ACHIEVEMENT EFFECTS OF DUAL LANGUAGE IMMERSION IN ONE-WAY AND TWO-WAY PROGRAMS: EVIDENCE FROM A STATEWIDE EXPANSION

April 2023

Jennifer L. Steele, Associate Professor of Education and Faculty Affiliate in the Dept. of Public Administration and Policy, American University, steele@american.edu

Johanna Watzinger-Tharp, Associate Professor of Linguistics, University of Utah, j.tharp@utah.edu

Robert O. Slater, Co-Director of the American Councils Research Center, American Councils for International Education, rslater@americancouncils.org

Gregg Roberts, Director of Dual Language Studies, American Councils for International Education, groberts@americancouncils.org

Karl Bowman, Dual Language Immersion and World Language Specialist, Utah State Board of Education, Karl.Bowman@schools.utah.gov

RUNNING HEAD: ACHIEVEMENT EFFECTS OF DUAL LANGUAGE IMMERSION

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, especially when demand exceeds supply. 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) and those comprising one-to-two thirds of the latter (two-way). Adjusting for school and cohort fixed effects, intent-to-treat estimates in one-way programs are null, but those in two-way programs reach 0.05 to 0.07 standard deviations in English, math, and science across grades 3-6. Benefits appear stronger, especially in math, for ELs whose primary (home) language matches the partner language. Our results highlight the special value of two-way DLI for primary speakers of the partner language.

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 & 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 world language proficiency 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 alike 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

[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.

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 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.

[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.

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, our school fixed-effects approach—essentially a difference-in-differences with staggered treatment timing—allows us to estimate plausibly causal effects of a statewide DLI program at scale. Using 17 cohorts of public school students across an entire state, we are able to 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. By examining ways in which one-way and two-way programs differ, including their instructional languages and student demographics, we shed new light on reasons that program effects may differ. The fact that Utah's DLI instructional model is constant across the state, with common curricula, teacher professional development, and instructional schedules (providing 50% of instruction in each language) allows us to explore partner language selection and student demographic composition as possible drivers of effect heterogeneity. We use the Bacon decomposition procedure and other robustness tests to ensure that our estimates not sensitive to high-variance timing groups outliers, effects of time-varying controls, or schools with non-parallel pre-treatment trends (Goodman-Bacon, 2021).

Our analysis finds null intent-to-treat (ITT) effects of one-way program access on student achievement in core content areas, and moderately positive effects of two-way programs in

English language arts (ELA), math, and science, ranging from 0.05 to 0.07 standard deviations averaged across grades 3-6. These findings are robust to a variety of sensitivity tests. Student heterogeneity tests show that two-way effects, especially in math, are larger among ELs whose primary language matches the partner language, and that as of grade 5, these language-matched ELs reach English proficiency at higher rates than their peers who lack DLI access. Because DLI-access effects seem to rise with the share of students in the school whose primary language matches the school's (current or eventual) partner language, they suggest that cultural and linguistic adjacency may enhance the benefits of DLI programs.

In this article, we briefly summarize the existing literature on DLI program effects and what this study adds. We then describe Utah's DLI program and our dataset. Next, we present our analytic approach, followed by descriptive, IV, and ITT results, robustness tests, heterogeneity tests, and explorations of reasons that one-way and two-way program effects may differ. 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 & Richards, 2020). Though this figure remains modest, it represents a five-fold increase from nine years ago, 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 & 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 &

Zoido, 2005), to 4-6% percent for French in Anglophone Canada (Christofides & 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, & 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 & Di Paolo, 2018; Vega-Bayo & 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 & Craik, 2010), though these laboratory studies are generally descriptive and not causal. Bilingualism has been linked to metalinguistic awareness (Cenoz, 2003; Keshavarz & Astaneh, 2004) and to children's social perceptive-taking (Fan, Liberman, Keysar, & Kinzler, 2016; Greenberg, Bellana, & 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 & Swain, 1978; Caldas & Boudreaux, 1999; Lapkin, Hart, & Turnbull, 2003; Marian, Shook, & Schroeder, 2013), and some studies used baseline matching on pre-interevention characteristics (Lambert, Genesee, Holobow, & Chartrand, 1993; Lambert, Tucker, & d'Anglejan, 1973). More recently, Authors (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 & Thomas, 2004; De Jong, 2004; Lindholm-Leary & 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 5,400 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 4,000 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 1,600 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 idea, 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, Milem, Jaquette, & Marx, 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, 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 a sample of about 49,000 unique students in ever-DLI schools (a quarter in ever-two-way 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.

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 $5,000 per year in program maintenance thereafter, though expenditures reportedly represented only an additional 1% of per-pupil funding across DLI schools. The funding was designed to incentivize DLI program launches across the state, but decisions about launching 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 typically 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 (Authors, 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 the extent to which the cultural heritages of students with primary languages other than English are respected within the school.

3.2 Analytic Sample

Our study uses an administrative dataset provided by the Utah State Board of Education that includes all public school students in the state of Utah for the entering kindergarten cohorts of 2001-2002 through 2017-2018. We restrict our analysis to a balanced panel of schools that have data for all 17 cohorts, observing ELA, math, and science test score outcomes from grades 3 through 6. We exclude charter schools from the analysis because they were not part of the state's DLI scale-up policies. Because our analytic strategy compares cohorts of students who attended eventual-DLI schools in years before and after the schools' launch of DLI programs, we focus on 33,476 unique students who attended ever-one-way DLI schools, and 15,340 students who attended ever-two-way DLI schools in grade 6 or below. We treat the students' first observed year in a Utah public school as his or her base year. The base year represents kindergarten for 59% of the sample, and first grade for 9%. Table 1 presents descriptive statistics for members of the analytic sample in their base observation year, as well as for 171,975 unique students who attended Utah public schools that never launched DLI programs.

<Insert Table 1 about here>

Table 1 shows that white, non-Hispanic students constituted 76% of students in never-DLI schools and 85% in ever one-way schools, but only 50% in ever two-way schools. Still, 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 free/reduced-price lunch (FRL) eligibility than their peers in ever one-way or never-DLI schools, at 57% versus 27% and 38%, respectively, and much higher rates of ever-EL enrollments, at 36% versus 8% and 14%, respectively. Students who attended ever two-way schools also lived in moderately less-educated zip codes than those in ever one-way schools, and in neighborhoods with higher shares of Limited English Proficiency and eligibility for the federal Supplemental Nutrition Assistance Program (SNAP).

Table 1 also shows the distribution of the ITT variables (dichotomous slot availability 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. Among students who attended ever one-way schools, about 76% attended schools that eventually offered Spanish or Mandarin Chinese, whereas French, Portuguese, and German programs accounted for a smaller share. 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 all test scores 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 about a tenth of an 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.

4. Econometric Strategy

Our analysis begins with a purely descriptive comparison of academic performance between DLI and non-DLI students in a given year, net of a rich set of baseline characteristics, as shown in equation 1.

$$y_{ics} = \alpha_1 + \beta_1 DLI_{ics} + \boldsymbol{\lambda_1}'\mathbf{C_c} + \boldsymbol{\delta_1}'\mathbf{S_s} + \boldsymbol{\varphi_1}'\mathbf{X_{ics}} + \boldsymbol{\eta_1}'\mathbf{K_{cs}} + \varepsilon_{1ics} \qquad (1)$$

Using ordinary least squares (OLS) regression, equation 1 estimates the relative performance, y_{ics}, of student *i* in cohort *c* from baseline school *s*, as a function of whether the student is currently enrolled in DLI. The average difference in y_{icst} between DLI and non-DLI students in the same school and cohort is given by β_1, holding constant vectors of fixed effects for kindergarten cohort ($\boldsymbol{\lambda_1}$), initial school ($\boldsymbol{\delta_1}$), and baseline student characteristics $\mathbf{X_{ics}}$, which include gender, race/ethnicity, subsidized meal eligibility at baseline, whether the student was ever classified as an EL, having a home language other than English (regardless of EL classification), special education status at baseline, and migrant status at baseline. Vector $\mathbf{K_{cst}}$ captures school-by-cohort attributes of baseline school *s* for cohort *c*, including the percent who are white, subsidized-meal eligible at baseline, ever classified as EL, and special education eligible at baseline. The dependent variable y_{ics} is a test score in ELA, math, or science for student I, standardized statewide by subject, grade, and year to mean 0 and SD 1. The error term is given by ε_{1ics}. We cluster standard errors at the base school level. 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.

Because virtually all schools in Utah that offer DLI programs also offer parallel non-DLI strands, students attending a DLI school must choose whether to take part in the DLI program at his or her school. This may happen through a choice process at the student's residentially zoned school or by the family's application to a DLI program, depending on district policy. Evidence suggests that families who do and do not take advantage of DLI programs may differ in terms of both observed and unobserved attributes (Lindholm-Leary & Block, 2010; Marian et al., 2013; Steele et al., 2017), including education levels and values, perceptions of their children's motivations and aptitudes, and access to information about local dual-language options. To avoid confounding by unobserved differences between DLI and non-DLI students in the same schools and cohorts, we employ a dichotomous ITT indicator of whether the school offered DLI slots in the student's first school when the student was in first grade. In this way, we leverage variation in first graders' access to DLI within schools over time to estimate the ITT effect of a school's launch of DLI on subsequent achievement in the school. In some specifications, we instead employ a continuous variable—the fraction of first grade DLI slots per pupil (SPP) in the student's first grade year, where the average in treated schools was 0.52, ranging from 0.29 to 0.99. For simplicity of interpretation and robustness checks, however, our preferred specification makes the ITT indicator dichotomous—slots offered or not—at the school-by-cohort level, but we compare these estimates to the SPP indicator in some cases. . In a school-by-cohort-level analysis, the ITT effect of first-grade DLI availability for cohort *c* in base school *s* (ITT_{sc}) is estimated as follows:

$$y_{cs} = \alpha_2 + \beta_2 ITT_{sc} + \boldsymbol{\lambda}_2{}'\mathbf{C}_c + \boldsymbol{\delta}_2{}'\mathbf{S}_s + \boldsymbol{\varphi}_2{}'\mathbf{X}_{cs} + \boldsymbol{\eta}_2{}'\mathbf{K}_{ic} + \varepsilon_{2cs} \qquad (2)$$

where coefficient β_2 is the average difference in y_{cs} associated with a school offering DLI slots to a given cohort.. We refer to this as an ITT variable because it indicates students' access to DLI slots as a function of their base school and cohort year, but it does not indicate their take-up of such slots, which may be vulnerable to selection on unobservable attributes.

Other terms in the model are interpreted as in equation 1. Unlike equation 1, which pertains only to the 2016-17 and 2017-2018 academic years, equation 2 is estimated for academic years 2001-2002 through 2017-2018, in schools that were observed for all 17 years.

To identify ITT effects of DLI access on school achievement in equation 2, we must assume that successive cohorts of students who began their educations in a given Utah school were, on average, the same from year to year, net of their observed baseline characteristics. If the population of students who enrolled in a school systematically changed in response to the availability of a DLI program, and the changes on unobserved attributes are associated with the outcomes of interest and uncorrelated with school-by-year controls , they could bias our estimates of the DLI-access effect. Because we are essentially employing a two-way fixed effects model with time-varying introduction of DLI access, we must also be concerned about non-parallel secular trends and control-variable effects in the shifting pool of treated versus un-treated schools in a given year (Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfœuille, 2020). To assess the extent to which our estimates reflect differences in the composition of treated and comparison units for each kindergarten cohort *c*, and the extent to which it may be driven by controls for time-varying attributes of each school, we use a Bacon decomposition test (Goodman-Bacon, 2021; Goodman-Bacon, Goldring, & Nichols, 2019), which indicates treatment-effect estimates and weights for within-school changes in DLI access, changes in school-by-cohort controls due to demographic shifts, and differences in time trends for never-treated versus ever-treated schools. In practice, we omit never-treated schools and in eventually-treated schools whose DLI-eligible cohorts were too young for the states' third-grade ELS and mathematics assessments.

Separately, we leverage the ITT indicator as an instrument for students' observed DLI enrollment, though clean DLI enrollment data are available only in the final two years of the dataset. In addition, we conduct robustness tests for families' selection int DLI schools beyond the first implementation year, and for extant secular trends or spillover effects beyond the treated cohorts using a placebo test. We investigate heterogenous ITT effects for ELs versus non-ELs in one-way and two-way programs using a student-level analysis disaggregated by grade. Finally, we investigate plausible mechanisms for differential effects between one-way and two-way DLI programs.

5. Results

5.1 Descriptive Within-Cohort Estimates

We begin with descriptive results from the within-cohort OLS regression models described in equation 1. As noted, these estimates pertain only to 2016-2017 and 2017-2018 data because those are the years in which reliable DLI enrollment indicators became available statewide. In Table 2, we present ELA estimates for three different model specifications to assess their sensitivity to controls for school-by-year demographic attributes and students' individual baseline attributes, respectively (Shadish, Clark, & Steiner, 2008). We find in columns 2 and 7 that controlling for school-by-year demographics has little effect on the estimates. This is consistent with a school-by-cohort selection test in Appendix Table 1 showing that neither the launch of a DLI program (the dichotomous ITT indicator) nor the first grade DLI slots per pupil for a given school and cohort (a granular ITT measure) strongly predicts demographic changes in school-by-cohort attributes.[3] However, controlling for students' baseline individual attributes in columns 3 and 8 of Table 2 reduces one-way program effect estimates from about 0.3 to 0.24 standard deviations and increases two-way program effect estimates from about 0.25 to 0.3 standard deviations. These patterns may reflect the fact that two-way programs typically serve a higher concentration of students from low-income, non-white, and EL backgrounds, as shown in Table 1.

OLS estimates for math and science scores include school-by-year and individual-level controls and are similar in magnitude to ELA estimates. Even with a full set of control variables, estimates of 0.25 standard deviations are large in magnitude and raise questions about possible selection bias. Still, they show that students enrolled in DLI programs substantially outperformed their non-DLI peers from the same base schools and cohorts in all three content areas.

<Insert Table 2 about here>

5.2 Instrumental Variables Estimates of Local Average Treatment Effects

In Table 3, we make further use of the plausibly endogenous regressor in Table 2: the student's DLI enrollment status in a given year. We use this as the first-stage dependent variable in a two-stage least squares model in which the arguably exogenous school-by-cohort-by-year ITT variable (DLI slots offered *or* DLI slots per first-grade pupil) serves as an instrumental variable. Insofar as the instrument strongly and monotonically predicts DLI enrollment and influences student achievement *only through* its effect on DLI enrollment, then it can be used to

[3] In prior versions estimated at the student level using school-by-cohort-by-grade controls, grade-level demographics showed more responsiveness than school-by-cohort demographics, with patterns that, if uncontrolled, would be expected to bias estimates toward zero due to modest selection of affluent, white, English-speaking families toward one-way and away from two-way programs.

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 & 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. The first-stage estimates for both the dichotomous and granular treatment indicators show that DLI enrollment is strongly predicted by both instruments, with F-statistics well above 10 (Bound, Jaeger, & Baker, 1995).

The question is whether the school-by-cohort variation in treatment access may also affect unobserved school attributes in ways that are confounded with rather than resulting from DLI enrollment, such as the systematic sorting of DLI-interested families into schools-by-cohorts that provide DLI access. As noted, we find limited evidence of sorting on observables in Table A1, and evidence in Table 2 that even within-cohort sorting on observables exerts only a modest effect on outcomes (Oster, 2019). This lends confidence that the exclusion restriction is plausible, especially conditional on school-by-cohort and individual controls (Altonji, Elder, & 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.

<Insert Table 3 about here>

In Table 3, we estimate that DLI enrollment may increase concurrent ELA test scores by 0.196 SD ($p<0.1$) using the more granular slots-per-pupil instrument, a plausible effect that is about 20% smaller than the OLS enrollment-effect estimate. We note that the magnitude of estimates for two-way programs are larger and more positive than those of one-way programs in all three subject areas, though most estimates do not reach statistical significance.

5.3 Intent-to-Treat Estimates of DLI Access Effects

To leverage the full panel of school-by-year cohorts from the 2001-2002 through 2017-2018 academic years, we turn now to estimates of the dichotomous school-by-cohort treatment effect on student achievement. Table 4 summarizes key ITT estimates in ELA, math, and science, pooled across grade levels, using school and cohort fixed-effects models and a Bacon decomposition analysis of the ITT estimates. Bacon decomposition results are estimated with the *bacondecomp* routine in Stata (Goodman-Bacon et al., 2019). Schools in this and all the school-by-cohort-level analyses are weighted by their 2002 school enrollments using analytic weights that do not artificially inflate statistical power.

<Insert Table 4 about here>

Substantively, access to one-way DLI programs yields no effect on aggregate achievement in ELA, math, or science, whereas access to two-way programs yields statistically significant higher achievement of 0.052 standard deviations in ELA, 0.071 in mathematics, and 0.064 in science. Since most DLI schools offer only about half of their slots as immersion slots, we might expect treatment-on-the-treated estimates to be larger, as suggested imprecisely for compliers in the IV analysis in Table 3. Bacon decomposition estimates are identical in magnitude to the fixed effect estimates, but are additionally weighted according to the variance in each school-by-cohort comparison block, yielding more precise estimates in ELA and math. The Bacon decomposition procedure allows us to establish that 93% to 98% of the ITT estimates are driven by within-school, pre/post treatment comparisons.[4] The models also include time-varying school-by-cohort covariates across time periods, which contribute only 3% to 5% of the treatment effect estimates. In science, the lack of testing until fourth grade results in a "never-treated" (here, eventually-treated) school comparison that yields an effect size comparable to the timing-group estimates.

5.4 Robustness Tests on ITT School-by-Cohort Estimates

Table 5 provides results from a series of additional robustness tests for the ITT estimates in Table 4. Expecting that families in the first DLI-eligible cohorts in a given school would have had less time to respond to a newly launched program, we run a robustness test in which, like Anghel et al. (2016) in Spain, we limit the ITT group to just the first DLI-eligible treatment group 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.

<Insert Table 5 about here>

The first-cohort results in Panel A of Table 5 are similar to those in the main analysis in Table 4, though the two-way program estimates for science are no longer significant. In Panel B, we focus on the first four ITT cohorts in a given school (out of nine that could have been treated in 2009-2010 through 2017-2018), to examine whether a school's implementation duration predicts outcomes. We find results that are again similar in magnitude, direction, and significance to those for the full panel in Table 4. In the next two panels, C and D, we focus on early-adopting (fall 2009 -fall 2012) versus late-adopting (fall 2013 – fall 2017) DLI schools in case early-adopting schools face different (as-yet-untreated) trends or are differently prepared than late adopters. Here, we find evidence of DLI achievement effects for ELA of 0.09 among late adopters ($p<0.1$), and insignificant effects otherwise. Finally, in Panel E, we conduct a placebo first-cohort analysis in which we designate the first treated cohort as being one year earlier than it actually was, and we limit the analyses to that cohort and all previous ones, similar to the first-cohort analysis. The placebo test examines the possibility that DLI effects were due to pre-existing capacity-building in treated schools that may or may not be linked to DLI planning. It could also be seen as testing possible whole-school spillover effects of state funding for DLI, since the funding could possibly be applied toward school-wide initiatives that benefit older grades that were ineligible for DLI. Estimates in the placebo test do suggest the existence of a possible pre-treatment improvement trend in two-way programs in math, where we find an estimate of 0.06 SD ($p<0.1$). Attributing this estimate to funding spillovers on older grades seems somewhat less plausible, since additional funding levels were low in per-pupil terms (about 1% across DLI schools), and since we might expect spillover effects of DLI funding to appear in more than one subject. In sum, these placebo test results do not suggest that the ITT estimates are not real, but that schools' capacity-building preceding the DLI launch, or their spreading of resources to pre-treated cohorts, could have played a contributing role, particularly in math.

5.5 Student-Level Estimates: Heterogeneity by EL Status and Grade

To understand heterogeneity of program effects for English learners versus primary English speakers, we continue with the ITT estimation strategy but expand it to a student-level analysis conducted separately by grade-level, adjusting for school-by-grade-by-year and time-invariant individual covariates. Results are presented in Table 6, where Panel A focuses on primary English speakers never classified as ELs, and Panel B focuses on ELs whose primary language (Spanish in all two-way programs and many one-way programs) matches that of the (current or eventual) partner language in the students' initial school. ITT at the school-by-cohort level is still defined as dichotomous.

<Insert Table 6 about here>

Here we find mostly null estimates across cells, though with two-way programs, we do find marginally significant estimates for primary English speakers in grade 4 ELA (0.063 SD) and grade 5 science (-0.079 SD) English speakers in ELs, noting that they are positive and negative, respectively. For language-matched ELs, however, we find substantially positive treatment-effect estimates in two-way programs ranging from 11% to 19% of a standard deviation in math. Why we see these effects for language-matched ELs in mathematics instead of ELA is not clear, except that mathematics is taught in the partner language in DLI elementary schools in Utah. Thus, a school's conversion to DLI may offer language-accessible, grade-level appropriate mathematics instruction for primary Spanish speakers more rapidly than they would otherwise receive.

As context, additional robustness tests for Table 6 (not shown) using the granular slots-per-pupil ITT variable instead of the dichotomous DLI availability variable yield statistically significant ($p<0.05$) two-way program benefits for language-matched ELs of 19% of a standard deviation in ELA and 18% of a standard deviation in math, and of 18% of a standard deviation in ELA ($p<0.05$) and science ($p<0.01$) for primary English speakers. These can be interpreted as predicted school-by-cohort effects of a two-way school shifting from 0% to 100% of first-grade slots as DLI.

<Insert Table 7 about here>

Table 7 focuses on English-language proficiency trajectories for language-matched ELs. It presents the effects of the dichotomous DLI ITT variable on the probability that a student ever classified as EL in Utah public schools remains classified as such in grades 1 through 6. For one-way programs, shown on the left, we find no statistically significant differences in rates of EL classification. In other words, ELs appear to be reclassified at similar rates before and after the launch of one-way DLI programs in their base schools, regardless of whether their primary language matches the partner language. For two-way programs, we find no differences in EL classification rates for ELs whose primary languages *do not* match the partner language (Spanish). For those whose home languages *do* match, we find similar rates of EL status persistence until grades 5, at which time students ever classified as EL are about 6 percentage points less likely to be classified EL ($p<0.05$). Estimates using the granular slots per pupil ITT measure (not shown) show larger two-way language-match effects in grades 4 through 6 by about one percentage point, but their statistical significance is consistent with Table 7. The finding that language-aligned DLI access increases English proficiency rates after several years of exposure is consistent with Steele et al. (2017), who find proficiency reclassification benefits beginning in grade 6, and with Umansky and Reardon (2014), who find it from grade 7 onward.

5.6 Interpreting One-Way versus Two-way Program Estimates

In Utah, both one-way and two-way programs use the same 50/50 instructional model. Programs also receive common dual-language curriculum and teacher professional development. From that perspective, we would expect similar average achievement effects in both program types, but estimates appear null in one-way programs and null-to-positive in two-way programs. We consider two possible explanations. One is that all two-way programs are Spanish programs, whereas one-way programs comprise Spanish, Mandarin, French, and German. Because Spanish is arguably the most phonetically accessible and English-adjacent of the partner languages, it is possible that the two-way effects are actually Spanish effects.

Another possibility is that one-way schools in Utah are more affluent and white than their two-way counterparts. To address whether program differences are driven by the baseline demographic attributes of the schools, we weight the one-way programs by their similarity to two-way schools in their pre-DLI launch year. We estimate a logistic regression model predicting whether a DLI school is two-way on the basis of the pre-DLI percentage of students who are white, subsidized-meal eligible, ever-ELs, and special-education eligible. Following Austin (2011), we calculate average treatment on the treated (ATT) weights as in equation 3:

$$w_{cs.ATT} = Z_s + \frac{(1-Z_s)p_{cs}}{1-p_{cs}} \qquad (3)$$

where Z_s is 1 if the ever-DLI school was two-way, and 0 if it was one-way, and where the fitted probability of the school being two-way school is p_{cs}. Applying these weights improves balance on pre-DLI school characteristics by about two-thirds.

<Insert Table 8 about here>

Table 8 shows that one-way program estimates remain null and close to 0 for all three subjects and with both restrictions—Spanish-only on the left and demographically weighted on the right. This suggests that the differences between two-way and one-way estimates in the main analysis are not Spanish-language effects, nor are they driven by demographic differences between one-way and two-way programs.

Finally, we consider the extent to which the ITT effect of DLI availability differs by the fraction of students in the school whose primary languages matches the partner language. The analysis includes all schools that eventually opened one-way or two-way programs, and it defines the language match in pre-treatment years based on the language that eventually became the partner language in the school. At the school-by-cohort level, we interact the dichotomous DLI access indicator with the fraction of language-match students in the school in the pre-DLI year, ranging from 0 to 0.83. (The average in ever-two-way schools is about 0.34.) In Table 9, we report the main effect of the dichotomous DLI access indicator and the interaction effect, which is interpreted as the differential effect of DLI access for each unit difference (here, from 0 to 1) in the fraction of language-match students in the last pre-DLI year.

<Insert Table 9 about here>

Table 9 shows a null main effect of DLI access on all three content areas. However, we find that moving from 0% to 100% percent of language-matched students in the school (not solely ELs, and perhaps not all in DLI) would increase predicted math achievement in the school by 0.193 standard deviations on average ($p<0.05$). The interaction terms also predict within-sample ITT benefits of fully-language matched schools of 0.058 and 0.154 in ELA and science, respectively, though these estimates do not approach statistical significance.

It is not clear why the interaction effect is larger in mathematics than in ELA and science, but this is consistent with the student-level benefits we find for language-matched ELs in two-way programs in mathematics in Table 6. In a study instrumenting English proficiency by immigrant children's age of arrival in the U.S., Aparicio Fenoll (2018) finds that math scores do not depend on exogenous variation in English proficiency. But it may still be the case that the ability to access mathematics instruction in their primary language throughout elementary school hastens children's acquisition of math skills and taps as strengths the primary language skills that might be marginalized or stigmatized in a non-DLI school context.

6. Discussion and Conclusion

As demand grows for public school programs that are both culturally inclusive and academically challenging, DLI programs show clear 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 & 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 state scale-up effort. Though Utah's DLI students spent half of their elementary instructional hours learning in a language other than English, we find no evidence of academic harm for one-way programs and modest evidence of academic benefit for two-way programs, especially for students whose primary language matches the partner language. Cross-grade estimates from school-by-cohort regressions and Bacon decomposition procedures are 5%, 7%, and 6% of a standard deviation in ELA, math, and science, respectively, and are robust to tests for selection and family sorting over time. Also, by grade 5, Spanish-speaking ELs with access to two-way DLI Spanish programs had English proficiency rates that were 6 percentage points higher than their counterparts without DLI access. Though a pre-treatment-year placebo test suggests that pre-program trends or spillover of funding to older grades could have played a role in the mathematics estimates, the data are generally consistent across models in showing modest ITT benefits of two-way programs Our IV analysis also finds a plausible local average treatment effect (reflecting actual enrollment) of 0.196 in ELA for DLI students whose enrollment was regulated by the school and cohort in which they landed. Moreover, the strongest mathematics effects are found for language-matched ELs in two-way programs across most elementary grades—a group whose families, on average, may be less prepared to navigate systems of residential zoning and school choice, and whose access to DLI is particularly important from an equity perspective (Lam & Richards, 2020).

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, Authors (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, & 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 (Authors, 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.

Our findings of stronger benefits in two-way programs for language-matched ELs comport with evidence about the academic benefits of culturally relevant instruction (Cabrera et al., 2014; Dee & Penner, 2016) and suggest a need to better understand language and cultural practices in these schools. Schools that offer two-way DLI 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 feasible in communities that serve students from diverse language backgrounds or from mostly English-speaking backgrounds. Future research should examine implementation differences by program type, and the extent to which effects covary with school cultural norms, parent communication practices, racial/ethnic alignment of teachers and students, and linguistically accessible content. In the interim, our study may be seen as reflecting the entwined nature of language and culture, and the complex ways in which they may reinforce one another.

Bibliography

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy, 113*(1), 151-184.

Anghel, B., Cabrales, A., & Carro, J. M. (2016). Evaluating a bilingual education program in Spain: The impact beyond foreign language learning. *Economic Inquiry, 54*(2), 1202-1223. doi:10.1111/ecin.12305

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Aparicio Fenoll, A. (2018). English proficiency and mathematics test scores of immigrant children in the US. *Economics of Education Review, 64*, 102-113. doi:10.1016/J.ECONEDUREV.2018.04.003

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. doi:10.1080/00273171.2011.568786

Authors. (2016). Academic achievement of students in dual language immersion. *International Journal of Bilingual Education and Bilingualism, 21*(8), 913-928. doi:10.1080/13670050.2016.1214675

Authors. (2018). Linguistic performance of dual language immersion students. *Foreign Language Annals, 51*(3), 575-595. doi:10.1111/flan.12354

Authors. (2021). Sustaining dual language immersion: Partner language outcomes in a state-wide program. *Modern Language Journal, 105*(1), 194-217. doi:https://doi.org/10.1111/modl.12694

Barik, H. C., & Swain, M. (1978). Evalutation of a French immersion program: the Ottawa study through grade five. *Canadian Journal of Behavioural Science, 10*(3), 192-201. doi:10.1037/h0081548

Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Science, 65*(4), 229-235. doi:10.1037/a0025406

Bialystok, E., & Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science, 19*(1), 12-23. doi:10.1177/0963721409358571

Bibler, A. (2020). Dual language education and student achievement. *Education Finance and Policy,* (online first). doi:https://doi.org/10.1162/edfp_a_00320

Bound, J., Jaeger, D. A., & Baker, R. M. (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), 443-450. doi:10.1080/01621459.1995.10476536

Burkhauser, S., Steele, J. L., Li, J., Slater, R. O., Bacon, M., & Miller, T. (2016). Partner-language learning trajectories in dual-language immersion: Evidence from an urban district. *Foreign Language Annals, 49*(3), 415-433.

Cabrera, N. L., Milem, J. F., Jaquette, O., & Marx, R. W. (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), 1084-1118.

Caldas, S. J., & Boudreaux, N. (1999). Poverty, race, and foreign language immersion: Predictors of math and English language arts performance. *Learning Languages, 5*, 4-15.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics, 225*(2), 200-230. doi:https://doi.org/10.1016/j.jeconom.2020.12.001

Cappellari, L., & Di Paolo, A. (2018). Bilingual schooling and earnings: Evidence from a language-in-education reform. *Economics of Education Review, 64*, 90-101. doi:https://doi.org/10.1016/j.econedurev.2018.03.007

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. Retrieved from http://www.cal.org/resources/immersion/

Center for Applied Linguistics. (2011b). Directory of two-way bilingual immersion programs in the U.S. . Retrieved from www.cal.org/twi/directory/

Chin, A., Daysal, N. M., & Imberman, S. A. (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. doi:10.1016/J.JPUBECO.2013.08.008

Christofides, L. N., & Swidinsky, R. (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., & Thomas, W. P. (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., & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review, 110*(9), 2964-2996. doi:10.1257/aer.20181169

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., & Penner, E. K. (2016). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal, 54*(1), 127-166. doi:10.3102/0002831216677002

Delaware Department of Education. (2011). Dual language immersion education in Delaware. Retrieved from https://www.doe.k12.de.us/immersion

Fabián Romero, E. (2017, October 25). With new research, policy shifts, bilingual education on rise. *Education Writer's Association Blog: Latino Ed Beat*. Retrieved from https://www.ewa.org/blog-latino-ed-beat/new-research-policy-shifts-bilingual-education-rise

Fan, S. P., Liberman, Z., Keysar, B., & Kinzler, K. D. (2016). The exposure advantage: Early exposure to a multilingual environment promotes effective communication. *Psychological Science, 26*(7), 1090-1097. doi:10.1177/0956797615574699

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* (pp. 9-14). New York: Asia Society.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254-277. doi:https://doi.org/10.1016/j.jeconom.2021.03.014

Goodman-Bacon, A., Goldring, T., & Nichols, A. (2019). bacondecomp: Stata module for decomposing difference-in-differences estimation with variation in treatment timing. Retrieved from https://ideas.repec.org/c/boc/bocode/s458676.html

Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective-taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive development, 28*(1), 41-50. doi:10.1016/j.cogdev.2012.10.002

Kamenetz, A. (2016, November 29). 6 potential brain benefits of bilingual education. *National Public Radio*. Retrieved from http://ripr.org/post/6-potential-brain-benefits-bilingual-education

Keshavarz, M. H., & Astaneh, H. (2004). The impact of bilinguality on the learning of English vocabulary as a foreign language (L3). *Bilingual Education and Bilingualism, 7*, 295-302.

Kuziemko, I. (2014). Human Capital Spillovers in Families: Do Parents Learn from or Lean on Their Children? *Journal of Labor Economics, 32*(4), 755-786. doi:10.1086/677231

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), 378-391. doi:10.2307/2295255

Lam, K., & Richards, E. (2020). More US schools teach in English and Spanish, but not enough to help Latino kids. *USA Today*. Retrieved from 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., Genesee, F., Holobow, N., & Chartrand, L. (1993). Bilingual education for majority English-speaking children. *European Journal of Psychology of Education, 8*(1), 3-22. doi:10.1007/BF03172860

Lambert, W. E., Tucker, G. R., & d'Anglejan, A. (1973). Cognitive and attitudinal consequences of bilingual schooling: The St. Lambert Project through grade five. *Journal of Educational Psychology, 65*(2), 141-159.

Lapkin, S., Hart, D., & Turnbull, M. (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., & Block, N. (2010). Achievement in predominantly low SES/Hispanic dual language schools. *International Journal of Bilingual Education and Bilingualism, 13*(1), 43-60. doi:10.1080/13670050902777546

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam: John Benjamins.

Marian, V., Shook, A., & Schroeder, S. R. (2013). Bilingual two-way immersion programs benefit academic achievement. *Bilingual Research Journal, 36*, 167-186. doi:10.1080/15235882.2013.818075

Maxwell, L. A. (2014, October 15). Successes spur push for dual-language classes. *Education Week,* pp. 14-15. Retrieved from https://www.edweek.org/ew/articles/2014/10/15/08dual.h34.html

Met, M. (1994). Teaching content through a second language. In F. Genesee (Ed.), *Educating second language children: The whole child, the whole curriculum, the whole community* (pp. 159-182). New York: Cambridge University Press.

Moll, L. C., & González, N. (1994). Lessons from research with language-minority children. *Journal of Reading Behavior, 26*(4), 439-456. doi:10.1080/10862969409547862

Mora, J. K. (2009). From the ballot box to the classroom. *Educational Leadership, 66*(7), 14-19.

North Carolina Department of Public Instruction. (2020). Dual language immersion. Retrieved from 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 and Economic Statistics, 37*(2), 187-204.

Padilla, A. M., Lindholm, K. J., Chen, A., Durán, R., Hakuta, K., Lambert, W. E., & Tucker, G. R. (1991). *The English-only movement--Myths, reality, and implications for psychology*. Washington, DC: American Psychological Association.

Paris, D., & Alim, H. S. (2014). What are we seeking to sustain through culturally sustaining pedagogy? A loving critique forward. *Harvard Educational Review, 84*(1), 85-100. doi:10.17763/HAER.84.1.982L873K2HT16M77

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. doi:10.1111/j.0026-7902.2004.00219.x

Saiz, A., & Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States. *Review of Economics and Statistics, 87*(3), 523-538.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*(484), 1334-1356.

Sleeter, C. E. (2012). Confronting the marginalization of culturally responsive pedagogy. *Urban Education, 47*(3), 562-584. doi:10.1177/0042085911431472

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal, 54*(1), 282S-306S. doi:10.3102/0002831216634463

Umansky, I. M., & Reardon, S. F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal, 51*(5), 879-912. doi:10.3102/0002831214545110

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. Retrieved from http://senatesite.com/utahsenate/dual-language-immersion/

Utah State Board of Education. (2020). Welcome to kindergarten: Frequently asked questions. Retrieved from https://www.schools.utah.gov/curr/kindergarten?mid=1179&tid=3

Valentino, R. A., & Reardon, S. F. (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), 612-637. doi:10.3102/0162373715573310

Vega-Bayo, A., & Mariel, P. (2022). Parents' Willingness to Pay for Bilingualism: Evidence from Spain. *Journal of Family and Economic Issues*. doi:10.1007/s10834-022-09852-1

Williams, C. (2017, December 28). The intrusion of white families into bilingual schools. *The Atlantic*.

Xu, X., Padilla, A. M., & Silva, D. M. (2015). Learner performance in Mandarin immersion and high school world language programs: A comparison. *Foreign Language Annals, 48*(1), 26-38. doi:10.1111/flan.12123

Yuki, K. (2022). Is bilingual education desirable in multilingual countries? *B.E. Journal of Economic Analysis and Policy, 22*(4), 889-949. doi:10.1515/bejeap-2022-0123

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

	Ever One-Way	Ever Two-Way	*Ever-DLI SD*	Never DLI
N Students	***33,476***	***15,340***	***48,816***	***171,975***
Individual Characteristics				
Female	0.49	0.49	*0.50*	0.49
Asian	0.03	0.07	*0.21*	0.04
Black	0.01	0.03	*0.13*	0.02
Hispanic	0.10	0.38	*0.39*	0.16
American Indian	0.01	0.01	*0.10*	0.02
White	0.85	0.50	*0.44*	0.76
Race Other/Missing	0.01	0.01	*0.08*	0.01
Base Free/Red. Lunch	0.27	0.57	*0.48*	0.38
Native Not English	0.09	0.39	*0.39*	0.15
Ever EL	0.08	0.36	*0.38*	0.14
Native/Partner Lang. Match	0.03	0.33	*0.18*	.
Base Special Education	0.11	0.12	*0.32*	0.12
Ever Migrant	0.00	0.02	*0.08*	0.01
Residential Zip Code Characteristics				
Pct. Bach. Deg.	35.14	26.86	*13.44*	29.53
Pct. Grad. Deg.	12.14	8.84	*6.24*	9.48
Pct. Limited Eng. Prof.	1.57	5.23	*2.89*	2.46
Pct. SNAP	6.61	10.53	*4.31*	8.67
Peer Attributes in Base School and Grade				
Pct. White	0.85	0.49	*0.22*	0.78
Pct. Free/Red. Lunch	0.25	0.58	*0.22*	0.35
Pct. Base EL	0.04	0.25	*0.09*	0.07
Pct. Base Special Ed.	0.09	0.09	*0.05*	0.11
DLI Access and Enrollment				
Slots available in gr. 1 (y/n)	0.42	0.41	*0.24*	0.00
Slots per first grader in gr. 1	0.22	0.21	*0.13*	0.00
Base School DLI Language				
Spanish	0.38	1.00	*0.49*	.
Chinese	0.42	0.00	*0.45*	.
French	0.13	0.00	*0.28*	.
Portuguese	0.02	0.00	*0.20*	
German	0.06	0.00	*0.11*	.
Mean Scores Across Obs. Years (Standardized within Subject, Grade, and Year)				
ELA	0.08	-0.32	*0.91*	-0.06
Math	0.14	-0.27	*0.89*	-0.02
Science	0.11	-0.35	*0.89*	-0.06

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

	One-way					Two-way				
	ELA	ELA	ELA	Math	Science	ELA	ELA	ELA	Math	Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
DLI enroll	0.306***	0.306***	0.236***	0.237***	0.232***	0.246***	0.245***	0.301***	0.309***	0.251***
	(0.023)	(0.023)	(0.020)	(0.024)	(0.021)	(0.061)	(0.060)	(0.047)	(0.052)	(0.057)
School-by-year controls		x	x	x	x		x	x	x	x
Individual controls			x	x	x			x	x	x
Observations	79,994	79,994	79,994	80,046	71,958	39,792	39,792	39,792	39,527	34,584
R-squared	0.015	0.015	0.101	0.080	0.075	0.010	0.011	0.140	0.115	0.124
Schools	55	55	55	55	55	27	27	27	27	27

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models include base school and cohort fixed effects, as well as individual and school-by-year controls, with standard errors clustered at the base school level. Estimates pertain to the two academic years for which clean DLI enrollment data were available statewide.

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	0.230***	0.230***	0.232***	0.224***	0.225***	0.218***
	(0.017)	(0.017)	(0.018)	(0.028)	(0.028)	(0.028)
Or Slots per pupil (gr. 1 in first-grade year)	0.458***	0.457***	0.455***	0.430***	0.430***	0.404***
	(0.034)	(0.034)	(0.034)	(0.060)	(0.060)	(0.064)
Second stage						
DLI enroll (Treated instrument)	0.032	-0.011	-0.033	0.179	0.179	0.165
	(0.085)	(0.085)	(0.089)	(0.127)	(0.138)	(0.161)
or DLI enroll (Slots per pupil instrument)	0.042	-0.029	-0.000	0.196~	0.178	0.005
	(0.082)	(0.083)	(0.096)	(0.116)	(0.127)	(0.158)
Observations	79,994	80,046	71,958	39,792	39,527	34,584
Schools	55	55	55	27	27	27

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models include base school and cohort fixed effects, as well as individual and school-by-year controls, with standard errors clustered at the base school level. Estimates pertain only to the academic years for which clean DLI enrollment data are available statewide.

Table 4. ITT estimates for the dichotomous DLI school-by-cohort ITT indicator

	One-way						Two-way					
	Fixed Effects			*Bacon Decomposition*			*Fixed Effects*			*Bacon Decomposition*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science	ELA	Math	Science
DLI offered	0.012	-0.012	-0.012	0.012	-0.012	-0.012	0.052~	0.071*	0.064*	0.052*	0.071**	0.064~
	(0.016)	(0.018)	(0.021)	(0.016)	(0.018)	(0.020)	(0.029)	(0.030)	(0.030)	(0.024)	(0.026)	(0.034)
School-by-cohorts	772	772	722	772	772	722	379	379	353	379	379	353
Schools	55	55	55	55	55	55	27	27	27	27	27	27
R-squared	0.385	0.394	0.384				0.235	0.314	0.167			
Attrib. to timing				0.004	-0.017	-0.022				0.039	0.059	0.027
Timing weight				*0.98*	*0.98*	*0.93*				*0.96*	*0.96*	*0.63*
Attrib. to covar.				0.361	0.447	0.555				0.383	0.372	0.587
Covariates weight				*0.02*	*0.02*	*0.02*				*0.04*	*0.04*	*0.05*
Relative to never-treated						0.032						0.034
Never-treated compar. weight						*0.05*						*0.32*

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models are estimated at the school-by-cohort level with analytic weights for students per school in 2002, and with controls for time-varying school-level fraction of students who are white, eligible for subsidized meals at school entry, ELs at school entry, and special education identified at school entry. They include base school and cohort fixed effects, with standard errors clustered at the base school level.

Table 5. Dichotomous ITT robustness to cohort restrictions and placebo test

	One-way			Two-way		
	(1) ELA	(2) Math	(3) Science	(4) ELA	(5) Math	(6) Science
A. First treated cohort vs previous	0.011	0.002	-0.009	0.059~	0.070~	0.044
	(0.022)	(0.022)	(0.021)	(0.032)	(0.034)	(0.039)
School-by-cohorts	566	566	565	293	293	289
R-squared	0.147	0.208	0.144	0.303	0.370	0.265
Schools	55	55	55	27	27	27
B. First 4 treated cohorts vs previous	0.017	-0.001	-0.000	0.059*	0.081**	0.054~
	(0.018)	(0.020)	(0.019)	(0.028)	(0.028)	(0.030)
School-by-cohorts	706	706	681	351	351	338
R-squared	0.395	0.259	0.211	0.258	0.325	0.324
Schools	55	55	55	27	27	27
C. Treated in early-adopting cohorts	0.018	0.027	0.011	0.020	0.004	0.056
	(0.027)	(0.025)	(0.026)	(0.041)	(0.057)	(0.047)
School-by-cohorts	533	533	497	211	211	196
R-squared	0.256	0.474	0.493	0.190	0.274	0.444
Schools	38	38	38	15	15	15
D. Treated in late-adopting schools	-0.059	-0.059	-0.088	0.088~	0.089	0.082
	(0.048)	(0.044)	(0.059)	(0.049)	(0.077)	(0.084)
School-by-cohorts	239	239	225	168	168	157
R-squared	0.626	0.371	0.305	0.402	0.458	0.505
Schools	17	17	17	12	12	12
E. Placebo first cohort vs previous	-0.004	-0.009	-0.017	0.029	0.062*	0.025
	(0.021)	(0.021)	(0.023)	(0.026)	(0.027)	(0.030)
Schools-by-cohorts	511	511	511	266	266	266
R-squared	0.178	0.216	0.143	0.301	0.390	0.284
Schools	55	55	55	27	27	27

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models are estimated at the school-by-cohort level with analytic weights for students per school in 2002, and with controls for time-varying school-level fraction of students who are white, eligible for subsidized meals at school entry, ELs at school entry, and special education identified at school entry. They include base school and cohort fixed effects, with standard errors clustered at the base school level. Early-adopting schools launched DLI in fall 2008-2012; late-adopting schools launched fall 2013-2017

Table 6. ITT estimates corresponding to Figure 2 for Never-ELs and Language-Match ELs

A. Never EL	One-way			Two-way		
Grade	ELA	Math	Science	ELA	Math	Science
	(1)	(2)	(3)	(4)	(5)	(6)
3	0.027	0.004		0.016	-0.037	
	(0.025)	(0.031)		(0.047)	(0.031)	
4	-0.004	-0.054*	-0.026	0.063~	0.017	0.002
	(0.023)	(0.023)	(0.031)	(0.035)	(0.046)	(0.047)
5	0.016	-0.000	-0.008	-0.027	-0.058	-0.079~
	(0.028)	(0.029)	(0.024)	(0.039)	(0.044)	(0.042)
6	-0.036	-0.034	-0.037	0.067	0.055	0.063
	(0.027)	(0.036)	(0.034)	(0.043)	(0.040)	(0.051)
Schools base gr.	55	55	55	27	27	27
Obs. base gr.	58,875	58,801	58,972	18,697	18,686	18,502
R-sq base gr.	0.070	0.058	0.068	0.088	0.074	0.080
B. Ever-EL: Language Match	One-way			Two-way		
Grade	ELA	Math	Science	ELA	Math	Science
	(1)	(2)	(3)	(4)	(5)	(6)
3	0.081	0.050		0.009	0.109**	
	(0.087)	(0.118)		(0.044)	(0.037)	
4	-0.005	-0.081	-0.054	0.087	0.178**	0.090
	(0.094)	(0.090)	(0.070)	(0.056)	(0.057)	(0.061)
5	0.076	0.063	0.076	0.047	0.095	0.064
	(0.088)	(0.079)	(0.079)	(0.062)	(0.081)	(0.082)
6	-0.004	0.065	0.129	0.091	0.188*	0.086
	(0.088)	(0.056)	(0.087)	(0.067)	(0.077)	(0.057)
Schools base gr.	51	52	51	27	27	27
Obs. base gr.	2,319	2,329	2,227	10,357	10,386	10,035
R-sq base gr.	0.076	0.061	0.062	0.051	0.033	0.031

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models are estimated at the student-by-grade level. They include base school and cohort fixed effects and individual and school-by-grade controls, with standard errors clustered at the base school level.

Table 7. Estimated dichotomous ITT effects on the probability of being classified as EL in each year among those ever classified

	One-way		Two-way	
Grade	Home/ School Language Match (1)	No Language Match (2)	Home/ School Language Match (3)	No Language Match (4)
1	0.011	-0.001	0.001	-0.014
	(0.016)	(0.020)	(0.009)	(0.024)
2	0.021	-0.041	0.004	0.022
	(0.035)	(0.044)	(0.016)	(0.029)
3	0.024	0.013	0.012	-0.025
	(0.033)	(0.045)	(0.023)	(0.032)
4	0.000	0.039	-0.020	0.026
	(0.030)	(0.043)	(0.024)	(0.036)
5	0.002	0.030	-0.058*	0.036
	(0.033)	(0.038)	(0.023)	(0.047)
6	-0.027	-0.040	-0.035	-0.033
	(0.050)	(0.040)	(0.024)	(0.043)
Schools gr. 1	49	55	27	27
Obs. gr. 1	2,406	2,624	10,913	1,680
R-sq gr. 1	0.062	0.099	0.113	0.108

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models include base school and cohort fixed effects, as well as individual and school-by-grade controls, with standard errors clustered at the base school level.

Table 8. ITT estimates investigating why one-way and two-way program effects may differ

	One-way Spanish-Only			One-way Weighted for 2002 Similarity to Two-way		
	(1)	(2)	(3)	(4)	(5)	(6)
	ELA	Math	Science	ELA	Math	Science
Treated school-by-cohorts	0.004	-0.031	0.008	0.028	0.004	-0.007
	(0.035)	(0.052)	(0.038)	(0.024)	(0.026)	(0.032)
School-by-cohorts	772	772	722	294	294	276
R-squared	0.462	0.417	0.341	0.182	0.173	0.365
Schools	55	55	55	21	21	21

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models are estimated at the school-by-cohort level with analytic weights for students per school in 2002, and with controls for time-varying school-level fraction of students who are white, eligible for subsidized meals at school entry, ELs at school entry, and special education identified at school entry. They include base school and cohort fixed effects, with standard errors clustered at the base school level.

Table 9. Interaction of DLI access with share of language-matched students in the school in the pre-treatment year

	(1) ELA	(2) Math	(3) Science
DLI offered (school-by-cohort)	0.019	-0.009	-0.006
	(0.017)	(0.018)	(0.026)
DLI offered * share of students with primary language match (0 to 1)	0.058	0.193*	0.154
	(0.083)	(0.096)	(0.169)
School-by-cohort obs. pooled	1,151	1,151	1,075
R-squared	0.278	0.343	0.246
Schools pooled	82	82	82

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models are estimated at the school-by-cohort level with analytic weights for students per school in 2002, and with controls for time-varying school-level fraction of students who are white, eligible for subsidized meals at school entry, ELs at school entry, and special education identified at school entry. They include base school and cohort fixed effects, with standard errors clustered at the base school level.

Appendix

Table A1. Selection test at school-by-cohort level: Regressing school-by-year attributes on dichotomous treatment or first-grade slots per pupil

	One-way				Two-way			
VARIABLES	(1) Frac white	(2) Frac FRL	(3) Frac ever-EL	(4) Frac sped	(6) Frac white	(7) Frac FRL	(8) Frac ever-EL	(9) Frc sped
Treated school-by-cohort	-0.019*	0.006	-0.000	0.002	-0.014	0.002	-0.060	-0.010
	(0.009)	(0.012)	(0.011)	(0.006)	(0.014)	(0.020)	(0.040)	(0.008)
Slots per pupil	-0.029	-0.025	-0.040	-0.000	-0.041	0.010	-0.043	-0.013
	(0.017)	(0.031)	(0.037)	(0.011)	(0.032)	(0.035)	(0.078)	(0.016)
Schools-by-cohort	935	935	935	935	459	459	459	459
R-squared (trt)	0.216	0.100	0.294	0.260	0.253	0.155	0.705	0.417
Schools	55	55	55	55	27	27	27	27

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 Models are estimated at the school-by-cohort level with analytic weights for students per school in 2002. They include base school and cohort fixed effects, with standard errors clustered at the base school level.