0.046)	(0.043)	(0.039)
Observations	12,278	11,892	9,883	58,214	56,574	46,498
R-squared	0.069	0.043	0.043	0.058	0.031	0.028
Schools	44	45	44	28	28	28

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the student-by-grade level. Following the specification in Equation (4), they include base school fixed effects and a linear cohort trend, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level. Early-adopting schools launched DLI in fall 2009–2012; late-adopting schools launched fall 2013–2017.

systematic sorting of families in later cohorts. Yet, for language-matched ELs, we still find statistically significant cross-grade math benefits of 0.12 SD ($p < 0.01$), relative to an estimate of 0.15 in Table 6. First-cohort estimates in the other columns of Panels A and B are null, similar to those for all cohorts.

Next, we examine whether a school's implementation duration is related to its outcomes by focusing on just the first four ITT cohorts in a DLI school relative to the cohorts that preceded them in the same school. This lets us gauge whether a longer implementation duration is associated with stronger outcomes, though our grade-level analyses in Tables 5 and 6 address this question in a more granular way. In Panels A and B of Table 7, we again find estimates very similar to those in Table 6, including a cross-grade math estimate of 0.151 ($p < 0.001$) for language-matched ELs.

We then focus on early-adopting (fall 2009–fall 2012) versus late-adopting (fall 2013–fall 2017) DLI schools in case effects are correlated with timing of program launches. For language-matched ELs in two-way programs, math estimates are positive and significant in early-adopting schools (0.135 SD, $p < 0.05$), and positive and

marginally significant in late-adopting schools (0.157 SD, $p < 0.1$). This suggests that effects are not driven by unobserved differences in the educational capacity of early versus late adopters.

Finally, we conduct a placebo first-cohort analysis in which we designate the first treated cohort as being one year earlier than it actually was. We limit the analyses to that cohort and all previous ones, similar to the first-cohort analysis. The placebo test examines the possibility that pre-existing capacity in treated schools may be mistaken for DLI effects. In the placebo test, we find null or negative and marginally significant estimates for never-ELs. For language-matched ELs, however, we find modest evidence of pre-existing positive trends in math in both one-way and two-way schools. In one-way schools, the placebo test yields a positive and marginally significant estimate in math of 0.143 ($p < 0.1$) for language-matched ELs – an effect not seen in the real ITT cohorts, and which we interpret as probably noise. In two-way schools, the positive estimate of 0.095 ($p < 0.05$) suggests that a portion of the positive math effect of 0.152 for language-matched ELs in Table 6 could plausibly be driven by pre-existing positive trends for ELs in language-matched schools. Yet, the placebo coefficient is also consistent with the notion that DLI confers benefits not *just* through primary language instruction, which is present for language-matched ELs in one-way programs, but through changes in the cultural responsiveness of school norms and practices. Such norms could plausibly have shown modest benefits for the final pre-treatment cohort in two-way schools.

5.5 Heterogeneity of Within-School Estimates by Student Composition

Guided by evidence in Table 6 that language-matched ELs in two-way programs, but not one-way programs, show mathematics benefits in response to DLI program launches, we examine schools' linguistic composition as a possible mediator of DLI launch effects. To do so, we examine the extent to which the launch effects of a DLI program vary with share of students in the school whose primary language matches the school's partner language. It should be noted that 96.4 % of students with primary/partner language matches are Spanish speakers; fewer than 1 % each are primary speakers of Chinese, Portuguese, French, or German. Specifically, we interact the intent-to-treat variable, ITT_{cs}, with the fraction of students in the school whose primary language matched the school's DLI partner language in the year preceding the DLI program launch ($match_{cs}$), as shown in Equation (5). The interaction coefficient, γ_5, represents the differential effect of DLI access for each unit increase (0–1) in the share of language-matched students, controlling for the other terms in the model:

Table 8: Interaction of DLI access with share of language-matched students in the school in the pre-treatment year, pooling one-way and two-way programs.

	(1) ELA	(2) Math	(3) Science
DLI offered (school-by-cohort)	−0.010	−0.012	−0.019
	(0.013)	(0.014)	(0.014)
DLI offered * share of students with primary language match (0–1)	0.111∼	0.164*	0.181∼
	(0.061)	(0.076)	(0.093)
Observations	3,099,723	2,992,455	2,451,025
R-squared	0.110	0.086	0.096
Schools	447	447	446

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the student-by-grade level. Following the specification in Equation (5), they include base school fixed effects and a linear cohort trend, as well as individual and school-by-grade-by-year controls, with standard errors clustered at the base school level. The fraction of language-matched students is measured on a 0 to 1 scale based on the year prior to the DLI program launch.

$$y_{igcs} = \alpha_5 + \beta_5 \mathrm{ITT}_{cs} + \gamma_5 \mathrm{ITT}(\mathrm{match})_{cs} + \lambda_5 c_c + \boldsymbol{\delta_5}' \boldsymbol{S_s} + \boldsymbol{\varphi_5}' \boldsymbol{X_{igcs}} + \boldsymbol{\eta_5}' \boldsymbol{K_{gcs}} + \varepsilon_{5icsg} \tag{5}$$

Our interaction coefficients in Table 8 show the differential, estimated effect of a DLI program launch on a hypothetical school with 100 % language-matched students as compared to a school with 0 % language-matched students. In practice, the average share of language matched students was 1.4 % in ever-one-way schools, and 13 % in ever-two-way schools, ranging as high as 33 % and 83 % in each group, respectively.

Main effects, denoting predicted effects of DLI program launches in schools without language-matched students, remain null. But we find substantial interaction effects of 0.11–0.18 SD across the three content areas, with marginally significant effects in ELA and science, and a statistically significant effect of 0.16 SD ($p < 0.05$) in math. This means that as a school's share of language-matched students rises by 10 percentage points (one-tenth of a one-unit change), the intent-to-treat effect of a DLI launch would be expected to increase by about 0.016 SD (one-tenth of the coefficient size), net of other terms in the model.

Of course, this analysis assumes a linear interaction effect. We then relax that assumption by instead interacting the ITT indicator in Equation (5) with a set of categorical variables representing natural increments in the fraction of language-matched students in the school. The categories, reflecting observed variation in the share of language-matched students, are [0–0.01) matched (the reference category),

Table 9: Interaction of DLI access with categorical fractions of language-matched students in the school in the pre-treatment year, pooling one-way and two-way programs.

	(1) ELA	(2) Math	(3) Science
DLI offered (y/n)	−0.011	−0.026∼	−0.024
	(0.016)	(0.015)	(0.018)
DLI offered × lang match group 2	0.007	0.041	0.021
	(0.021)	(0.025)	(0.025)
DLI offered × lang match group 3	0.034	0.045	0.061
	(0.040)	(0.046)	(0.052)
DLI offered × lang match group 4	0.073*	0.117**	0.119*
	(0.030)	(0.039)	(0.051)
Combined effect group 1	−0.011	−0.026∼	−0.024
	(0.016)	(0.015)	(0.018)
Combined effect group 2	−0.004	0.015	−0.003
	(0.015)	(0.020)	(0.018)
Combined effect group 3	0.023	0.019	0.037
	(0.037)	(0.043)	(0.049)
Combined effect group 4	0.062*	0.091*	0.095*
	(0.026)	(0.036)	(0.048)
Observations	3,099,723	2,992,455	2,451,025
R-squared	0.110	0.086	0.096
Schools	447	447	446

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the student-by-grade level. Adapting Equation (5) to make the match_{cs} variable categorical, the model includes base school fixed effects and a linear cohort trend, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level. Note that group 1 (omitted) is schools with <0.01 language-matched students in the final pre-DLI year ($n = 39$ schools, plus 408 never-DLI schools). Group 2 has [0.01–0.2) matched students ($n = 30$ schools); group 3 has [0.2–0.4) matched (9 schools); and group 4 has at least 0.4 matched (12 schools).

as well as [0.01–0.2) matched, [0.2–0.4) matched, and [0.4–1] matched.[6] We use the *lincom* command in Stata 15.1 to test the statistical significance of the linear combination of the ITT main effect and interaction effects in each content area. The ITT-by-category interaction coefficients and the linear combinations of main and interaction effects are shown in Table 9.

In Table 9, we find that the ITT effects of DLI access are driven by schools with at least 40 % of students whose primary languages match the partner language. In

6 These categories include 443, 30, 9, and 12 schools, respectively. The lowest-matched category of 443 schools includes 404 never-DLI schools.

this category, access to a DLI program predicts an additional 0.06 SD in ELA scores, 0.09 SD in math scores, and nearly 0.1 SD in science scores, as shown in the bottom panel of the table. We can reject the null hypothesis at the 5 % level for all three subjects. In schools with 20 % to almost 40 % language-matched students (category 3), predicted effects are positive in magnitude but do not reach statistical significance. These estimates suggest that DLI benefits may indeed rise in a somewhat linear fashion with the fraction of language-matched students in the school.

5.6 English Learner Reclassification Estimates

In light of the test score benefits we find for language-matched ELs with access to two-way DLI programs, we also examine EL students' persistence in English learner status over time. Given evidence from prior studies, we would anticipate that students with DLI access would be reclassified as English-proficient from grades 6 or 7 onward (Steele et al. 2017; Umansky and Reardon 2014). When we examine reclassification rates to English proficiency among students ever classified as ELs, that is roughly what we find.

Table 10 focuses on English-language proficiency trajectories for language-matched ELs. Here, adapting Equation (4) as a linear probability model, the dependent variable is that a student ever classified as EL in Utah public schools remains classified as EL in grades 1 through 6. Thus, a negative coefficient means that students in the ITT group are exiting EL status at a higher rate. Our analysis anchors students to their base school, cohort group, and initial EL status. It estimates the probability of their subsequent classification as ELs in each grade level.

For one-way programs, shown in the left two columns, we find no statistically significant differences in rates of EL classification as a function of DLI program launches until grade 6. At that point, non-language matched ELs with DLI access show lower rates of EL persistence by 5.1 percentage points, a marginally significant effect ($p < 0.1$). In two-way programs, for students with a primary-partner language match, rates of EL persistence by ITT status are similar until grade 5, at which point the fraction of ITT students who are still classified as ELs is about 6.2 percentage points lower than for other students ($p < 0.05$). If these reclassification effects are real, we would expect them to persist in subsequent grades. For these language-matched ELs in two-way programs, the fifth-grade effect is followed by an estimate similar in magnitude (−0.041), though not significance, in grade 6. For ELs whose primary and school partner languages *do not* match, we see those in two-way programs exiting EL status 5.3 percentage points faster in grade 3. However, this effect does not seem to persist into subsequent grades, suggesting that it may be an anomaly of the third-grade non-language-matched sample.

Table 10: Estimated ITT effects on the probability of being classified as EL in each year among those ever classified as EL.

Grade	One-way		Two-way	
	Primary/partner language match (1)	No language match (2)	Primary/partner language match (3)	No language match (4)
1	0.011	0.003	0.001	0.000
	(0.016)	(0.012)	(0.008)	(0.013)
2	0.021	−0.010	0.005	−0.010
	(0.035)	(0.020)	(0.015)	(0.024)
3	0.023	−0.017	0.010	−0.053*
	(0.033)	(0.027)	(0.022)	(0.027)
4	0.000	−0.018	−0.022	−0.015
	(0.030)	(0.027)	(0.023)	(0.032)
5	0.002	−0.017	−0.062*	−0.015
	(0.033)	(0.023)	(0.023)	(0.042)
6	−0.027	−0.051∼	−0.041	−0.033
	(0.049)	(0.027)	(0.024)	(0.030)
Schools gr. 1	50	410	28	380
Obs. gr. 1	2,410	49,610	11,223	48,546
R-sq gr. 1	0.062	0.077	0.111	0.077

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models estimate the ITT effect of an ever-EL student remaining as EL-classified in each of grades 1 through 6. Following Equation (4), in which English proficiency status is the dichotomous, dependent variable of interest in each grade, the models include base school fixed effects and a linear cohort trend, as well as individual and grade-by-cohort-by-school controls, with standard errors clustered at the base school level.

6 Discussion and Conclusion

As demand grows for public school programs that are both culturally inclusive and academically challenging, DLI programs hold natural appeal. Demand for these programs is growing in the U.S., with lotteries and wait lists in many localities, raising concerns about gentrification and the crowding out of students whose primary languages match the schools' partner languages (Lam and Richards 2020; Williams 2017). This study adds to the growing research on DLI programs by examining the effects of DLI program launches on schools' subsequent achievement across a large scale-up effort in one U.S. state. Though Utah's DLI students spent half of their elementary instructional hours learning content in a language other than English, we find little evidence of academic harm for one-way or two-way programs. Most estimates are null, suggesting that schools' launches of DLI programs had little effect

on their students' academic performance in English, math, or science, all of which are tested in English. However, for Spanish-speaking English learners, the launch of a two-way Spanish DLI program in their base schools predicted notably higher mathematics performance than we would have anticipated, at 13 %–15 % of a standard deviation in grades 3, 4, and 6. For these students, it also predicted higher rates of EL reclassification to English-proficient by grade 5.

Most striking is the role that a school's language composition plays in moderating the DLI effect. Raising the fraction of language-matched students by 10 percentage points predicts an additional 0.016 SD of student achievement in math, 0.011 SD in ELA, and 0.018 in science, all significant at the 0.1 level or better. In schools in which at least 40 % of students have primary languages matching the partner language, DLI availability boosts scores by 0.062 SD in ELA, 0.091 in math, and 0.095 in science, all of which are significant at the 0.05 level.

Our findings about primary and partner language alignment comport with evidence about the academic benefits of culturally relevant instruction (Cabrera et al. 2014; Dee and Penner 2016). They suggest a need to better understand language and cultural practices in these schools. Schools that offer two-way DLI or serve a larger share of primary speakers of the partner language may be more responsive to the needs of language-minority students and families, creating a more culturally and linguistically sustaining environment. Of course, from a policy perspective, creating two-way programs depends on having a critical mass of students who share a common, non-English language. They may be less feasible in communities that serve students from diverse language backgrounds or from mostly English-speaking backgrounds. Future research should examine implementation differences as a function of schools' language composition. It would be useful to understand how DLI-access effects covary with schools' cultural norms, parent communication practices, and the racial/ethnic alignment of teachers and students.

It is worth acknowledging that our dependent variables are not the only foci of DLI programs in Utah or elsewhere. Utah's stated intention in rapidly scaling DLI was to prepare a bilingual and biliterate workforce. Because students not enrolled in DLI were not tested in bilingualism or biliteracy, our analysis focuses on the effects of program launches on students' achievement in core content tested in English. Fortunately, given that the study is part of a broader research-practice partnership, we can interpret these estimates alongside companion research in Utah. Specifically, Watzinger-Tharp, Rubio, and Tharp (2018) found that Utah students in Chinese, French, and Spanish DLI programs were meeting or exceeding partner-language performance benchmarks in grades 3, 6, and 8, with average eighth-grade skill attainment of Intermediate Mid-to-High in Spanish and French and Intermediate Low in Chinese. These levels already exceed what would be expected in traditional secondary school language electives (Burkhauser et al. 2016;

Xu, Padilla, and Silva 2015). In a follow-up study, the team found that well over 80 % of ninth graders reached all four of the state's proficiency benchmarks in Spanish and French, and over 60 % achieved listening and reading benchmarks in Chinese (Watzinger-Tharp, Tharp, and Rubio 2021). In other words, Utah DLI students appear to meet the state's goals of moving students toward bilingualism and biliteracy. Given this progress, future work should examine ITT effects on AP language credit completion, high school graduation, postsecondary attainment, and labor market outcomes.

In the interim, our study offers several key takeaways. First, DLI students in Utah outperformed their same-cohort, same-school peers by at least a fifth of a standard deviation on average, but much of this was likely due to the unobserved preferences and skills of families who chose these programs. A larger policy question is how these programs affected subsequent student achievement in schools that launched them. Our estimates suggest that the programs' average effects on students' learning of English, math, and science through grade 6 were minimal. However, effects on these subjects were positive in schools where many students were primary speakers of the partner language, and especially for ELs in two-way programs whose partner languages matched their primary languages. If we view schools' mission as preparing a well-informed global citizenry, then the ability of these programs to maintain students' core learning while equipping them with two languages may represent a promising model in its own right.

Acknowledgments: This work was made possible by research-practice partnership grant #R305H170005 from the U.S. Department of Education's Institute of Education Sciences. We are grateful to Kristin Campbell at the Utah State Board of Education for preparing the anonymized administrative datasets. This work has benefitted from feedback we received from Utah dual language program leaders, participants at meetings of the Association of Public Policy Analysis and Management, the Association of Education Finance and Policy, the American Educational Research Association, the University of Pennsylvania's IES Doctoral Fellowship Speaker Series, the University of Arkansas Department of Education Reform Speaker Series, and insightful peer reviewers. Any errors are our own.
Research funding: This work was supported by the Institute of Education Sciences, U.S. Department of Education (R305H170005).

Appendix

Table A1: Selection test regressing grade-by-cohort-by-school attributes on DLI program launches.

Variables	One-way				Two-way			
	(1) Frac white	(2) Frac FRL	(3) Frac ever-EL	(4) Frac sped	(5) Frac white	(6) Frac FRL	(7) Frac ever-EL	(8) Frc sped
DLI offered	0.002	−0.031***	−0.032***	−0.001	−0.061***	0.047**	0.110***	−0.005
	(0.005)	(0.008)	(0.007)	(0.003)	(0.011)	(0.014)	(0.020)	(0.003)
Linear cohort covariate	−0.002***	0.003***	0.009***	−0.003***	−0.003***	0.004***	0.011***	−0.003***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
Obs.	4,346,674	4,346,674	4,346,674	4,346,674	3,949,017	3,949,017	3,949,017	3,949,017
R-sq.	0.006	0.008	0.149	0.050	0.014	0.014	0.188	0.048
Schools	419	419	419	419	389	389	389	389

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\sim p < 0.1$. Models are estimated at the student-by-grade level using an adaptation of Equation (4) in which each grade-by-cohort-by-school covariate in matrix $\boldsymbol{K}_{gcs}$ is individually regressed on the ITT indicator (ITT_{cs}), the linear cohort term (c_c), and the base school indicators ($\boldsymbol{S}_s$), with no additional controls. Standard errors are clustered at the base school level.

References

Altonji, J. G., T. E. Elder, and C. R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113 (1): 151—84.

Anghel, B., A. Cabrales, and J. M. Carro. 2016. "Evaluating a Bilingual Education Program in Spain: The Impact beyond Foreign Language Learning." *Economic Inquiry* 54 (2): 1202—23.

Angrist, J. D. 1993. "The Effect of Veterans Benefits on Education and Earnings." *Industrial and Labor Relations Review* 46 (4): 637—52.

Angrist, J. D., and S. H. Chen. 2011. "Schooling and the Vietnam-Era GI Bill: Evidence from the Draft Lottery." *American Economic Journal: Applied Economics* 3 (2): 96—118.

Angrist, J. D., and J.-S. Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Aparicio Fenoll, A. 2018. "English Proficiency and Mathematics Test Scores of Immigrant Children in the US." *Economics of Education Review* 64: 102—13.

Barik, H. C., and M. Swain. 1978. "Evaluation of a French Immersion Program: the Ottawa study through Grade Five." *Canadian Journal of Behavioural Science* 10 (3): 192—201.

Bialystok, E. 2011. "Reshaping the Mind: The Benefits of Bilingualism." *Canadian Journal of Experimental Science* 65 (4): 229—35.

Bialystok, E., and F. I. M. Craik. 2010. "Cognitive and Linguistic Processing in the Bilingual Mind." *Current Directions in Psychological Science* 19 (1): 12—23.

Bibler, A. 2020. "Dual Language Education and Student Achievement." *Education Finance and Policy*. (online first). https://doi.org/10.1162/edfp_a_00320.

Bound, J., D. A. Jaeger, and R. M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443—50.

Burkhauser, S., J. L. Steele, J. Li, R. O. Slater, M. Bacon, and T. Miller. 2016. "Partner-language Learning Trajectories in Dual-Language Immersion: Evidence from an Urban District." *Foreign Language Annals* 49 (3): 415—33.

Cabrera, N. L., J. F. Milem, O. Jaquette, and R. W. Marx. 2014. "Missing the (Student Achievement) Forest for All the (Political) Trees: Empiricism and the Mexican American Studies Controversy in Tucson." *American Educational Research Journal* 51 (6): 1084—118.

Caldas, S. J., and N. Boudreaux. 1999. "Poverty, Race, and Foreign Language Immersion: Predictors of Math and English Language Arts Performance." *Learning Languages* 5: 4—15.

Callaway, B., and P. H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200—30.

Cappellari, L., and A. Di Paolo. 2018. "Bilingual Schooling and Earnings: Evidence from a Language-In-Education Reform." *Economics of Education Review* 64: 90—101.

Cenoz, J. 2003. "The Additive Effect of Bilingualism on Third Language Acquisition: A Review." *International Journal of Bilingualism* 7: 1—87.

Center for Applied Linguistics. 2011a. Directory of Foreign Language Immersion Programs in U.S. Schools. http://www.cal.org/resources/immersion/.

Center for Applied Linguistics. 2011b. Directory of Two-Way Bilingual Immersion Programs in the U.S. www.cal.org/twi/directory/.

Chin, A., N. M. Daysal, and S. A. Imberman. 2013. "Impact of Bilingual Education Programs on Limited English Proficient Students and Their Peers: Regression Discontinuity Evidence from Texas." *Journal of Public Economics* 107: 63—78.

Christofides, L. N., and R. Swidinsky. 2010. *The Economic Returns to a Second Official Language: English in Quebec and French in the Rest-Of-Canada* (2010-04). Retrieved from Nicosia, Cyprus https://core.ac.uk/download/files/153/6625815.pdf.

Collier, V. P., and W. P. Thomas. 2004. "The Astounding Effectiveness of Dual Language Education for All." *NABE Journal of Research and Practice* 2 (1): 1—20.

Commission on Language Learning. 2017. *America's Languages: Investing in Langauge Education for the 21st Century*. Retrieved from Cambridge, MA.

Committee for Economic Development. 2006. *Education for Global Leadership: The Importance of International Studies and Foreign Language Education for U.S. Economic and National Security*. Retrieved from Washington, DC:.

de Chaisemartin, C., and X. D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *The American Economic Review* 110 (9): 2964—96.

De Jong, E. 2004. "L2 Proficiency Development in a Two-Way and a Developmental Bilingual Program." *NABE Journal of Research and Practice* 2 (1): 77—108.

Dee, T. S., and E. K. Penner. 2016. "The Causal Effects of Cultural Relevance: Evidence from an Ethnic Studies Curriculum." *American Educational Research Journal* 54 (1): 127—66.

Delaware Department of Education. 2011. Dual Language Immersion Education in Delaware. https://www.doe.k12.de.us/immersion.

Dynarski, S. 2003. "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion." *The American Economic Review* 93 (1): 279—88.

Fabián Romero, E. 2017. With New Research, Policy Shifts, Bilingual Education on Rise. Education Writer's Association Blog: Latino Ed Beat. https://www.ewa.org/blog-latino-ed-beat/new-research-policy-shifts-bilingual-education-rise.

Fan, S. P., Z. Liberman, B. Keysar, and K. D. Kinzler. 2016. "The Exposure Advantage: Early Exposure to a Multilingual Environment Promotes Effective Communication." *Psychological Science* 26 (7): 1090—7.

Fortune, T. W. 2012. "What the Research Says about Immersion." In *Chinese Language Learning in the Early Grades: A Handbook of Resources and Best Practices for Mandarin Immersion*, 9—14. New York: Asia Society.

Goodman-Bacon, A. 2021. "Difference-in-differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254—77.

Greenberg, A., B. Bellana, and E. Bialystok. 2013. "Perspective-taking Ability in Bilingual Children: Extending Advantages in Executive Control to Spatial Reasoning." *Cognitive Development* 28 (1): 41—50.

Imai, K., and I. S. Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29: 405—15.

Kamenetz, A. 2016. *6 Potential Brain Benefits of Bilingual Education*. National Public Radio. http://ripr.org/post/6-potential-brain-benefits-bilingual-education.

Keshavarz, M. H., and H. Astaneh. 2004. "The Impact of Bilinguality on the Learning of English Vocabulary as a Foreign Language (L3)." *Bilingual Education and Bilingualism* 7: 295—302.

Kropko, J., and R. Kubinec. 2020. "Interpretation and Identification of Within-Unit and Cross-Sectional Variation in Panel Data Models." *PLoS One* 15: 1—22.

Krueger, A. B., and P. Zhu. 2004. "Another Look at the New York City School Voucher Experiment." *American Behavioral Scientist* 47 (5): 658—98.

Kuziemko, I. 2014. "Human Capital Spillovers in Families: Do Parents Learn from or Lean on Their Children?" *Journal of Labor Economics* 32 (4): 755—86.

Ladson-Billings, G. 1992. "Liberatory Consequences of Literacy: A Case of Culturally Relevant Instruction for African American Students." *The Journal of Negro Education* 61 (3): 378—91.

Lam, K., and E. Richards. 2020. *More US Schools Teach in English and Spanish, but Not Enough to Help Latino Kids*. USA Today. https://www.usatoday.com/in-depth/news/education/2020/01/06/english-language-learners-benefit-from-dual-language-immersion-bilingual-education/4058632002/.

Lambert, W. E., F. Genesee, N. Holobow, and L. Chartrand. 1993. "Bilingual Education for Majority English-speaking Children." *European Journal of Psychology of Education* 8 (1): 3—22.

Lambert, W. E., G. R. Tucker, and A. d'Anglejan. 1973. "Cognitive and Attitudinal Consequences of Bilingual Schooling: The St. Lambert Project through Grade Five." *Journal of Educational Psychology* 65 (2): 141—59.

Lapkin, S., D. Hart, and M. Turnbull. 2003. "Grade 6 French Immersion Students' Performance on Large-Scale Reading, Writing, and Mathematics Tests: Building Explanations." *Alberta Journal of Educational Research* 49 (1): 6—23.

Lindholm-Leary, K. J., and N. Block. 2010. "Achievement in Predominantly Low SES/Hispanic Dual Language Schools." *International Journal of Bilingual Education and Bilingualism* 13 (1): 43—60.

Lyster, R. 2007. *Learning and Teaching Languages through Content: A Counterbalanced Approach*. Amsterdam: John Benjamins.

Marian, V., A. Shook, and S. R. Schroeder. 2013. "Bilingual Two-Way Immersion Programs Benefit Academic Achievement." *Bilingual Research Journal* 36: 167—86.

Maxwell, L. A. 2014. *Successes Spur Push for Dual-Language Classes*, 14—5. Education Week. https://www.edweek.org/ew/articles/2014/10/15/08dual.h34.html.

Met, M. 1994. "Teaching Content through a Second Language." In *Educating Second Language Children: The Whole Child, the Whole Curriculum, the Whole Community*, edited by F. Genesee, 159—82. New York: Cambridge University Press.

Moll, L. C., and N. González. 1994. "Lessons from Research with Language-Minority Children." *Journal of Reading Behavior* 26 (4): 439—56.

Mora, J. K. 2009. "From the Ballot Box to the Classroom." *Educational Leadership* 66 (7): 14—9.

North Carolina Department of Public Instruction. 2020. Dual Language Immersion. https://www.dpi.nc.gov/districts-schools/classroom-resources/k-12-standards-curriculum-and-instruction/programs-and-initiatives/dual-language-immersion.

Oster, E. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187—204.

Padilla, A. M., K. J. Lindholm, A. Chen, R. Durán, K. Hakuta, W. E. Lambert, and G. R. Tucker. 1991. *The English-only Movement–Myths, Reality, and Implications for Psychology*. Washington: American Psychological Association.

Paris, D., and H. S. Alim. 2014. "What Are We Seeking to Sustain through Culturally Sustaining Pedagogy? A Loving Critique Forward." *Harvard Educational Review* 84 (1): 85—100.

Potowski, K. 2004. "Student Spanish Use and Investment in a Dual Immersion Classroom: Implications for Second Language Acquisition and Heritage Language Maintenance." *The Modern Language Journal* 88 (1): 75—101.

Saiz, A., and E. Zoido. 2005. "Listening to what the World Says: Bilingualism and Earnings in the United States." *The Review of Economics and Statistics* 87 (3): 523—38.

Shadish, W. R., M. H. Clark, and P. M. Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment." *Journal of the American Statistical Association* 103 (484): 1334—56.

Sleeter, C. E. 2012. “Confronting the Marginalization of Culturally Responsive Pedagogy.” *Urban Education* 47 (3): 562—84.

Small, D. S., and P. R. Rosenbaum. 2008. “War and Wages: The Strength of Instrumental Variables and their Sensitivity to Unobserved Biases.” *Journal of the American Statistical Association* 103 (483): 924—33.

Steele, J. L., R. J. Murnane, and J. B. Willett. 2010. “Do Financial Incentives Help Low-Performing Schools Attract and Keep Academically Talented Teachers? Evidence from California.” *Journal of Policy Analysis and Management* 29 (3): 451—78.

Steele, J. L., R. O. Slater, G. Zamarro, T. Miller, J. Li, S. Burkhauser, and M. Bacon. 2017. “Effects of Dual-Language Immersion Programs on Student Achievement: Evidence from Lottery Data.” *American Educational Research Journal* 54 (1): 282S—306S.

Steele, J. L., R. O. Slater, J. Li, G. Zamarro, T. Miller, and M. Bacon. 2018. “Dual-language Immersion Education at Scale: An Analysis of Program Costs, Mechanisms, and Moderators.” *Educational Evaluation and Policy Analysis* 40 (3): 420—45.

Umansky, I. M., and S. F. Reardon. 2014. “Reclassification Patterns Among Latino English Learner Students in Bilingual, Dual Immersion, and English Immersion Classrooms.” *American Educational Research Journal* 51 (5): 879—912.

UNESCO. 2016a. *Education 2030: Incheon Declaration and Framework for Action for the Implementation of Sustainable Development Goal 4*. Retrieved from New York.

UNESCO. 2016b. *If You Don’t Understand, How Can You Learn? (Global Education Monitoring Report, Policy Paper 24)*. Retrieved from New York.

Utah Senate. 2016. Dual Language Immersion: Origin Story. http://senatesite.com/utahsenate/dual-language-immersion/.

Utah State Board of Education. 2020. Welcome to Kindergarten: Frequently Asked Questions. https://www.schools.utah.gov/curr/kindergarten?mid&tnqx3d;1179&tid&tnqx3d;3.

Valentino, R. A., and S. F. Reardon. 2015. “Effectiveness of Four Instructional Programs Designed to Serve English Learners: Variations by Ethnicity and Initial English Proficiency.” *Educational Evaluation and Policy Analysis* 37 (4): 612—37.

Vega-Bayo, A., and P. Mariel. 2022. “Parents’ Willingness to Pay for Bilingualism: Evidence from Spain.” *Journal of Family and Economic Issues*. https://doi.org/10.1007/s10834-022-09852-1.

Watzinger-Tharp, J., F. Rubio, and D. S. Tharp. 2018. “Linguistic Performance of Dual Language Immersion Students.” *Foreign Language Annals* 51 (3): 575—95.

Watzinger-Tharp, J., K. Swenson, and Z. Mayne. 2016. “Academic Achievement of Students in Dual Language Immersion.” *International Journal of Bilingual Education and Bilingualism* 21 (8): 913—28.

Watzinger-Tharp, J., D. S. Tharp, and F. Rubio. 2021. “Sustaining Dual Language Immersion: Partner Language Outcomes in a State-wide Program.” *The Modern Language Journal* 105 (1): 194—217.

Williams, C. 2017. *The Intrusion of White Families into Bilingual Schools*. The Atlantic.

Xu, D., and S. S. Jaggars. 2013. “The Impact of Online Learning on Students’ Course Outcomes: Evidence from a Large Community and Technical College System.” *Economics of Education Review* 37: 46—57.

Xu, X., A. M. Padilla, and D. M. Silva. 2015. “Learner Performance in Mandarin Immersion and High School World Language Programs: A Comparison.” *Foreign Language Annals* 48 (1): 26—38.

Yuki, K. 2022. “Is Bilingual Education Desirable in Multilingual Countries?” *The B.E. Journal of Economic Analysis & Policy* 22 (4): 889—949.