



# Partner-Language Learning Trajectories in Dual-Language Immersion: Evidence From an Urban District

Susan Burkhauser  
*RAND Corporation*

Jennifer L. Steele  
*American University*

Jennifer Li  
*RAND Corporation*

Robert O. Slater  
*American Councils for International Education*

Michael Bacon  
*Portland Public Schools*

Trey Miller  
*RAND Corporation*

**Abstract:** *Research has demonstrated that students in dual-language immersion programs perform as well as, or better than, their peers in core academic content areas by late elementary school. However, the extent to which immersion education fosters bilingualism has received less attention in the literature. Using data from a four-year*

---

*Susan Burkhauser (PhD, Pardee RAND Graduate School) is Assistant Policy Analyst at the RAND Corporation, Santa Monica, CA.*

*Jennifer L. Steele (EdD, Harvard University) is Associate Professor of Education at American University, Washington, DC.*

*Jennifer Li (PhD, University of California, Los Angeles) is Management Scientist and Applied Linguist at the RAND Corporation, Santa Monica, CA.*

*Robert O. Slater (PhD, American University) is Senior Consultant for Policy Research at American Councils for International Education, Washington, DC.*

*Michael Bacon (MEd, Portland State University) is Assistant Director of Dual Language for the Portland Public Schools, Portland, OR.*

*Trey Miller (PhD, Stanford University) is Economist at the RAND Corporation, Santa Monica, CA.*

*Foreign Language Annals, Vol. 49, Iss. 3, pp. 415–433. © 2016 by American Council on the Teaching of Foreign Languages.*

DOI: 10.1111/flan.12218

*efficacy study of dual-language immersion education in the Portland Public Schools in Oregon, this study reports the skill levels that 1,284 dual-language immersion students achieved in their classroom partner languages (Spanish, Japanese, and Mandarin Chinese) between third and eighth grades. The authors found that by eighth grade, the average dual-language immersion student, regardless of language, performed at least at the Intermediate Low sublevel, and often higher, on Standards-Based Measurement of Proficiency assessments of nearly all language skills tested (listening, reading, writing, and speaking). In comparison, eighth graders in the Portland Public Schools who began taking Spanish as an elective in upper elementary or middle school scored only at about the Novice Mid sublevel. After four years of immersion learning (grades K–3), fourth-grade students whose home language was Spanish scored similarly in reading and speaking to their immersion peers whose home language was not Spanish; however, they outperformed their immersion peers in listening and writing.*

**Key words:** *dual-language immersion, one-way immersion, performance assessment, quantitative research, two-way immersion*

The academic benefits of dual-language immersion programs are becoming increasingly clear. In dual-language immersion programs, students receive general academic instruction in two languages from early grades onward. They include both two-way programs, in which about half of the students in a classroom are native speakers of each of the two classroom languages (in the United States, typically English and another language, termed the “partner language”), and one-way programs, in which most students in the classroom are native speakers of English. A key objective of both types of programs is to produce students who emerge bilingual and biliterate, regardless of their first language. Due to a considerable body of observational research on the performance of native English speakers (Lambert, Tucker, & d’Anglejan, 1973; Lapkin, Hart, & Turnbull, 2003; Marian, Shook, & Schroeder, 2013; Padilla, Fan, Xu, & Silva, 2013) and of English learners (ELs) (Collier & Thomas, 2004; Lindholm-Leary & Block, 2010; Marian et al., 2013; Thomas & Collier, 2015) in dual-language immersion programs, the language education community has long understood that students in dual-language immersion can perform as well as, if not better than, their peers on standardized tests of language arts and mathematics administered in English.

In addition, there is evidence that bilingualism, which requires students to switch attention rapidly from one representational system to another, is correlated with enhanced cognitive skills and may therefore positively affect learning in a wide range of subject areas (Bialystok & Craik, 2010). On a national level, such findings have been slow to influence public policy discussions about the importance of access to dual-language immersion programs, perhaps in part because many immersion education studies have been unable to adjust for the unobserved motivation and knowledge levels of families entering immersion programs in lieu of other programs. In recent years, however, a spate of newer studies have been able to use longitudinal, administrative data from large school districts to apply contemporary research methods to the estimation of immersion effects. Using data from the San Francisco Unified School District and employing extensive statistical controls to adjust for baseline selection, Umansky and Reardon (2014) found that native Spanish-speaking ELs placed in two-way Spanish immersion classrooms were initially slower to be reclassified as English-proficient but attained higher rates of English proficiency by high school. Using the same dataset, Valentino and Reardon (2015) found that ELs placed in a bilingual environment—be

it dual-language immersion or other bilingual education programs—showed faster academic growth in English language arts than peers placed in monolingual English programs. Steele et al. (in press) capitalized on the randomization of more than 1,600 students—ELs as well as native English speakers—in immersion programs in Portland, Oregon, and found positive effects in English reading in grades five and eight, accompanied by no evidence of negative effects in other subjects or grades. Although the estimated effect magnitudes in the newer studies are in some cases more modest than those in prior studies, the general conclusions are similar: Immersion students perform as well as, or better than, their peers in English, and the benefits of immersion, especially for English learners, may not be immediately evident but appear to grow over time.

Meanwhile, progress in establishing the benefits of dual-language immersion *has* been evident on some fronts. In the last half decade, Utah, North Carolina, and Delaware have made concerted statewide efforts to increase the number of dual-language immersion programs, typically framing these policies as ways to enhance the global competitiveness of their workforces (Delaware Department of Education and Office of the Governor, 2012; Fasciano, 2013; Utah State Office of Education, n.d.). In other words, the policy justification for immersion rests not only on the fact that it may confer medium- to longer-term benefits on students' English skills but also on the more obvious argument that bilingualism carries advantages in a global marketplace. Rigorous estimates of the earnings returns due to bilingualism in North America range from 2 to 3% for non-English languages in the United States (Saiz & Zoido, 2005) to 4 to 6% for French in Anglophone Canada to 7 to 8% for English in Francophone Quebec (Christofides & Swidinsky, 2010). Bilingualism also appears to confer advantages in terms of social perspective-taking (Fan, Liberman, Keysar, & Kinzler, 2016; Greenberg,

Bellana, & Bialystok, 2013) and intercultural competence (Genesee, 1977; Genesee & Gándara, 1999; Lambert & Tucker, 1972) that are less easily measured in dollars but may be both personally and societally beneficial.

If a key rationale for immersion education is rooted in the benefits of bilingualism, then a critical question for parents and policy makers alike is the extent to which immersion education fosters bilingualism. Although some research does exist that addresses this question, the number of studies is small. Because state and national requirements for standardized testing in the United States focus on tests administered in English, there are typically fewer measurement requirements for—and consequentially less large-scale testing of—students' proficiency in the non-English languages, which, as noted above, we term the “partner language.”<sup>1</sup> Testing students in the partner languages is both time-consuming and expensive and may therefore receive lower priority in a policy context that emphasizes annual testing in mathematics and English language arts.

Drawing on data from a larger study that investigated the causal effects of dual-language immersion in Portland, Oregon, public schools, this article reports the average levels of partner-language proficiency (Spanish, Chinese, or Japanese) that dual-language immersion students achieved. Specifically, the study traced the learners' language proficiency trajectories from third through eighth grade, although the particular grades that were tested varied somewhat by partner language. The study builds on an important but small body of research describing the learning of the partner language among students in dual-language immersion programs. The study differs from previous studies in two ways: The sample was larger than most of the existing studies and, in lieu of strictly cross-sectional analysis, the study tracked the 1,284 individual students' development over time.

## Measuring Partner-Language Proficiency: Key Frameworks and Assessments

In considering the measurement of partner-language proficiency, it is useful to understand how proficiency is often conceptualized and measured in the United States. The U.S. government sets guidelines for measuring language proficiency in reading, writing, listening, and speaking using the Interagency Language Roundtable (ILR) scale (Association of the United States Army, 2010; Interagency Language Roundtable, n.d.a), a six-point scale ranging from 0 (no proficiency) to 5 (functionally native proficiency) in each of four skill areas. If a rating is between two numeric levels, it is denoted with a “+”; for example, an individual may be rated as a 2+ in reading if the person is above a 2 but below a 3 (Interagency Language Roundtable, n.d.b). The Defense Language Institute, operated by the U.S. Department of Defense, estimates that it takes an adult about 26 weeks of intensive study to reach a Level 2 (limited working proficiency) in reading and listening and a Level 1+ (elementary plus) in speaking for the simplest Category 1 languages, such as Spanish, French, Italian, and Portuguese. The Institute estimates that it takes about 64 weeks (two-and-a-half times as long) to reach similar levels of proficiency in the most difficult Category 4 languages, such as Mandarin Chinese, Japanese, and Arabic (Association of the United States Army, 2010). Note that Level 2 listening proficiency is described as the ability to understand speech as it evolves in face-to-face interactions that are spoken at a normal rate, concern routine social or work topics, include some repetition, and include a variety of verb tenses (time frames) (Interagency Language Roundtable, n.d.b). Although these estimates of learning time clearly mask enormous variation among individuals, and, for that matter, were normed on individuals who had been selected for their language learning aptitude, they do offer some guidance about the relative

time needed to reach limited working proficiency in selected languages of varying difficulty.

In 1986, the ACTFL adapted the ILR scale for the academic community, establishing the ACTFL Proficiency Guidelines (ACTFL, 2012). These guidelines divide proficiency into five broad levels: Novice, Intermediate, Advanced, Superior, and Distinguished. Each of the first three levels—Novice through Advanced—is further divided into three subcategories: Low, Mid, and High, resulting in 11 sublevels altogether. The ACTFL guidelines specify the amount of knowledge and skill required to move from one sublevel to the next. Due to the increasingly complex, nuanced, and sophisticated nature of the tasks that are required at each successive proficiency level, the amount of knowledge and skill required to move from Novice Low to Novice Mid is much smaller than the amount of knowledge and skill required to move from Advanced Mid to Advanced High. Thus, while the relationship among levels is algebraic, not linear, any individual user of a language, regardless of the circumstances under which he or she learned the language, can be rated on this scale.

Several assessments of test takers’ knowledge and skills have been developed based on these guidelines. The ACTFL developed and validated the Oral Proficiency Interview, Oral Proficiency Interview—Computer, Business Writing Test, Writing Proficiency Test, Listening Test for Professionals, and Reading Test for Professionals for adult learners. The ACTFL also recently developed a language proficiency assessment for children in grades five through 12, the ACTFL Assessment of Performance Toward Proficiency in Languages, which measures performance in interpersonal listening/speaking, presentational writing, and interpretive reading and listening (ACTFL, n.d.). Another commercial test of world language knowledge and skill in reading, writing, listening, and speaking is the Standards-Based Measurement of Proficiency (STAMP), which is published by

Avant Assessment (2015a). There are also two versions of the STAMP: version 4SE, which is designed for elementary school students, and version 4S, which is designed for students ages 13 through adult (Avant Assessment, 2015a).<sup>2</sup> Topics on each assessment are intended to be age appropriate, with items on the 4SE pertaining to daily school life and items on the 4S addressing a much wider range of topics and contexts. For example, a sample reading item from the STAMP 4SE Spanish language test presents a picture of a living room filled with various items and asks the test taker, in Spanish, “Where is the mother?” (Avant Assessment, n.d.d). In contrast, a sample reading item from the STAMP 4S Spanish language test presents a picture of a driver’s license with information on the license written in Spanish and asks the test taker, in English, “What kind of card is this?” (Avant Assessment, n.d.c). Since much of the existing research on partner-language acquisition in the United States has used the STAMP 4SE, and because both versions of STAMP are used by the Portland Public Schools, STAMP assessment data were used to measure learners’ language development in the current study.

### **Extant Research on Partner-Language Proficiency in Immersion Programs**

Existing research has shown that there is variation in the proficiency levels that are attained by students in dual-language immersion programs, some of which may be attributable to the intensity and length of exposure to the partner language. In one of the earliest North American studies of immersion education, Lambert and colleagues (1973) assessed native English-speaking fourth and fifth graders enrolled in French immersion since kindergarten, finding that “[b]y the end of Grade 4. . . the experimental children as a group had attained a functional bilingualism in French and English.” (p.143). Although the 50 immersion students had not reached the same average

level of French proficiency as a matched group of 30 native French-speaking students, their performance was not statistically different on several measures of French skills, including written composition. It is notable, however, that immersion students in the study received virtually 100% of their instruction in French through grade four, which exceeds the number of contact hours that are available in most of the existing immersion studies cited above.

In the largest U.S. study of dual-language immersion students’ partner-language proficiency outcomes, researchers at the University of Oregon’s Center for Applied Second Language Studies (2013), where the STAMP was developed, used a national database of STAMP-takers to examine the reading, writing, and speaking performance of 1,477 students in grades six through 12 who were studying Chinese, French, Japanese, or Spanish in immersion programs. The study was limited to students whose home languages did not match the classroom partner language and who had participated in an immersion program since at least third grade. In sixth grade, the percentage of students scoring at least at the Intermediate Low level was 17% in reading, 56% in writing, and 41% in speaking (p. 2). Among twelfth graders, however, the corresponding percentages were 91, 94, and 97%, respectively (p. 2). In the two domains for which Intermediate High skills could be measured, 15% of students scored Intermediate High in writing and 3% did so in speaking (p. 5).

Focusing on a high-performing school district in Northern California, Xu, Padilla, and Silva (2015) used the STAMP to compare the Mandarin skills of 48 fourth and fifth graders in a two-way Mandarin immersion program to those of 119 high school students taking fourth-year or fifth-year (AP) Mandarin in the same school district. Researchers used the STAMP 4SE for the elementary students and the 4S for the high school students. The immersion students received 80% of their instruction in Mandarin through first grade, a share that

gradually declined to 50% by grades 4 and 5 (p. 31). In reading, 66% of the fifth graders performed at the Intermediate Low level or better, including 23% who scored at Intermediate High, whereas only 59% of the AP Mandarin students performed at Intermediate Low or better and none reached Intermediate High (p. 31). In speaking, 71% of the fifth graders met or exceeded the Intermediate Low level and 69% did so in writing, although the AP students modestly exceeded these performance levels in speaking and writing (pp. 32–33). Similar patterns were seen in comparing the fourth graders to the fourth-year Mandarin speakers. Moreover, while the proficiency of heritage speakers (defined as those whose home language included Mandarin) noticeably exceeded that of nonheritage speakers in the AP program, the two groups scored almost equivalently in the immersion program. The latter finding was consistent with earlier results reported for the same school by Padilla and colleagues (2013), which showed proficiency gaps between heritage and nonheritage speakers closing by fifth grade.

Fortune and Zhang-Gorke (2014) also used the STAMP 4SE to assess the performance of 81 fifth graders in one-way Mandarin immersion programs in two Minnesota districts. Students in these districts received all instruction in Mandarin until third grade, at which point English was gradually introduced, meaning that students had more classroom-based Mandarin exposure than in the Xu et al. (2015) study. In Chinese reading, 55% scored Intermediate Low or better, and 64% did so in writing—modestly lower rates than were found by Xu et al. (2015). However, in speaking, 84% scored Intermediate Low or better, which exceeded the performance in Xu et al., and 97% reached that threshold in listening.

In a largely ethnographic study examining the language use of native Spanish speakers and native English speakers in a two-way immersion school in Chicago, Potowski (2007) reported that in eighth grade, the

31 native Spanish speakers in the study scored markedly higher in oral language proficiency than the 16 native English speakers on the Language Assessment Scales Oral, with a mean of 85.5 (out of 100) for the former group and 64.9 for the latter. Native Spanish speakers also outperformed native English speakers in writing (24.9 vs. 17.5 out of 30) and also in reading vocabulary (49<sup>th</sup> vs. 34<sup>th</sup> percentile) and comprehension (67<sup>th</sup> vs. 58<sup>th</sup> percentile), although the reading differences were not statistically significant (p. 145). It is not clear why Xu et al. (2015) found little difference in the performance of fifth-grade immersion students by home language while Potowski found substantial differences among the eighth graders. Two possible explanations may lie in the relative amounts of community reinforcement for the students' home languages (Mandarin vs. Spanish) or in the fact that heritage speakers in the Xu et al. study were defined based on the presence of Mandarin in the home and not necessarily on the requirement that it be their first language, whereas the heritage and nonheritage groups in the Potowski study were defined based on the students' own first languages.

The study reported here aimed to contribute to the literature on partner-language acquisition in dual-language immersion programs by examining proficiency growth trajectories for students in Spanish, Mandarin Chinese, and Japanese programs who were observed between third and eighth grades. Specifically, the study sought to provide insight into the following questions:

1. What levels of partner-language achievement do Portland's dual-language immersion students achieve in reading, listening, speaking, and writing, on average, in elementary and middle grades?
2. How do these levels of achievement differ by immersion partner language (Spanish, Japanese, and Chinese)?
3. Do immersion students whose home language is Spanish show different Spanish-language achievement than those whose home language is not Spanish?

## Methods

### *Context*

Data for this study were drawn from a larger study of dual-language immersion in the Portland Public School system in Oregon. The Portland Public Schools began implementing immersion programs in 1986 and have since expanded these programs substantially. During the 2012–2013 academic year, when the larger study commenced, 3,860 individuals, or approximately 8%, of the district's students were enrolled in immersion programs. At this time, the district maintained programs in 11 elementary schools, four middle schools, and five high schools, with instruction in Spanish, Mandarin Chinese, Japanese, and Russian.<sup>3</sup> During the three academic years reflected in our language assessment data (2011–2012 through 2013–2014), all but one of the Spanish programs used a two-way dual-language immersion model in which approximately half of the students were native speakers of Spanish and half were native speakers of English or another language. These programs used a 90/10 instructional time allocation, with 90% of instruction in the partner language in kindergarten declining over time to about 50% by fifth grade and about 29% by middle school. The Chinese and Japanese programs, as well as one of the Spanish programs, used a one-way dual-language immersion model in which most of the classroom participants were native speakers of English. The one-way programs used a 50/50 instructional time allocation, with about half of the instruction in English and half in the partner language until middle school, at which time partner-language instruction was about 29%. Both types of programs continued through high school, but the students in the sample that was used in the study were not yet old enough to be tracked into high school.

### *Participants*

This article focuses on 1,284 students from 14 schools who began as immersion students in the Portland Public School system

in kindergarten. Although the students in our study started kindergarten from fall 2005 through fall 2010, our language assessment data are limited to academic years 2011–2012 through 2013–2014. In 2013–2014, the majority of the oldest cohort of students—those who began kindergarten in fall 2005—were in eighth grade. Disaggregated by language but not by program type (two-way immersion or one-way immersion), data are reported for 728 students in Spanish immersion programs (503 in two-way programs and 225 in the one-way program), 324 in the one-way Japanese program, and 237 in the one-way Chinese program (see Table 1). A handful of students were tested in more than one language. More than half of the children in all of the programs were female. Thirty-two percent ( $N = 230$ ) of students in the Spanish dual-language immersion programs reported Spanish as their native or home language. Among the other 68% ( $N = 498$ ), the home language was almost always English. Because the district had not historically offered set-aside slots for children whose native language was Japanese or Chinese, the share of students in those programs whose home language matched the partner language was quite small, about 3% in each language ( $N = 10$  and  $N = 8$ , respectively).

### *Assessment*

Students in the Spanish program were tested in Spanish in fourth, seventh, and eighth grade; students in the Japanese program were tested in Japanese in third, fourth, fifth, and eighth grade; and students in the Chinese program were tested in Chinese in third, fourth, fifth, seventh, and eighth grade. Students' performance in all four skills (reading, listening, speaking, and writing) was measured with the STAMP 4SE (grades K–6) and STAMP 4S (grades 7–8) assessments. Both versions of the assessment are Web-based, although it is expected that the STAMP 4SE is delivered in a proctored environment. Listening and reading

TABLE 1

Demographic Characteristics for Immersion Students, by Language

	Immersion Program			
	All	Spanish	Chinese	Japanese
N	1,284	728	237	324
Hispanic (%)	24.8	41.8	1.3	3.4
White (%)	47.7	46.4	35.0	59.3
Black (%)	3.7	5.1	2.1	1.5
Asian (%)	17.8	3.0	56.1	24.4
Other race (%)	4.6	2.1	5.1	9.9
Female (%)	55.5	52.9	59.9	59.0
EL in kindergarten (%)	18.7	28.7	8.0	4.0
Home language not English (%)	19.3	28.3	12.7	3.7
Home language matched partner language (%)	19.3	31.6	3.4	3.1

subtests on each version are computer-adaptive, meaning that the difficulty of questions students receive depends on their performance on prior questions. The listening items consist of progressively more difficult dialogues and monologues performed by fluent speakers; the reading items involve answering questions about a realistic scenario (e.g., a graphic of a cell phone screen with a text message in the tested language). Test takers answer questions that are presented on the computer screen in English. For the writing and speaking subtests, tasks are given aurally in English for the 4SE and in print for the 4S. To respond in writing, students type a response into the computer; when responding orally, students record a response directly into the computer using a microphone (Clark, 2012a–c).

The resulting STAMP ratings consist of the major levels Novice, Intermediate, and Advanced and the sublevels Low, Mid, and High (Avant Assessment, 2015b). The STAMP literature states that STAMP levels are “aligned to” (Avant Assessment n.d.a, n.d.b) and “similar to” (Avant Assessment, 2015b) the proficiency levels and sublevels that are described in the ACTFL Proficiency Guidelines (e.g., Novice High, Intermediate

Mid). Some STAMP results are reported using those designations; however, they are not equivalent to official ACTFL ratings. The levels and sublevels are scored using integers ranging from 1 to 9. Sublevels 1–3 represent Novice Low, Mid, and High ratings, respectively. Sublevels 4–6 represent Intermediate Low to Intermediate High, and Sublevels 7–9 represent Advanced Low to Advanced High. On the STAMP for writing and speaking, Advanced Mid and High are compressed at Sublevel 8.

Analyses

Student performance on STAMP assessments of the partner languages over time was estimated using student random-effect models, which adjust for the nesting of observations within students. The simplest specification is shown in Model 1:

$$y_{it} = \alpha_1 + \beta_1' G_{it} + u_{1i} + \varepsilon_{1it} \tag{1}$$

where  $y_{it}$  represents the language performance sublevel score (reading, writing, speaking, or listening) for student  $i$  in grade level  $t$ , and  $G_{it}$  represents a vector of dichotomous grade-level indicators, with effects given by parameter vector  $\beta_1$ . The intercept



term is  $\alpha_1$ . Parameters  $u_{1i}$  and  $\varepsilon_{1it}$  represent student-level and observation-level error terms, respectively, each with mean 0 and variance  $\sigma^2$ . It is important here to acknowledge that this model treats the dependent variable,  $y_{it}$ , as a continuous variable with assumed interval properties even though this is not the case, as explained above. However, interpreting the scores as a linear transformation of an underlying, curvilinear growth process, it is possible to comment on growth curves much as one would comment on earnings curves using the natural log of wages rather than wages in raw dollars.

To examine differences in performance trajectories for immersion students whose home languages did, or did not, match the partner language, Model 1 was extended as follows:

$$y_{it} = \alpha_2 + \beta_2' G_{it} + \delta_2 p_i + \gamma_2' (p_i G_{it}) + u_{2i} + \varepsilon_{2it} \tag{2}$$

In Model 2, the variable  $p_i$  is a dichotomous indicator of whether the student's native or home language matched the partner language, with effects given by  $\delta_2$ . Parameter  $\gamma_2$  expresses the differential effect of each grade level on the performance of

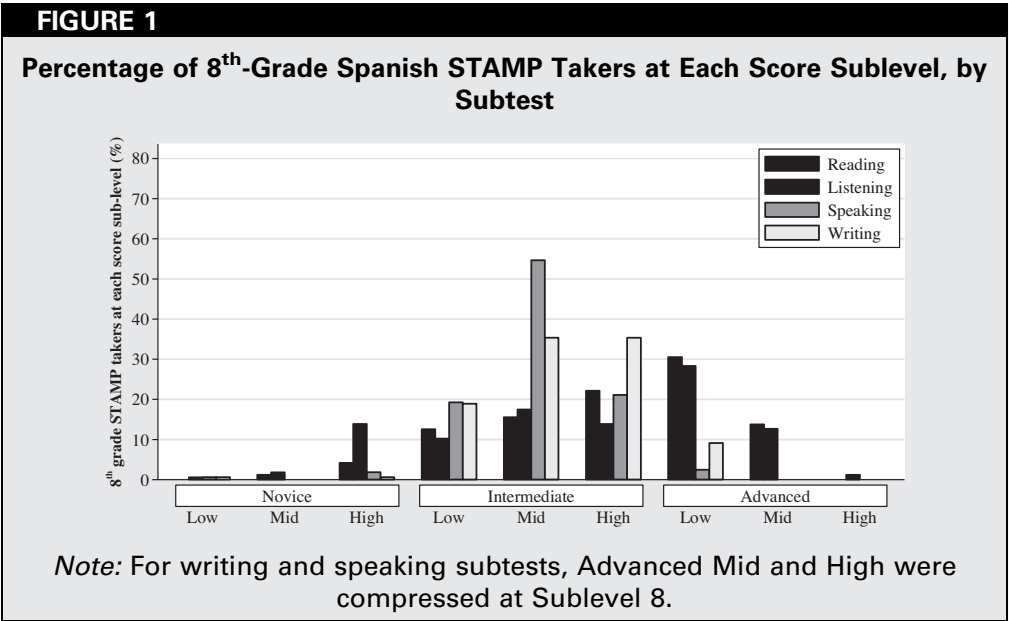
students whose native or home language was the same as the classroom partner language.

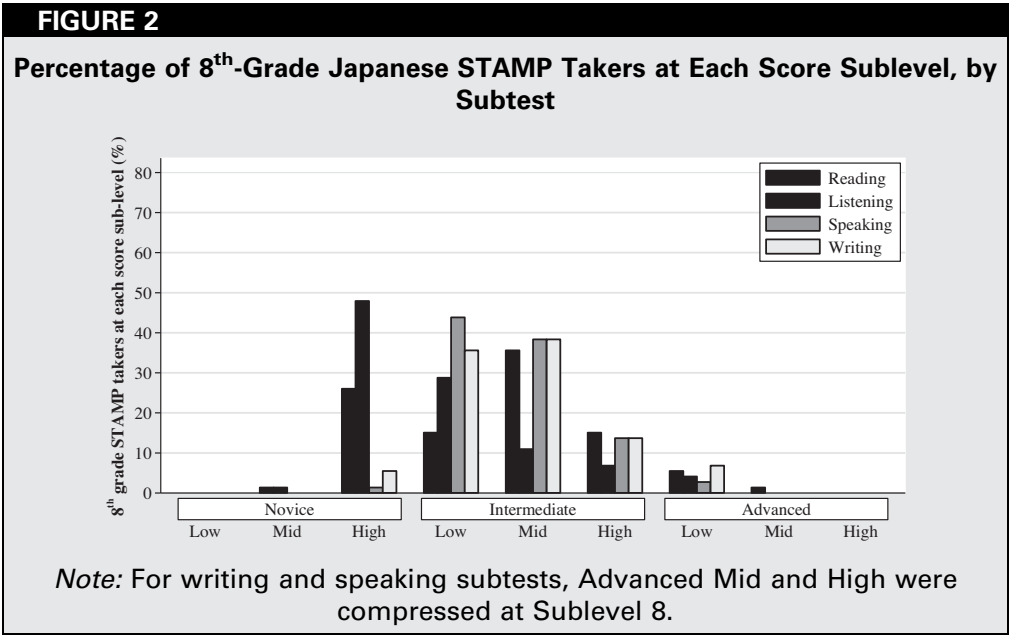
Results

Figure 1, Figure 2, and Figure 3 present the percentage of eighth-grade STAMP takers at each STAMP performance sublevel and skill subtest for Spanish, Japanese, and Chinese, respectively. While the regression analysis provides estimates of student language performance on average, these figures present the variation of student test results at the last point at which they were tested. As shown in these figures, by eighth grade, at least three quarters of the dual-language immersion students were scoring at Sublevel 4 (Intermediate Low) or above, with the exceptions of Chinese reading (29% of students) and Japanese listening and reading (51 and 73%, respectively). It was also possible to trace students' learning trajectories across their K–8 careers using the statistical approach described in Models 1 and 2.

Spanish

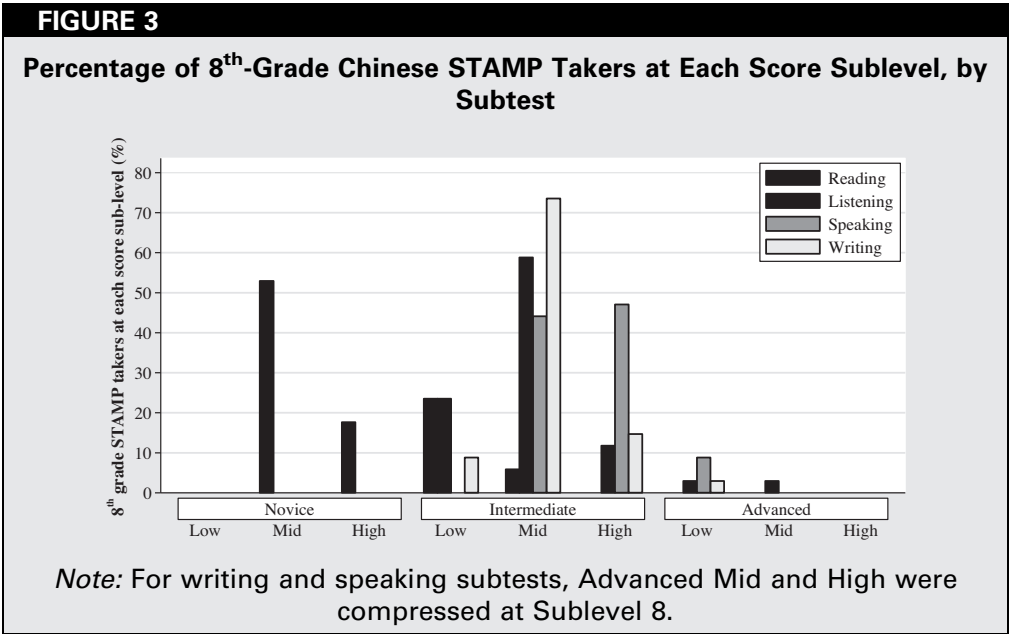
Based on estimates from Model 1 as shown in Figure 4, by fourth grade, students





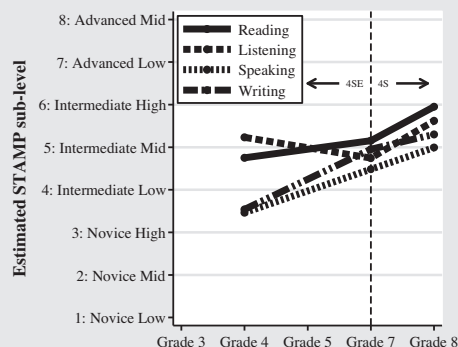
approached or slightly exceeded Sublevel 5 (Intermediate Mid) with estimates as high as 4.8 in reading and 5.2 in listening. Their speaking and writing sublevels were lower, approximately 3.5 each (between Novice High and Intermediate Low). There was

little change from grades four to seven on the receptive language assessments (reading and listening). However, students gained at least one sublevel in the productive skills, speaking and writing, from grade four to grade seven. Despite the change in



### FIGURE 4

## Estimated Spanish STAMP Performance by Grade Level



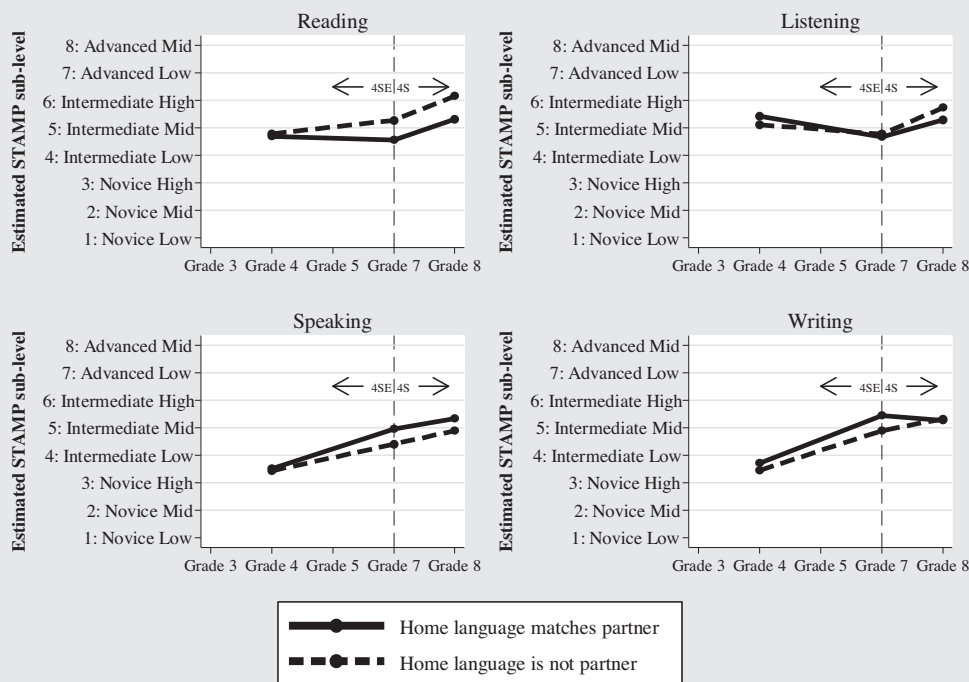
version of the STAMP exam that was used, by eighth grade, students performed at nearly Sublevel 6 (Intermediate High) in reading, 5.6 (between Intermediate Mid

and Intermediate High) in listening, and at 5 and 5.3 respectively (about Intermediate Mid) in speaking and writing.

As depicted in Figure 5, estimates from statistical model 2 showed that by fourth grade (the earliest tested grade) there already appeared to be little difference in the language skills of learners whose home language was, and was not, Spanish for reading and speaking. For listening and writing, home-language speakers of Spanish had an advantage of about a third of a sub-level. In addition, there was little difference in the learning trajectories for these two groups of students until grades seven and eight: A disadvantage seemed to appear for those whose home language was Spanish in seventh and eighth grades for reading and in eighth grade for listening. Because fewer students in the sample were old enough to be observed in eighth grade and because the

### FIGURE 5

### Estimated Spanish STAMP Performance by Grade Level, Home Language Status



subgroup of home-language speakers of Spanish was only about a third of the Spanish immersion sample, the eighth grade results for this subgroup were estimated somewhat less precisely than for the full sample.

*Japanese*

Data for three skills (listening, reading, and speaking) in Japanese were available beginning in third grade, one year earlier than for students in the Spanish program. At that point in time, as shown in Figure 6, the highest performance levels were in listening (4.2, Intermediate Low), and the lowest performance levels—about 3 or Novice High—were in speaking. By fourth grade, students’ performance had improved only modestly (by half of a sublevel or less) in their strongest areas of listening and reading, but it had improved by a full sublevel, to Intermediate Low, in their initially weak area of speaking. By fourth grade, students’ scores were also available in writing, which were the lowest of the four skills at about 3.5, or about Novice High.

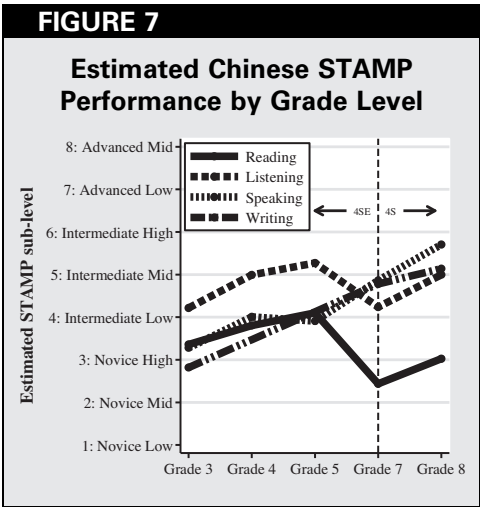
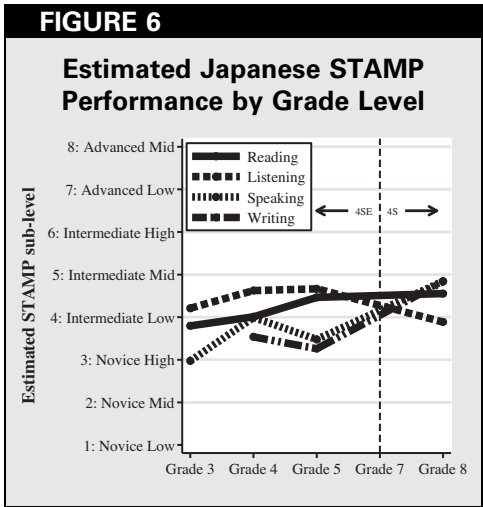
Only modest progress was made in the receptive language skills (reading and listening) by grade eight; however, students’ learning trajectories over time showed marked progress in the productive skills

(speaking and writing). By eighth grade, students achieved scores in the high 4s (Intermediate Low to Mid) in reading, speaking, and writing, and about 3.8 (Novice High to Intermediate Low) in listening. Because the number of students in the Japanese programs whose home language matched the partner language was very small, the disaggregated estimates by home language were too imprecise to report.

*Chinese*

Data for all four skills in Chinese were available beginning in third grade. At that point in time, as shown in Figure 7, the highest performance levels were in listening (4.2, Intermediate Low), and the lowest performance levels—about 3 or Novice High—were in writing. Students’ performance in Chinese rose by nearly a full sublevel in the tested skills between third and fourth grades such that by fourth grade they were scoring at about Intermediate Mid in listening and at approximately Intermediate Low in speaking and reading. Writing was not tested in grade four.

By eighth grade, only modest progress was made in the receptive language skills (reading and listening). In fact, seventh-grade scores in reading declined by more



than one sublevel on average. However, in the productive skills (speaking and writing), students' learning trajectories showed marked improvement over time. In the eighth-grade sample, students scored at or above Sublevel 5 (Intermediate Mid) on average in listening, speaking, and writing but only at Sublevel 3 (Novice Mid to High) in reading. As with Japanese, the small number of students in the Chinese programs whose home language matched the partner language rendered the disaggregated estimates by home language too imprecise to report.

In summary, the data reported here indicated that Portland students in Spanish immersion on average achieved Intermediate Mid skill in listening, speaking, and writing and Intermediate High skill in reading by eighth grade. Those in Japanese immersion on average achieved Intermediate Low skill in listening and reading and about Intermediate Mid proficiency in speaking and writing by grade eight. Those in Chinese immersion on average achieved Novice High skill in reading, Intermediate Mid skill in listening and writing, and nearly Intermediate High skill in speaking. (See Appendix for tables with full model results for each partner language and skill subtest.)

## Discussion

Dual-language immersion schools are on the rise in the United States. While research has demonstrated that students who enroll in dual-language immersion programs perform as well as, or better than, their peers in core content areas by late elementary school, the extent to which immersion students develop skills in listening, speaking, reading, and writing in the partner language has received less attention.

Evidence presented in this article shows that by eighth grade, the average dual-language immersion student, regardless of language, performs at least at the Intermediate Low sublevel, and often higher, on STAMP assessments in nearly

all of the four language skills. This suggests that the average student can understand the main ideas and explicit details in both written and spoken material, short monologues, and conversations; express independent thoughts orally and in writing using an assortment of different vocabulary words; and show good accuracy when using relatively formulaic sentence structures (Avant Assessment, 2015a).

Although one might argue that language performance levels of Intermediate Low to High are not impressive for eighth-grade students who have been enrolled in immersion programs since kindergarten, it is worth noting that Portland students who began taking Spanish as an elective in upper elementary or middle school scored at about Sublevel 2 (Novice Mid) on average in eighth grade, indicating that they were only able to communicate using learned phrases and simple vocabulary (Avant Assessment, 2015a).<sup>4</sup> Moreover, according to the study by the Center for Applied Second Language Studies (2013), fewer than one third of 12<sup>th</sup> graders nationally who took the STAMP and had *not* participated in immersion programs were able to perform at the Intermediate level. Thus, the immersion advantage is clear. In addition, immersion students' learning trajectory across the nine years of instruction provides hope that students are poised to move into and even through the Advanced Low and Mid sublevels with continued instruction in high school. Another key finding is that, after four years of immersion learning (grades K–3), fourth-grade students who spoke Spanish at home no longer scored significantly higher than their counterparts whose home language was not Spanish for two of the four language skills, reading and speaking. Those whose home language was Spanish still outperformed their counterparts in listening and writing. This supports findings from Xu et al. (2015) that by fifth grade there was no clear heritage speaker advantage for Mandarin immersion student performance as measured by STAMP assessments.

When interpreting the results of this study, one should keep a number of limitations in mind. First, two different forms of the STAMP were used—the 4SE through sixth grade and the 4S in seventh and eighth grade. While each version was appropriate to the age and developmental level of the test takers, differences between the instruments should be taken into consideration. Given that the 4S is normed on an older population than the 4SE, evidence of a leveling off between sixth and seventh grades could in part reflect differences in the tests themselves. In addition, one might expect the slopes of the growth trajectories to flatten naturally over time as a function of the nonlinear STAMP scoring scale, given that progression between consecutive sublevels reflects greater amounts of learning at higher sublevels.

The differences in skill-specific achievement levels between Spanish on the one hand and Japanese and Chinese on the other may reflect natural differences in the difficulty of the languages and in their degree of similarity to English. In particular, reading appeared to be the strongest skill among eighth-grade Spanish immersion students, but it was the weakest skill for eighth-grade Chinese immersion students and the second weakest for the Japanese immersion students. It seems plausible that because written Chinese and to a large extent written Japanese are symbolic rather than alphabetic, students' reading ability may be especially dependent on the range of *written* vocabulary they have encountered. In addition, it is important to remember that all participants came from the same school district and that their families purposefully applied for their children to be enrolled in immersion programs. Thus, participants in this study may not be fully representative of the larger population of learners either within the school district under consideration or in other school districts.

Moreover, while one strength of this study lies in its district-wide scope, the district in question is particularly well

established in the area of dual-language immersion education. The Portland Public Schools have 30 years of experience operating immersion programs, provide professional development offerings specifically tailored to the needs of dual-language immersion principals and instructors, and are involved in selecting, purchasing, and creating suitable curriculum materials in the partner languages. Insofar as curricular and instructional quality mediate immersion students' ability to learn partner languages, these contextual factors are important to bear in mind.

The district-wide scope also makes it somewhat difficult to generalize about the instructional practices that may have produced these results, as the study included a diverse array of instructors, grades, and schools. For readers interested in the types of language pedagogy observed in the study, Li, Steele, Slater, Bacon, and Miller (2015) described pedagogical observation data from a sample of two-way immersion classrooms in the district. However, because it was not possible to document the pedagogy of most of the classrooms for which STAMP data were available, we were not able to link pedagogical practices to partner-language skill levels in this study.

## Conclusion

This study reports the results of standardized assessments of language performance for six cohorts of students who began kindergarten in dual-language immersion programs in Spanish, Chinese, or Japanese and were subsequently tracked through eighth grade. Although policy makers may emphasize students' performance on high-stakes tests of mathematics and English language arts, dual-language immersion programs also aim to produce bilingual and biliterate citizens. Because the levels of skill were markedly higher than the skills exhibited by Portland students tested in elective (nonimmersion) Spanish classes in eighth grade and since they compared favorably to those of 12<sup>th</sup> graders across the country who have taken

STAMP assessments within nonimmersion language programs (Center for Applied Second Language Studies, 2013), the findings support those of previous studies and further demonstrate the promise of the dual-language immersion model.

## Notes

1. We define the non-English language as the “partner language” in this study because it does not imply that either English or the non-English language is the “target” or “second” language. Rather, the aim is for students, regardless of their native language, to become bilingual and biliterate. This is achieved through teaching that encompasses both English and a classroom partner language, which is the first language for some students (especially those in two-way programs) and the second or third language for others.
2. All of the STAMP 4S assessments used in the study were normed using Rasch models (one-parameter item response theory models) on students from a volunteer sample of schools across the United States. Avant Assessment’s technical reports noted the Ns of the norming populations (5,000 for Spanish, 1,000 for Chinese, and 150 for Japanese), but they also noted that no demographic information was collected on the norming populations, nor did they clarify what proportion of participating schools were middle schools, high schools, or schools serving adult learners.
3. The Russian program is not included in this analysis due to limited data about students’ proficiency in Russian.
4. The amount of time students may have been exposed to Spanish up to that point varies considerably by school.

## References

- ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Alexandria, VA: Author. Retrieved June 20, 2016, from [http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012\\_FINAL.pdf](http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf)
- ACTFL. (n.d.). *The ACTFL assessment of performance toward proficiency in languages: AAPPL measure FAQs*. Retrieved March 31, 2016, from <http://aappl.actfl.org/aappl-measure-faqs#>
- Association of the United States Army. (2010, August 1). *DLI’s language guidelines*. Retrieved March 31, 2016, from <http://www.ansa.org/publications/ausanews/specialreports/2010/8/Pages/DLI%E2%80%99slanguageguidelines.aspx>
- Avant Assessment. (2015a). *STAMP 4S benchmark and rubric guide*. Eugene, OR: Author.
- Avant Assessment. (2015b). *Avant STAMP score interpretation guide*. Retrieved June 20, 2016, from <http://avantassessment.com/docs/avant-stamp4s-score-interpretation.pdf>
- Avant Assessment. (n.d.a). *Avant STAMP 4S proficiency assessments: Grades 7 through university*. Retrieved June 20, 2016, from <http://avantassessment.com/avant-stamp4s.html>
- Avant Assessment. (n.d.b). *Avant STAMP 4Se proficiency assessments: Grades 3 through 6*. Retrieved June 20, 2016, from <http://avantassessment.com/avant-stamp4se.html>
- Avant Assessment. (n.d.c). *Take an Avant STAMP 4S sample test*. Retrieved June 20, 2016, from <https://stamp4s.avantassessment.com/stamp4s/do/samplelogin>
- Avant Assessment. (n.d.d). *Take an Avant STAMP 4Se sample test*. Retrieved June 20, 2016, from <https://stamp4s.avantassessment.com/stamp4s/do/samplelogin>
- Bialystok, E., & Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science*, 19, 12–23. doi: 10.1177/0963721409358571
- Center for Applied Second Language Studies. (2013). *What levels of proficiency do immersion students achieve?* Eugene, OR: Center for Applied Second Language Studies, University of Oregon. Retrieved August 6, 2015, from <https://casls.uoregon.edu/wp-content/themes/caslstheme/pdfs/tenquestions/TBQImmersionStudentProficiencyRevised.pdf>
- Christofides, L. N., & Swidinsky, R. (2010). *The economic returns to a second official language: English in Quebec and French in the Rest-of-Canada*. Institute for the Study of Labor. Retrieved April 4, 2016, from <https://core.ac.uk/download/files/153/6625815.pdf>

- Clark, M. (2012a). *Avant STAMP 4S (STAndards-based Measurement of Proficiency – 4 Skills): Chinese Technical Report*. Eugene, OR: Center for Applied Second Language Studies (CASLS).
- Clark, M. (2012b). *Avant STAMP 4S (STAndards-based Measurement of Proficiency – 4 Skills): Japanese Technical Report*. Eugene, OR: Center for Applied Second Language Studies (CASLS).
- Clark, M. (2012c). *Avant STAMP 4S (STAndards-based Measurement of Proficiency – 4 Skills): Spanish technical report*. Eugene, OR: Center for Applied Second Language Studies (CASLS).
- Collier, V. P., & Thomas, W. P. (2004). *The astounding effectiveness of dual language education for all*. Fairfax, VA: George Mason University.
- Delaware Department of Education and Office of the Governor. (2012, August 1). *World language immersion program prepares to launch*. Retrieved March 29, 2016, from <http://news.delaware.gov/2012/08/01/world-language-immersion-program/>
- 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, 1090–1097. doi: 10.1177/0956797615574699
- Fasciano, H. (2013). *North Carolina dual language: Ongoing research findings*. Raleigh, NC: North Carolina Department of Public Instruction. Retrieved April 4, 2016, from <http://www.ncpublicschools.org/docs/ccsa/conference/2013/presentations/115.pdf>
- Fortune, T. W., & Zhang-Gorke, Y. (2014, May 8). Early total Mandarin immersion: Academic achievement and Mandarin language development. Presentation at the annual National Chinese Language Conference (NCLC), Los Angeles, CA.
- Genesee, F. (1977). *French immersion students' perceptions of themselves and others: An ethno-linguistic perspective*. Montreal: McGill University.
- Genesee, F., & Gándara, P. (1999). Bilingual education programs: A cross-national perspective. *Journal of Social Issues*, 55, 665–685. doi: 10.1111/0022-4537.00141
- Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective-taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive Development*, 28, 41–50. doi: 10.1016/j.cogdev.2012.10.002
- Interagency Language Roundtable. (n.d.a). History of the ILR scale. Retrieved March 31, 2016, from <http://www.govtilr.org/Skills/IRL%20Scale%20History.htm>
- Interagency Language Roundtable. (n.d.b). *Interagency Language Roundtable language skill level descriptions—Listening*. Retrieved March 31, 2016, from <http://www.govtilr.org/skills/ILRscale3.htm>
- Lambert, W. E., & Tucker, G. R. (1972). *Bilingual education of children: The St. Lambert experiment*. Rowley, MA: Newbury House.
- 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, 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, 6–23.
- Li, J., Steele, J., Slater, R., Bacon, M., & Miller, T. (2015). Teaching practices and language use in two-way dual language immersion programs in a large public school district. *International Multilingual Research Journal*, 10, 31–43. doi: 10.1080/19313152.2016.1118669
- Lindholm-Leary, K. J., & Block, N. (2010). Achievement in predominantly low SES/Hispanic dual language schools. *International Journal of Bilingual Education and Bilingualism*, 13, 43–60. doi: 10.1080/13670050902777546
- 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
- Padilla, A. M., Fan, L., Xu, Z., & Silva, D. (2013). A Mandarin/English two-way immersion program: Language proficiency and academic achievement. *Foreign Language Annals*, 46, 661–679. doi: 10.1111/flan.12060
- Potowski, K. (2007). *Language and identity in a dual immersion school*. Clevedon, UK: Multilingual Matters.
- Saiz, A., & Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States. *Review of Economics and Statistics*, 87, 523–538.
- Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (in



press). The effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal* (Centennial Issue), 53.

Thomas, W. P., & Collier, V. P. (2015). *English learners in North Carolina dual language programs: Year 3 of this study: School year 2009-10*. Fairfax, VA: George Mason University.

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, 879–912. doi: 10.3102/0002831214545110

Utah State Office of Education. (n.d.). *Utah dual language immersion: Why immersion?*

Retrieved March 29, 2016, from <http://www.utahdli.org/whyimmersion.html>

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, 612–637. doi: 10.3102/0162373715573310

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, 26–38. doi: 10.1111/flan.12123

*Submitted May 19, 2016*

*Accepted June 24, 2016*



TABLE A2				
Coefficients (and standard errors) From Student Random-Effects Regressions of Japanese STAMP Performance on Grade Level				
Variables	(1) Reading	(2) Listening	(3) Speaking	(4) Writing
Constant (grade 3)	3.796*** (0.110)	4.211*** (0.095)	2.974*** (0.124)	
Grade 4	0.212* (0.106)	0.414*** (0.085)	1.020*** (0.128)	3.539*** (0.063)
Grade 5	0.668*** (0.176)	0.455** (0.156)	0.503* (0.198)	-0.280* (0.142)
Grade 8	0.756*** (0.163)	-0.324* (0.146)	1.872*** (0.179)	1.297*** (0.126)
Home language matched partner				
Observations	425	424	420	352
Number of students	323	324	324	317

Notes: Home language matched partner language for about 3% of the Japanese language sample. For Japanese writing the comparative grade (i.e., constant in the regression) was fourth grade, and for all other subtests it was third grade. \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ .

TABLE A3				
Coefficients (and standard errors) From Student Random-Effects Regressions of Chinese STAMP Performance on Grade Level				
Variables	(1) Reading	(2) Listening	(3) Speaking	(4) Writing
Constant (grade 3)	3.360*** (0.117)	4.217*** (0.094)	3.279*** (0.100)	2.818*** (0.171)
Grade 4	0.444*** (0.103)	0.776*** (0.085)	0.728*** (0.111)	
Grade 5	0.742*** (0.155)	1.056*** (0.127)	0.629*** (0.149)	
Grade 7	-0.923*** (0.158)	0.020 (0.129)	1.584*** (0.150)	1.963*** (0.206)
Grade 8	-0.333~ (0.202)	0.779*** (0.166)	2.426*** (0.191)	2.322*** (0.220)
Home language matched partner				
Observations	460	467	456	140
Number of students	235	237	234	108

Notes: Home language matched partner language for about 3% of the Chinese language sample. \*\*\* $p < 0.001$ , ~ $p < 0.1$ .