ABSTRACT

Title of Dissertation:	INDIVIDUAL VARIABLES IN CONTEXT: A LONGITUDINAL STUDY OF CHILD AND ADOLESCENT ENGLISH LANGUAGE LEARNERS
	Jason Robert Struck, Doctor of Philosophy, 2023
Dissertation directed by:	Professor Nan Jiang, Second Language Acquisition Dr. Martyn Clark, Second Language Acquisition

Millions of public-school students in the United States are identified as English language learners (ELLs), whose academic success is tied to their second language (L2) English education. Previous research in adult populations indicates that L2 proficiency is related to the contextual variable of the prevalence of one's first language (L1) among their peers, called L1 density, which may also moderate the effects of individual variables such as age and exposure to the L2. Despite its substantial impact on the amount and quality of adult learners' exposure to the L2, the variable of L1 density has received little attention in child and adolescent populations, even though it is unknown what role, if any, L1 density plays in L2 acquisition in a school context. Other outstanding questions concerning individual variables include the nature of the purported rate advantage of later starters and whether the similarity of one's L1 and L2 is related to L2 proficiency.

The current study addressed these questions by analyzing longitudinal L2 proficiency assessment records of 10,879 ELLs in grades 1–12 in the United States. The assessment was WIDA's ACCESS for ELLs Online Test, a national, standardized test with scores for each of the four domains of listening, reading, speaking, and writing. Multilevel models were used to estimate the effects of several variables: age of enrollment in a United States school, length of enrollment, language similarity, and L1 density.

In the fitted model estimates, age of enrollment had a small, positive effect. Length of enrollment had a sizable, positive effect but attenuated over time. ELLs enrolling at a later age progressed slightly slower than ELLs enrolling at an earlier age, contrary to the widely accepted notion that later starters enjoy a rate advantage. Little to no evidence was found for a relationship between test scores and language similarity or L1 density, or that the effects of age of enrollment or length of enrollment varied with L1 density.

The results of this study give evidence for the following conclusions for ELLs in United States schools: an earlier age of enrollment is associated with greater gains in L2 proficiency over time, speakers of different L1s are not expected to become differentially proficient in L2 English, and ELLs' levels of L2 proficiency are not expected to vary with how many of their peers speak the same L1.

INDIVIDUAL VARIABLES IN CONTEXT: A LONGITUDINAL STUDY OF CHILD AND ADOLESCENT ENGLISH LANGUAGE LEARNERS

by

Jason Robert Struck

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Nan Jiang, Chair
Dr. Martyn Clark, Co-Chair
Professor Kira Gor
Dr. Megan Masters
Professor Tracy Sweet, Dean's Representative

Acknowledgements

Special gratitude is due

To my committee members for their guidance and assistance. I want to thank Dr. Nan Jiang for helping me improve my writing and argumentation over my graduate career, Dr. Martyn Clark for helping me frame and structure this dissertation and for pushing me to always consider the bigger picture, Dr. Kira Gor for her contagious enthusiasm for studying bilingualism and multilingualism, Dr. Megan Masters for comments that helped improve and clarify my analyses, and Dr. Tracy Sweet for believing in me and teaching me so much of what I know of multilevel modeling.

To Dr. Steven Ross for inspiring me to research the topic of L1 density and for directing the study of adult immigrants summarized in Chapter 1.

To EDMS faculty who made statistics make sense: Dr. Greg Hancock, Dr. Jeff Harring, Dr. Ji Seung Yang, Dr. Yang Liu, Dr. Tracy Sweet, Dr. Laura Stapleton, and Dr. Peter Steiner.

To those at ARLIS who helped me learn R: Valerie Karuzis (who helped me write my first script) and Meredith Hughes (who taught me script and project management).

To my colleagues at the University of Wisconsin–Madison who supported me in my studies and helped me to expand my statistical expertise: Russell Dimond, Doug Hemken, and Andrew Arnold.

To those at WIDA who provided the data and answered my many questions: Stephen O'Connell, Mark Chapman, Fabiana MacMillan, Nick Kraninger, Aaron Burreson, and Alicia Kim.

To Nicholas Hammond for meticulous proofreading and comments that greatly improved the readability of this dissertation.

To the ELL teachers who rejected the possibility of "holding all else constant" and taught me to see the uniqueness of the individual student: Daniel Struck and Joanna Fischer.

To my wife and daughter for their patience and support as I constantly found myself being pulled between competing responsibilities to family, school, full-time work, and everything else in life.

To Jesus for being the God Who Sees me, who is merciful and answers me.

Thank you all.

Table of Contents

Chapter 1: Introduction

Motivation

In the United States, around 10% of public-school students are identified as English language learners (ELLs), counting five million students, and this number is expected to continue to increase (National Center for Education Statistics, 2022). ELLs who receive specialized English language education tend to outperform their counterparts in English-only programs (Genesee et al., 2005; Hill et al., 2019; Pope, 2016). In order to maintain and improve the quality of ELL education, more research into ELLs' language development is crucial for their success. Individual variables, including the age at which one begins acquiring a second language (L2) and the amount of time one has been exposed to that language, have been the subject of a great number of second language acquisition (SLA) studies. Less attention has been given to the effects of contextual variables on SLA and to the question of whether the effects of individual variables are stable across different contexts. One of these contextual variables is first language (L1) density, the proportion of one's peers who speak the same native language, which may have a moderating effect on the individual variables of age and exposure. The current author's previous study of adult immigrants in Australia found that L1 density had an effect on language learning and moderated an individual variable's effect (Struck, 2020). While that study was with adults, the current dissertation examined a younger population of children and adolescents in grades one through twelve to further explore the moderating effect of language context on individual variables in SLA. Given the size of the ELL population and the importance of language education for their academic achievement, it is critical that additional research be done on their language development trajectories so that educators and administrators can better understand and facilitate the education of their students.

The results of Struck (2020) concerning adult learners are discussed in the next section to introduce and contextualize the concept of L1 density. In this context, qualitative (a teacher's case study) and quantitative data (Struck, 2020) corroborate the L1 density effect. Although important differences exist between its population and that of the current study (children and adolescents), it, along with the literature reviewed in Chapter 2, motivated the current study.

L1 Density and Second Language Acquisition

The Adult Migrant English Program (AMEP) assists immigrants to Australia by teaching them L2 English and helping them integrate into Australian society (Martin, 1998). In the period of interest (1993–1997), individuals were organized into geographically based venues where English language instruction occurred, such as schools, hospitals, or community centers. At these venues, the prevalence of one's L1 varied, so that some individuals learned English as an L2 in an environment where many of their peers spoke the same L1, while others were one of only a few speakers of their L1 at their venue. Kohn (1997), a teacher in the AMEP, observed a correlation between progress in learning the L2 in the classroom, and the reliance on the L1 outside of the classroom. Students who lived among many other speakers of the same L1 (i.e., higher L1 density) tended to use the L1 for socializing and daily needs such as shopping and banking at the expense of using their developing L2. This phenomenon is referred to as "enclosure," representing the degree of overlap between the institutions of the L1 and L2 groups (Schumann, 1978). As language groups increase in size, they tend to develop their own institutions (e.g., workplaces, schools, churches), leading to less overlap between the L1 and L2 groups, with the effect that speakers of that language do not need to acquire and use another language (Chiswick & Miller, 2005). In contrast, for speakers of less common L1s (i.e., learners in lower L1 density environments), learning an L2 remains valuable and necessary for daily

needs, employment, adjustment, and communication, whether that L2 is spoken with native speakers of that language or with speakers of other L1s.

Struck (2020) modeled adult immigrants' SLA in the AMEP context, focusing specifically on the L1 density effect. Data for the study came from a research project led by Dr. Steven Ross (Ross, 1997). The dataset was cross-sectional and contained data for 33,859 individuals with a mean age of arrival of 36.2 years (SD 12.6, median 33.4, range 14.2–97.5), a mean length of residence of 1.7 years (SD 1.0, median 1.5, range 0.1–8.7), and a mean education of 10.5 years (SD 2.6, median 12.0, range 0.0–15.0). Individuals' L2 progress was assessed using the Certificate in Spoken and Written English curriculum. This version of the curriculum was competency-based and used task-based assessments. These tasks were activities learners could do with language, such as taking notes from oral presentations or participating in discussions (Burrows, 1993). The assessments targeted one domain each (reading, writing, listening, or speaking), but involved multiple domains as integrated tasks (e.g., listen to a recording and then write a response). Competency achievement records were scored with dichotomous two-parameter logistic item response theory models to obtain a latent ability estimate for each individual. The L1 density variable was calculated as the proportion of individuals at a venue with the same L1, equal to the number of speakers of a given language divided by the total number of individuals at that venue. Multilevel models predicted these L2 proficiency estimates from individual variables of age of arrival, length of residence, and years of education, and contextual variables of L1 density and its interactions with the individual variables.

In the model results, in addition to a positive effect of education and a negative effect of age of arrival, a key finding of this study was that L1 density had a negative effect on the language outcome. The proposed mechanism was that a greater L1 density was related to a

decrease in opportunities for input and interaction with native speakers of the L2, as well as speakers of other L1s, with whom one would presumably need to use the L2 to communicate. This decreased contact with and use of the L2 would have then led to decreased acquisition of the L2. L1 density was also found to have a negative interaction with the individual variable of age of arrival. The main effects of these two terms were negative, suggesting that the negative effect of age was more pronounced for individuals living alongside many speakers of the same L1, or that the L1 density effect was greater among older individuals. A possible reason for this is, in economic terms, younger individuals would have seen a greater return on investment for any efforts made in learning English, whereas older individuals in majority-L1 communities would have had less incentive given the decreased long-term utility of the L2.

The primary contribution of this study of adult language learners in the AMEP context was that it found a relationship between linguistic context and the individual's language-learning outcomes, and this points to the crucial role of input and interaction in SLA. Given that L1 density played a role in adult SLA, the next question to be answered is, does L1 density also show a relationship with a language outcome in a younger population? The current study attempted to answer this question using many of the same variables, including age of onset and length of exposure. New variables, such as language similarity and the interaction of age of onset and length of exposure, were introduced in order to assess the role of the L1 in SLA and how SLA varied as a function of age. The focus on L1 density, however, remained the same.

The Current Study

Motivated by the findings of the effects of L1 density in adult SLA, the purpose of the current study was to examine individual and contextual variables in the L2 English acquisition of children and adolescents. Although previous studies with similar aims have been completed,

these studies either used samples of adults (e.g., Struck, 2020), had small samples (e.g., Jia & Aaronson, 2003), used cross-sectional data (e.g., Chiswick & Miller, 1992), or made use of unreliable measures of self-assessed L2 proficiency (e.g., Raijman, 2013). The current study aimed to improve upon each of these potential shortcomings by using a large, longitudinal sample of language learners with data from a standardized, national language assessment. It also measured the effect of language similarity, testing whether starting from an L1 more or less similar to the L2 was associated with higher scores on the assessment.

Chapter 2 presents a review of the literature on relevant variables, and it concludes with a discussion of remaining research questions. These questions formed the basis of the research questions of the current study. Chapter 3 describes the context of the data, which consisted of assessment records of children and adolescents who were identified as ELLs in grades one through twelve in the United States. A description of the data and preparation processes are also presented. After this is a formal presentation of the research questions of the current study along with corresponding hypotheses derived from findings in the literature. Support for or against each hypothesis came from statistical models' patterns of statistical significance and strengths of associations. Chapter 4 details these statistical models and highlights the key results from the models. In Chapter 5, these results are related to the literature that was presented in Chapter 2, and each of this dissertation's research questions are answered. Chapter 6 contains a summary of the findings of this dissertation alongside a discussion of limitations and suggestions for future research.

Chapter 2: Child and Adolescent Second Language Acquisition

What follows are summaries and syntheses of select studies on several variables in the SLA of children and adolescents. The literature review below focuses on the issues relevant to the current study in order to contextualize and motivate the research questions. The individual variables of age of onset, exposure to the L2, and rate of acquisition are discussed first, followed by language similarity. Then, the contextual variable of L1 density is examined. While the population of interest in the current study was child and adolescent language learners, the literature review also discusses studies with adults where helpful or where the literature on younger learners is limited. Preference is given to empirical, quantitative studies that estimated partial effects of variables of interest while controlling for others (e.g., multiple regression models as opposed to bivariate correlations). Some small-scale qualitative studies are also included provided they shed light into the possible mechanisms behind the observed effects of SLA variables.

Age of Onset and Length of Exposure

Delimitations

Two fundamental, time-related variables in SLA are age of onset and length of exposure. The age of onset of SLA, the age at which one begins meaningful exposure to the L2, tends to explain the largest proportion of variance in SLA outcomes (Granena & Long, 2013). Actual exposure to and use of the L2 are both necessary to acquire the target language, and this occurs over time. The magnitude of the effect of length of exposure on SLA, how much is acquired over time, is referred to as "rate." Age of onset and length of exposure interact so that the rate at which the L2 is acquired is not the same for all learners. Older learners generally acquire the L2 faster, a phenomenon termed "rate advantage" (Long, 1990; see also Krashen et al., 1979). It is

important to note that the issue of rate is separate from that of ultimate attainment, and consequently the literature on sensitive periods in SLA. Rate refers to the slope of language learning, or the amount of progress made in some length of exposure, usually close to the age of onset. Ultimate attainment, on the other hand, is the eventual level attained by language learners, from which there is no further progress, usually ten or more years after onset (DeKeyser, 2000).

The following discussion of child and adolescent learners' SLA is limited to studies taking place in second language contexts. In these contexts, the onset of SLA is marked through events such as enrollment in a school where the L2 is spoken or immigration to a country where the L2 is a dominant language. Much research has been done with respect to SLA within foreign language contexts (e.g., Muñoz, 2006), but such contexts are very different from the context of the current study. In foreign language studies, exposure to the L2 typically occurs in the classroom for only an hour or so per day, and differences by age in rate of learning are observed after a few hundred hours. In second language contexts, where learners are exposed to the L2 for several hours per day, hundreds of hours are reached in mere months. In the current study, the average learner had been enrolled in a school where the L2 was spoken for several years (mean 4.8 years; see Table 5, p. 44), making the amount of exposure magnitudes larger than that of learners in foreign language contexts, so findings from foreign language contexts cannot be easily generalized to language learners in a second language context.

Distinguishing Second from First Language Acquisition

An initial question in child SLA is when *second* language acquisition begins, or, the age after which a language is considered an L2 and before which it is considered an L1. To answer this question, Meisel (2009) recorded speech samples from 22 L1 German learners of L2 French every three to five months over two years. The subjects all lived in France where their home

language was German, and their exposure to French was at a French preschool, which began around age three. Ages at enrollment ranged from 2.7 to 4.0 years. For children whose age of enrollment was at least 3.6 years, their errors in finiteness and grammatical gender resembled those of adult learners of L2 French and were not attributable to the application of L1 grammatical rules to the L2 (i.e., L1 transfer). Children who began learning French before this time made errors similar to those of simultaneous bilingual children. Based on these findings, a possible division between L1 and L2 acquisition (i.e., simultaneous and sequential bilingualism) may be around 3.6 years.

Studies with Children and Adolescents

Several studies examining age of onset and length of exposure effects in child and adolescent SLA are summarized below. The essential facts and findings from these studies are presented in Table 1 (p. 14).

Golberg et al. (2008) examined the vocabulary acquisition of 19 child learners of L2 English in Canada. Five of the subjects were born in Canada, but none received significant exposure to English until enrollment in school. The mean age at the first round of data collection was 5.3 years (SD 0.8, range 4.2–6.8) with an average length of enrollment at that time of 0.8 years (SD 0.4). The authors measured subjects' receptive vocabulary with the Peabody Picture Vocabulary Test Third Edition (PPVT–III) and productive vocabulary with spontaneous speech samples every six months for two years. Age of enrollment was dichotomized to before and after five years in the analyses. Results showed that the group with later starters had a rate advantage for receptive vocabulary, but there was no difference between the two groups for productive vocabulary.

Hopp (2011) assessed the morphosyntax of 60 child learners of L2 German. The mean age of first exposure to German was 3.3 years (SD 0.8, range 1.2–5.0) and the mean length of exposure was 2.0 years (SD 0.9 years, range 0.4–5.3). An elicited production task was used to measure accuracy of case markings and gender markings in determiner phrases, and gender agreement. After controlling for length of exposure, age of onset was not significantly correlated with accuracy on any measure, but length of exposure was moderately correlated with each measure after controlling for age of onset.

Jia and Aaronson (2003) followed 10 child L1 Chinese learners of L2 English for three years. The mean age of arrival in the United States was 9.9 years (SD 3.7, range 5–16). The length of exposure was not provided, but subjects immigrated in the summer, and data collection began in December, meaning that subjects' lengths of exposure all began at around six months (assuming arrival in June) and ended three years later at 3.5 years. The authors used a grammaticality judgment task, an L1-L2 translation task, and surveys of language use. For both the grammar and translation tasks, no effect of age of arrival was found for the first two years, but a lower age of arrival was associated with a higher score for both tasks in the third year. In the language use surveys, it was found that children who arrived in the L2 country before nine years of age preferred using the L2 over the L1, while those arriving later continued to prefer the L1. Younger arrivals also had more L2-speaking friends and higher rates of L2 media consumption. Therefore, the apparent age-of-arrival effect was confounded with variables of language exposure and preference.

Other researchers have made findings similar to Jia and Aaronson (2003), who found that a later age of arrival was correlated with a preference for the L1 and lower consumption of L2 media. Szuber (2007) found with 59 L1 Polish adolescent learners of L2 English, who had a

mean age of arrival of 15.2 years (SD 1.5, range 11.3–18.6) and mean length of residence of 2.1 years (SD 1.7, range 0.3–7.6), that an older age of arrival was correlated with less exposure to the L2 outside of the home. Likewise, Singleton (2003) argues that adolescent immigrants, in comparison with child immigrants, have more autonomy in self-selecting their language environment. This would have the consequence of heterogeneity in L2 exposure by age of arrival, with later arrivals showing greater variability.

Jia and Fuse (2007) studied the same subjects as Jia and Aaronson (2003), following them for five years and assessing their morphosyntax on 16 occasions. The authors used spontaneous speech and elicited production tasks, and they coded speech samples for accuracy in use of the third person singular, past tense markers, the "be" and "do" copula and auxiliaries, and the progressive aspect. After controlling for a language environment composite measure (reflecting language of media and home language), an earlier age of arrival was found to correlate with higher accuracy. The advantage was consistent, but it became statistically significant only at the end of the study, and only for third person singular and past tense. Another finding was that different morphemes showed different patterns over time, such as the progressive aspect showing early, rapid acquisition, and the regular past tense showing no significant change over time.

Paradis (2011) examined the vocabulary and morphosyntax of 169 child learners of L2 English. The mean age of first exposure to English was 4.2 years (SD 0.9, range 0.8–6.3), and the mean length of exposure was 1.6 years (SD 1.0, range 0.3–5.2). Vocabulary was assessed with the PPVT-III, and morphosyntax was assessed with an elicited production task, which was scored for accuracy on the use of third person singular, past tense, and "be" and "do". Stepwise

multiple regression analyses for both measurements found positive effects of length of exposure, providing evidence for the importance of exposure to the L2.

Creagh et al. (2019) investigated how long it would take ELLs to achieve parity with their native English speaker peers on a standardized reading assessment in Australia's National Assessment Program – Literacy and Numeracy. This assessment was administered to all children in grades three, five, seven, and nine. Their subjects were 1,872 learners of L2 English in Australia, divided into three groups based on age of arrival. Because the test was administered periodically, the age-of-arrival groups and time-to-parity can only be described as ranges. The age-of-arrival groups were less than five to eight years old (which included being born in Australia), between seven and ten years old, and between nine and twelve years old. The authors used propensity score matching to find a comparable group of native English speakers (matched on age, gender, and parental education), and parity was operationalized as the lack of a statistically significant difference between the two groups' reading scores (i.e., a failure to reject the null hypothesis of no difference). The first group (age of arrival less than five to eight years) achieved parity in four to six years, and the second group (age of arrival seven to ten) did so in two to four years. The third group (age of arrival nine to twelve) did not achieve parity within the two to four years of observation but continued to score below the native speaker group. While this study is limited by the lack of granularity in age-of-arrival measurements, its findings suggest that the rate of learning may not be monotonic: the middle of three age-of-arrival groups showed the quickest improvement. As a matter of speculation, the oldest group's slow progress may owe itself to a preference for the L1 and self-selection into L1 groups (Jia & Aaronson, 2003; Singleton, 2003; Szuber, 2007). If that was the case, it would explain why the oldest group

did not show a rate advantage over the second oldest group, since the oldest group may have had the least amount of exposure to the L2, slowing SLA and leading to lower reading scores.

Two summaries of child and adolescent SLA have argued for a rate advantage for older learners. Krashen et al. (1979) found that older children (generally, age of onset over 10) enjoy a faster rate of learning morphosyntax than younger children in second language contexts, but younger children eventually catch up and surpass older children. As for vocabulary acquisition, Paradis (2007) maintains that, on the basis of a shared conceptual system, L2 lexical acquisition is faster if the concepts have already been acquired in the L1. Psycholinguistic research provides evidence that L2 words can be mapped to existing concepts (Meade & Dijkstra, 2017). That is, older learners who have already learned many concepts and words in the L1 must only learn the L2 word form, while younger learners need to learn both the word and the concept, with the result that older learners are faster. Vocabulary acquisition is further sped up if the L2 words are cognates of the L1 words, meaning that a greater similarity between the L1 and L2 in the form of loanwords or language typology aids vocabulary learning (see next section, "Language Similarity").

In summary of the studies just reviewed, a later age of onset is often associated with a faster rate of learning, but individuals with an earlier age of onset may eventually catch up and surpass those with a later start. Studies that did not find a statistically significant effect of age of onset recruited subjects with a minimum age of onset of around one year (Hopp, 2011; Paradis, 2011), so it is possible age plays less of a role in the very early years as the line between first and second language acquisition is blurred. (The current study analyzed learners with a minimum age of onset of four years to try to limit the scope to *second* language acquisition.) Length of exposure had a positive effect, so more exposure to the L2 was associated with a higher outcome,

but only in those studies with very young subjects. The findings of Jia and Aaronson (2003) point toward an interaction between age of onset and length of exposure so that earlier arrivals may prefer the L2 while later arrivals prefer the L1, resulting in a suppression of the effect of length of exposure in some populations because exposure increases little with time. These studies recruited relatively small populations with limited ranges of age of onset and length of exposure though, so this hypothesis cannot be directly tested.

Table 1

Summary of Studies on Age of Onset and Length of Exposure Effects in Children and Adolescents

Study	Subjects	Age of onset (years): mean (SD) min-max	Length of exposure (years): mean (SD) min-max	Outcome	Findings
Golberg et al. (2008)	19 learners of L2 English	5.3 (0.8) 4.2–6.8	0.8 (0.4)	Receptive vocabulary, productive vocabulary	Rate advantage for later age of onset for receptive vocabulary
Hopp (2011)	60 learners of L2 German	3.3 (0.8) 1.2–5.0	2.0 (0.9) 0.4–5.3	Productive accuracy in case markings, gender markings, gender agreement	Positive effect of length of exposure
Jia and Aaronson (2003)	10 L1 Chinese learners of L2 English	9.9 (3.7) 5–16	Around 0.5 at beginning of study and 3.5 at end of study	Grammaticality judgment task, translation task	Earlier age of onset surpassed later after two years; later age of onset preferred L1 and earlier preferred L2
Szuber (2007)	59 L1 Polish learners of L2 English	15.2 (1.5) 11.3–18.6	2.1 (1.7) 0.3–7.6	Survey of language exposure	Later age of onset correlated with less exposure to L2
Jia and Fuse (2007)	10 L1 Chinese learners of L2 English (same subjects as Jia and Aaronson, 2003)	9.9 (3.7) 5–16	Around 0.5 at beginning of study and 5.5 at end of study	Productive morphosyntax (accuracy in third person singular, past tense, be, do, progressive aspect)	Earlier age of onset surpassed later after five years for some outcomes
Paradis (2011)	169 learners of L2 English	4.2 (0.9) 0.8–6.3	1.6 (1.0) 0.3–5.2	Receptive vocabulary, productive morphosyntax (accuracy in third person singular, past tense, be, do)	Positive effect of length of exposure
Creagh et al. (2019)	1872 learners of L2 English	Ranges: less than 5–8, 7–10, 9–12	Varied by group, range 2–15	Reading score from National Assessment Program – Literacy and Numeracy test	Rate advantage for middle age-of-onset group

Language Similarity

Metrics

The L1 plays a role in learning an L2, with the result that an L2 is easier to learn if it is more similar to the L1 (Odlin, 2003). The relative similarity of two languages is called language similarity or language distance, depending on which way the scale runs. The term "language similarity" is favored even though "language distance" is more prevalent, since the measure used in the current study was one of language similarity. Language similarity can be quantified theoretically according to a wide range of features, including geographical distance between supposed centers of languages (e.g., Beijing as the center of Chinese), language typology, script, character frequencies, and a variety of morphosyntactic features such as question word order and verbal person marking (Cysouw, 2013). Metrics can be as complex as continuous measures that are composites of many of these features, or as simple as two broad categories (e.g., Indo-European versus non-Indo-European). Language similarity can also be quantified through empirical means by assessing the L2 ability of language learners after some time learning the language.

The Gateway Languages Database is a theoretical language similarity measure that was created for the purpose of identifying languages suitable for cross-training in crisis events (Gnanadesikan & van Rossum, 2016). It quantifies the relative ease of learning one language given another language as the starting point. The database contains 101 languages, and each language pair has a similarity metric. Language similarity can range from zero to one hundred and is a composite of several subscores: related words and languages, script, phonology, morphology, word types, syntax, and sociolinguistic factors. The similarity score is based on the notion that another language is easier to learn if it is more similar to another already known

language. Clark et al. (2016) found empirical support for the scale, in that the mean Gateway Languages Database similarity scores decrease across the Defense Language Institute Foreign Language Center's increasing language difficulty categories, which were based on observations of learner progress (i.e., an empirical metric; see Masters, 2023).

Another empirical language similarity measure comes from Chiswick and Miller (2005), who reported language score data from Hart-Gonzalez and Lindemann (1993). The scores were obtained from L1 English learners of foreign languages at the Foreign Service Institute, such that lower scores represent more difficult languages, and higher scores represent less difficult languages. Chiswick and Miller (2001) took the inverse of the language score measure to create a language distance measure. The authors modeled the L2 English use of immigrants to Canada in their study, which is summarized in more detail in the section "L1 Density." Individuals who spoke L1s more distant from English were predicted to have a lower likelihood of speaking English.

Table 2 has correlation coefficients for all six pairwise combinations of the four language similarity metrics from the Gateway Languages Database, Defense Language Institute, and Chiswick and Miller's (2005) paper. All correlations were statistically significant ($ps < .001$). Because some metrics are of similarity or ease while others measure distance or difficulty, the signs of the coefficients vary. While the Gateway Languages Database has statistics for all pairs of languages, the other three measures only have statistics for pairs including English, so only similarities with English were used to calculate the correlations. The high correlations provide evidence for the construct validity of language similarity across theoretical and empirical measures, and they support the use of any one of these four measures.

Table 2

Correlations of Theoretical and Empirical Language Similarity Measures

	Gateway	DLIFLC	Language Score
DLIFLC	−0.80		
Language Score	0.78	−0.85	
Language Distance	−0.68	0.73	−0.93

Note. Gateway = Gateway Languages Database similarity score; DLIFLC = Defense Language Institute Foreign Language Center language difficulty categories; Language Score = language score reported in Chiswick and Miller (2005); Language Distance = Chiswick and Miller's (2005) language distance score, calculated as the inverse of Language Score.

Relationship with Second Language Acquisition Outcomes

Some have argued that language similarity has a stronger correlation with basic skills like phonological awareness than complex skills like listening comprehension (Jeon & Yamashita, 2014; Melby-Lervåg & Lervåg, 2011). That is, language similarity is more likely to play a role in early SLA than in its later stages. A separate issue with these two studies, both of which were meta-analyses, was that they operationalized language similarity as dichotomous variables of Indo-European languages versus non-Indo-European languages, and alphabetic versus nonalphabetic script. Jeon and Yamashita (2014) note that such broad categories group together dissimilar languages such as English and German. Language similarity was used to predict L2 morphological knowledge, but since these two languages represent decidedly different starting points, where German has a richer case marking system than does English, grouping together such different languages is hardly appropriate. This underscores the importance of using more fine-grained measurements of language similarity.

Ross (2001) found evidence for a weak relationship between language similarity and general language skills. This study included models of the effects of individual variables on task-

based assessment achievement in Australia's Adult Migrant English Program, the same context described in Chapter 1 of this dissertation. The outcome was a count of how many of the task-based competency assessments were achieved, where example tasks included giving directions to one's home, calling a television repair service, and rescheduling a job interview. A series of identical structural equation models were fit to subsets of the data with learners at different levels in the program (Ns = 9447, 7555, and 4417), predicting competency achievement counts by education, hours enrolled in the program, length of residence, age, and language similarity. The theoretical similarity of each L1 and L2 English was calculated as a function of script, canonical word order, and language typology. Languages more similar to English were assigned higher scores. In each of the three models, language similarity's effect was positive, and its standardized coefficient was one of the smallest at 0.04 (it was the same in each model after rounding), meaning that language similarity was only weakly related to language task performance.

Studies with Children and Adolescents

Other studies have looked at the role of language similarity in younger populations. Chen et al. (2012) examined the role of L1 in L2 vocabulary acquisition for 89 L1 Spanish and 77 L1 Chinese learners of L2 English in Canada. Subjects were in either fourth or seventh grade. All had lived in Canada for at least two years, and some were born in Canada: 52% of the Spanish speakers and 17% of the Chinese speakers. One part of the study was the assessment of vocabulary with the PPVT-III. The test was shortened from 180 to 60 items, and then divided into Spanish-English cognates (35 items) and non-cognates (25 items), and the non-cognates had a slightly higher mean frequency. Regression models predicted vocabulary scores, separated into cognates and non-cognates, by individual variables for the two L1 groups. After controlling for age, maternal education, and nonverbal reasoning (as assessed by Raven's Standard Progressive

Matrices), length of residence in Canada showed different patterns for each group. For the L1 Spanish children, length of residence had a statistically significant and positive effect on non-cognate vocabulary ($\hat{\beta} = 0.22$), but a non-significant and near-zero effect for the cognate vocabulary ($\hat{\beta} = -0.01$). For the L1 Chinese children, length of residence was statistically significant and had a positive effect for both cognate ($\hat{\beta} = 0.58$) and non-cognate vocabulary ($\hat{\beta} = 0.61$). The difference between scores on the two vocabulary types was smaller for the L1 Spanish group. The authors argued that the L1 Spanish children used cognates to quickly learn L2 English vocabulary, thus nullifying the effect of length of residence on this item type. L1 Chinese children were not able to use such a strategy since Chinese and English do not share cognates like Spanish and English do, so length of residence remained an important predictor for both vocabulary types.

In Paradis' (2011) elicited production task (summarized in the section "Age of Onset and Length of Exposure"), verb morphology accuracy was partially predicted by whether one's L1 marked tense. Of the children's L1s, Chinese (Mandarin and Cantonese) was the only one that did not mark tense. The other L1s were Hindi, Punjabi, Urdu, Spanish, and Arabic. After controlling for other demographic and cognitive factors, the unstandardized coefficient of the L1 tense indicator variable was 0.197. The task was scored as proportion correct and had a mean score of 0.52 (SD 0.28, range 0.05–1.00), so whether one's L1 marked tense was associated with a difference in scores of nearly 20 percentage points. Because the only non-tense-marking language was Chinese, it is not possible to generalize the findings to all languages that do not mark tense, but it can at least be claimed that the speakers of different L1 groups may vary in their acquisition of L2 morphology.

Likewise, Paradis (2005) examined morphology errors of 24 child learners of L2 English in Canada. Subjects were first exposed to English in school, and they had a mean age of enrollment of 4.7 years (SD 1.0, range 3.3–7.4) and a mean length of enrollment of 0.8 years (SD 0.3, range 0.2–1.5). The task was a forty-five-minute interview designed to elicit spontaneous speech with a variety of target structures. Questions were coded on which structure they required, and responses from subjects were coded for errors of omission and commission. Target structures were divided into two types: tense (third person singular, regular and irregular past tense, *be*, and *do*) and non-tense (progressive aspect, prepositions, plural, articles). Subjects were split into two groups based on their L1s, with Chinese (Mandarin and Cantonese) in the non-richly inflected L1 group and all others (including Arabic, Japanese, and Spanish, among others) in the richly inflected L1 group. The non-richly inflected group had a lower score for both target structure groups, and Mann-Whitney *U* tests showed the difference was statistically significant for the non-tense morphemes ($z = -2.27, p = .024$) but not the tense morphemes ($z = -1.53, p = .126$). The author noted that the difference for the tense morphemes, while non-significant, trended in the same direction as the non-tense morphemes and would have possibly become statistically significant with greater power.

The studies reviewed in this section provide evidence that language similarity plays a role in SLA, but two unknowns remain. The studies with children and adolescents have had limited numbers and types of L1s represented, so that the language similarity measure was often an indicator of whether one's L1 was Chinese (Chen et al., 2012; Paradis 2005, 2011). This confounds L1 effects with group effects, making it unknown whether analyses revealed an effect of language similarity or some experience of a certain L1 group (e.g., socioeconomic status). Additionally, studies with younger populations focused on particular areas of language, such as

vocabulary or morphosyntax, so the question remains what would be found with more broadly defined task-based assessments such as listening comprehension, since such later-stage language skills may have a small correlation with language similarity only detectable with larger samples (Jeon & Yamashita, 2014; Melby-Lervåg & Lervåg, 2011; Ross, 2001).

L1 Density

The concept of L1 density, as introduced in Chapter 1, refers to the concentration of an individual's L1 in some area, and it indicates whether an individual is more or less likely to require using an L2 to communicate with others. Vervoort et al. (2012) used structural equation modeling to examine the relationships between L1 density, contact with speakers of the same L1, and L2 proficiency and use. They surveyed 2,163 immigrants from Turkey and Morocco to the Netherlands who were at least 15 years old. Rather than L1 density, they used the very similar measure of ethnic residential concentration. The relationships between ethnic residential concentration and (self-rated) majority language (L2) proficiency and use were mediated by contact with natives and co-ethnics. A simplified version of the model may be represented as follows:

L1 density → Contact with L1/L2 → L2 proficiency and use

Living near more people of the same ethnicity (i.e., the same L1) led to more contact with them and less contact with native speakers of the L2, which led to lower proficiency in and use of the L2. The reverse was also true; living near fewer co-ethnics led to more contact with natives and greater L2 proficiency and use.

Psycholinguistic research has proposed mechanisms underlying the relationship between contact with and use of the L2 and L2 proficiency. Language phonetic representations may be shaped through use, so greater L1 use may cause L2 sounds to be categorized like their closest

L1 counterparts, leading to L1-like (accented) L2 speech (Piske et al., 2001). In this way, L1 density may affect L2 proficiency by influencing mental representations of the L2.

To the author's knowledge, only two studies have examined the role of L1 density in L2 proficiency in child and adolescent populations (Ahn & Jepsen, 2015; Friesen & Krauth, 2011). Before discussing these and other relevant papers, it is helpful to first review the findings related to the L1 density effect in studies of adult SLA. The studies on the L1 density effect in adult populations are summarized in Table 3 (p. 26).

Studies with Adults

Chiswick and Miller (2001) modeled the L2 use of immigrants to Canada. They used 1991 census data of 32,168 adult males who were born outside Canada in non-English speaking countries. The authors used multinomial logit models to predict whether individuals would report speaking an official language (English or French) but not at home, or speaking an official language including at home, versus the reference category of not speaking an official language. The models controlled for several demographic and geographic variables, and of particular interest are age of arrival, length of residence, and "minority language concentration," or the percent reporting speaking the same L1 in the metropolitan area or province (i.e., L1 density). Age of arrival had a negative effect on the predicted likelihood of speaking an official language. Length of residence had a positive effect, but this diminished over time, as evidenced by a negative quadratic term, meaning the rate of SLA slowed over time. L1 density had a negative effect, meaning that individuals in regions with a greater share of the population speaking the same L1 were less likely to use an official language.

The same authors have co-authored similar publications using census data to model the L2 proficiency of immigrants in the United States (N = 32,255), Canada (N = 23,741), and

Australia (N = 7,288; Chiswick & Miller, 1992, 1995). For the United States and Australian data, L2 proficiency was a dichotomous variable indicating whether an individual reported speaking only English or a high level of self-rated proficiency, as opposed to a low level of proficiency or no proficiency in English. In studies of these two contexts, when calculating the L1 density variable, individuals who reported speaking only English were assigned a value of zero. That is, speakers of the most common language in these two contexts (English) were treated as if they spoke the least common language. L1 density was found to have a negative effect on L2 proficiency, but this is obviously biased due to the way monolingual English speakers were handled. The magnitude of the bias is unknown, but it suggests that the L1 density effect estimate is untrustworthy.

The authors gave no indication of treating the L1 density variable in the Canadian data in this way, however (Chiswick & Miller, 1992). The Canadian census asked individuals whether they spoke French and English, just English, or just French well enough to conduct a conversation. Responses were dichotomized, so that individuals who responded they could use one or both languages well enough were assigned a one, and individuals who spoke neither well enough were assigned a zero. As in Chiswick and Miller's (2001) study, L1 density was the proportion in a metropolitan area or province with the same L1. The results generally matched those of the United States and Australian analyses, with a negative effect of age of arrival, a positive linear but negative squared effect of length of residence, and a negative effect of L1 density. The model included interactions between L1 density and other individual variables, and the signs of the interactions matched the signs of the main effects: a negative interaction of L1 density and age of arrival, and a positive interaction of L1 density and length of residence.

In another multilingual context, van Tubergen and Wierenga (2011) examined the L2 proficiency of 2,250 Turkish and Moroccan male immigrants in Belgium. L2 proficiency was again a self-rated dichotomized scale, but separate measures were created for each combination of language (Dutch and French) and productive language domain (speaking and writing). They approximated L1 density as the percent same-ethnicity in the municipality. For each of the four outcomes, they found negative effects of age of arrival and L1 density, and a positive effect of length of residence with a negative quadratic term.

Raijman (2013) modeled the L2 proficiency of 593 Jewish South African adult immigrants to Israel. Two features of this study are its outcome and L1 density measure. L2 proficiency was again self-assessed, but in three separate dimensions of conversation, reading, and writing abilities, which were combined into a composite score. For the L1 density measure, subjects were asked what percent of their friends were Israeli. Answers were dichotomized into greater or less than 50% to act as an L2 contact variable, where greater values indicated greater contact with the L2. In addition to replicating the findings of a negative effect of age of arrival and a positive effect of length of residence, this study also found a positive effect of the L2 contact variable.

Raijman et al. (2015) also examined the L2 proficiency of 3,373 adult immigrants to Israel, this time from multiple places of origin. The outcome was measured in the same way, a composite self-rated L2 proficiency variable. The authors took another unique approach to the L1 density variable, and it was an indicator of whether one had at least one Israeli friend, making this another dichotomous L2 contact variable. Age of arrival had a negative effect on L2 proficiency, and length of residence and L2 contact had positive effects.

Struck's (2020) study of immigrants to Australia was presented in Chapter 1 but is summarized again here for comparison with the studies just discussed. In this study, subjects were 33,859 adult immigrants to Australia. L2 proficiency was measured as latent ability estimates derived from an item response theory model that used teacher-assessed competency achievement records. L1 density was the proportion of individuals at one's language learning venue who spoke the same L1. This measure is unlike those in all the studies above, which used the proportion of all individuals, regardless of immigrant status or L1, in the denominator of the L1 density calculation. Nevertheless, negative effects of age of arrival and L1 density were again found.

The literature on the L1 density effect in adults establishes a link between L1 density and L2 proficiency, regardless of whether L1 density is operationalized as a proportion of the overall population or of only those non-native speakers in the sample. These studies show consistency in the presence and sign of their effects across contexts and measurements. A common limitation of most of these studies is the use of self-rated proficiency, which is only moderately correlated with objective measures of proficiency (meta-analytic $r = .53$; Masters, 2014).

Table 3

Summary of Studies on L1 Density in Adults

Study	Subjects	Age of arrival (years): mean (SD)	Length of residence (years): mean (SD)	L1 density operationalization	Outcome	L1 density findings
Chiswick and Miller (2001)	32,168 male immigrants to Canada	24.2 (11.5)	19.8 (12.6)	Percent same-L1 in region	Self-report whether official language spoken	Negative main effect
Chiswick and Miller (1992)	23,741 male immigrants to Canada	(Not provided[a])	19.5 (10.6)	Percent same-L1 in region	Self-report whether official language spoken	Negative main effect; positive interaction with length of residence
van Tubergen and Wierenga (2011)	2,250 Turkish and Moroccan male immigrants to Belgium	19.3 (10.1)	18.8 (8.9)	Percent same-ethnicity in region	Self-report whether official language spoken	Negative main effect
Raijman (2013)	593 Jewish South African immigrants to Israel	27.6 (8.3)	18.9 (11.7)	Whether majority of friends were Israeli (native speakers of L2)	Self-rated proficiency composite score	Negative main effect
Raijman et al. (2015)	3,373 immigrants to Israel	41.6 (13.1)	13.2 (4.8)	Whether one had at least one Israeli friend (native speaker of L2)	Self-rated proficiency composite score	Negative main effect
Struck (2020)	33,859 immigrants to Australia	36.2 (12.6)	1.7 (1.0)	Proportion same-L1 among language learners at venue	Ability estimates derived from task-based assessment records	Negative main effect; negative interaction with age

[a]Mean age at data collection was 42.6 (10.5), similar to that of subjects in Chiswick and Miller (2001): 44.0 (10.9). Lengths of residence were also comparable for these two studies.

Studies with Children and Adolescents

The findings of Vervoort et al. (2012) suggest the relationship between L1 density and L2 proficiency is mediated; L1 density correlates with contact with L2 speakers, which in turn correlates with L2 proficiency. Since only two studies have quantified the relationship between L1 density and L2 proficiency in child and adolescent SLA (Ahn & Jepsen, 2015; Friesen & Krauth, 2011), it is useful to also discuss other papers that have studied the association between L2 contact and L2 proficiency in child and adolescent populations (the latter two variables in the causal chain proposed by Vervoort et al., 2012). If such a relationship can be established, then L1 density should correlate with L2 proficiency in child and adolescent populations, if L1 density correlates with L2 contact in these populations. The final three studies reviewed below support the presence of a relationship between L2 contact and L2 proficiency in child and adolescent populations (Carhill-Poza, 2015; Genesee & Hamayan, 1980; Paradis, 2011). These five studies that examined the effect of either L1 density or L2 contact are summarized in Table 4 (p. 33). The relationships between these variables may again be represented with a simplified version of the model proposed by Vervoort et al. (2012). An uncertainty in generalizing this model to younger learners is indicated with the substitution of an arrow for a question mark:

L1 density (?) Contact with L1/L2 → L2 proficiency and use

It is unknown how L1 density relates to L1/L2 contact in a school environment, as discussed below in the section "Differences in L1 Density's Effect Between Adults and Children."

L1 Density and L2 Proficiency. Friesen and Krauth (2011) studied all students in English-speaking schools in British Columbia who were in seventh grade in the years 2002–2004 (N = 139,610), including 9,865 L1 Chinese and 5,076 L1 Punjabi learners of L2 English. The dataset contained both fourth and seventh grade standardized reading and numeracy assessment

scores. They included a measure of L1 density for the L1 Chinese and L1 Punjabi students, calculated as the proportion of same-L1 speakers in one's grade at one's school. Regression models predicted assessment scores from individual variables of gender and L1 indicators, and contextual variables of the percent of peers for gender, Aboriginal status, and L1. Of interest are interaction terms between the L1 indicators and the share of peers speaking different L1s. Adding together this interaction term (e.g., L1 Chinese times Percent Peers L1 Chinese) with the main effect of peer L1 (e.g., Percent Peers L1 Chinese) yields a percent same-L1 variable. Some of the summed coefficients were given in the text, and the other coefficients and their standard errors can be calculated from the model summary tables. Doing so reveals that L1 density's effect was small and statistically non-significant. A 10% increase in L1 density was associated with a 0.2% to 0.9% of a standard deviation decrease in reading scores for the two L1 groups, and a 0.2% decrease to a 2.1% increase on the math assessment, also statistically non-significant. The authors interpreted these findings as evidence for the relative non-importance of L1 density while controlling for group effects, since the L1 indicator variable served as a proxy for socioeconomic status, with the Chinese and Punjabi groups belonging to higher and lower socioeconomic groups, respectively.

An alternative explanation for the lack of an L1 density effect in Friesen and Krauth (2011) can be found in their table of descriptive statistics. Only around 70% of non-native English-speaking students were classified as ELLs in fourth grade (73.6% for L1 Chinese, 69.8% for L1 Punjabi), and this figure dropped to around 20% in seventh grade (24.8% for L1 Chinese, 14.2% for L1 Punjabi). Assuming students exit from an ELL program or are otherwise not classified as ELLs because their level of proficiency in English is deemed sufficient to function in school, then it follows that for these students, the effect of L1 density should have been greatly

reduced since they could have operated in the L2 with all their peers, and may have even preferred to do so.

Ahn and Jepsen (2015) estimated the L1 density effect in middle school student learners of L2 English in North Carolina. The data was longitudinal and included four cohorts of students beginning sixth grade in 2006–2009 and finishing eighth grade in 2009–2012, and it included all students in this time frame. One of their models predicted end-of-grade reading assessment scores for 26,889 students with limited English proficiency (i.e., all classified as ELLs). The L1 density variable was calculated as the percentage of all students in one's grade at one's school who spoke the same home language. After controlling for several grade-level characteristics (e.g., percent limited English proficiency, percent female, percent Black) and student-level characteristics (e.g., previous year's score, free or reduced-price lunch, transfer history), L1 density was found to have a positive effect on reading scores. The effect size was very small; a 10% increase in L1 density was associated with a 0.8% of a standard deviation increase in reading scores. A model predicting mathematics scores found the opposite, a negative effect of L1 density, but the effect size was similarly small (−0.7% of a standard deviation). The authors gave no interpretation of their findings, but they argued earlier in the paper that a positive effect would have resulted from the opportunity to use the L1 with same-L1 peers, and a negative effect would have arisen from decreased incentives to learn the L2.

The findings of Ahn and Jepsen (2015) are difficult to interpret because of the lack of other studies estimating the effect of L1 density and because of the divergent results within the paper with the two outcomes. Neither of the author's hypotheses (an advantage of support within an L1 group or a disadvantage of decreased utility of the L2 with an increased presence of the L1) can simultaneously account for the positive and negative effects of L1 density. In fact, these

hypotheses assume that the two outcomes of reading and mathematics are fundamentally similar, ignoring that reading requires the L2 much more than does mathematics. Furthermore, the study did not control for any common SLA variables like age or length of exposure, outside of the panel fixed effects for students. Putting aside these limitations and the results of the mathematics model, the findings seem to suggest that the support of same-L1 peers outweighs the decreased opportunities for exposure to and interaction in the L2, and the reduced incentives for learning the L2. This would be an extension of Carhill-Poza (2015), whose study is discussed below and who found benefits of academically oriented peers, regardless of their L1. The implication of this study would be a positive interaction between academic orientation and L1, so that as the proportion of same-L1 peers increases, so do scores.

L2 Contact and L2 Proficiency. This discussion now turns to three studies examining the relationship between L2 contact and L2 proficiency. Genesee and Hamayan (1980) studied individual differences in the acquisition of L2 French of 52 children in a French full immersion program in Montreal, Canada. Subjects were all in first grade and had been first exposed to French the year prior in kindergarten. Outcomes included a standardized language test assessing multiple areas (including literacy, semantics, and vocabulary), a listening comprehension test that involved matching aural prompts to pictures, and an interview-based oral production test. Principal component analysis was used to combine predictors. The most relevant of these was a factor reflecting teachers' ratings of students' use of French inside and outside of the classroom and of students' relationships with teachers and peers. Other factors included measures of attitude, personality, and cognition. In multiple regression analyses that controlled for all these factors, the factor of interest was found to be a statistically significant predictor for the standardized language test and listening comprehension test, but not for the oral production test.

Its standardized coefficient was sizable and one of the largest in each model ($\hat{\beta}$s ranged 0.26–0.39).

Paradis (2011), whose study was summarized in the section "Age of Onset and Length of Exposure" and who examined the vocabulary and morphosyntax of young children, also included a "richness of the [L2] English environment outside school" variable in predicting children's L2 vocabulary and morphosyntax (p. 222). This quantitative measure was created by surveying parents about which languages their children used for media and activities and with friends, and how often they participated in these. Responses were used to create a composite score. After controlling for age and length of exposure, a positive effect of the richness of the L2 environment was found for both the vocabulary and morphosyntax outcomes, with standardized coefficients of 0.16 and 0.21, respectively.

Carhill-Poza (2015) examined the effect of peer networks on oral academic English proficiency, as measured by a standardized test. Subjects were 102 immigrant L1 Spanish learners of L2 English who averaged 16.5 years old (SD 1.3) and had lived in the United States for an average of 3.5 years (SD 2.0, range 1–7). Subjects were asked to name peers and indicate the languages used and activities done with them. Peers were coded as "English-oriented" if at least 50% of interactions were in English, and they were counted as "academic" if they engaged in academic activities (e.g., doing homework, studying for tests) together at least a few times a week. In multilevel models, after controlling for age of arrival, length of residence, gender, and maternal education, the number of English-oriented peers was positively associated with academic English proficiency ($\hat{\beta} = 0.25, p < .01$). However, after adding the number of academic peers to the model, the coefficient on English-oriented peers became statistically non-significant while the academic peers' coefficient was statistically significant ($\hat{\beta} = 0.30, p <$

.001). The number of English-oriented peers and academic peers were correlated ($r = 0.59, p <$ .001), and qualitative results and analyses of variance teased apart the interaction between the two types of peers. Both English-oriented and academic peers were associated with higher proficiency, and the combination of the two was associated with the highest levels of academic English proficiency. The findings highlight the importance of both L2 support and academically oriented peers who speak either the L1 or the L2.

An issue with both Genesee and Hamayan (1980) and Paradis (2011) is the use of composite measures, making it impossible to conclude whether L2 proficiency was related to contact with speakers of the L2 or to the other indicators (e.g., media consumption, classroom relationships). A common problem with all three studies is their cross-sectional nature and the fact that L2 proficiency and use are both individual-level variables. The result is that it cannot be known whether it is the proficient learners who use the L2, or those who use the L2 become proficient. To disentangle the two factors of L2 proficiency and use, therefore, it is necessary that the two are handled in such a way that the possibility of causality flowing in one direction can be logically eliminated, either through temporal precedence or through exogenous assignment to different L2 use environments. The current study had both: a longitudinal design and a measurement not of individual L2 use but of prevalence of the L1 among one's peers.

Table 4

Summary of Studies on L1 Density and L2 Contact Effects in Children and Adolescents

Study	Subjects	Language predictor	Outcome	Findings
Friesen and Krauth (2011)	9,865 L1 Chinese and 5,076 L1 Punjabi learners of L2 English	Proportion same-L1 in grade in school	Standardized reading assessment	L1 density not statistically significant
Ahn and Jepsen (2015)	26,889 learners of L2 English	Percentage same-L1 in grade in school	Standardized reading assessment	Small, positive effect of L1 density
Genesee and Hamayan (1980)	52 learners of L2 French	Teachers' ratings of students' L2 use and relationships	Standardized language test, listening comprehension test, oral production test	Sizable, positive effect of L2 use and relationships for language and listening test models, but not for oral production
Paradis (2011)	169 learners of L2 English	Parents' ratings of children's L2 use	Vocabulary and morphology measures	Positive effect of L2 use for both outcomes
Carhill-Poza (2015)	102 L1 Spanish learners of L2 English	Numbers of English-oriented peers and academic peers	Standardized speaking assessment	Positive effects of both peer types

Differences in L1 Density's Effect Between Adults and Children. While the L1 density effect is robust in adult learners, the question remains whether it can explain some variance in child and adolescent SLA. Two differences between these populations prevent a naïve generalization of the findings with adults to younger L2 learners. First, children are unable to self-select into wholly L1 environments since they are (assumedly) attending a school where subjects are taught in the L2. This means the L2 has some nonzero minimum value, in contrast to situations where adults can live in environments where they can exclusively use their L1 for their daily needs. That is, children have some incentive to learn the L2 regardless of their peers' L1s. Second, children of different ages may show different patterns of language preference and use, following Jia and Aaronson (2003) who found older child and adolescent immigrants preferred using the L1 and had less L2 contact than did younger arrivals. Therefore, the effect of L1 density on L2 proficiency in child and adolescent populations may vary as a function of age. The only two studies that analyzed the effect of L1 density used data from similar ages: grades four and seven (Friesen & Krauth, 2011) and grades six through eight (Ahn & Jepsen, 2015). The current study expanded these ranges to cover all grades one through twelve, and in contrast to these two papers, it included important individual-level variables and their interactions with L1 density to capture L1 density effects across childhood and adolescence. Additionally, these two studies' only language-related outcome was reading comprehension, but the current study had four outcomes, adding the other three language domains of listening, writing, and speaking.

It may be that L1 density's effect differs between older and younger populations, and a weaker or no relationship between L1 density and L2 proficiency is observed in children and adolescents. This could signal one of several possibilities, including that contact with the L2 is

invariant in a school context, or that the path between L1 density and L2 proficiency is mediated or moderated by other variables.

Summary and Remaining Questions

This review of child and adolescent SLA research can be summarized with three pairs of general conclusions and questions. First, age of onset and length of exposure interact to favor older starters with a rate advantage, but the stability of these effects is unclear, especially given that language preference may vary as a function of age. Second, SLA outcomes vary between learners of different L1 backgrounds, but it is unknown whether these findings generalize to individuals of other L1 backgrounds and to more global language outcomes. Third, while L2 contact seems to predict L2 proficiency, questions remain whether L1 density predicts L2 contact and proficiency in a school environment, and how L1 density moderates the effects of the other individual variables of age of onset and length of exposure. The current study addressed these exact questions.

Chapter 3: The Current Study

The current study's objective was to begin addressing some of the gaps in the literature on child and adolescent SLA by modeling the effects of age, length of exposure, language similarity, and L1 density. Features of this study include its large sample size of thousands of subjects, the measurement of language proficiency with a standardized outcome, the longitudinal nature that allowed analyses of effects over time, a language similarity measure in combination with a wide variety of L1s represented, and the inclusion of an L1 density measure with children and adolescents, one of the first few of its kind.

The data for this study was deidentified assessment records from World-Class Instructional Design and Assessment's (WIDA) Assessing Comprehension and Communication in English State-to-State (ACCESS) for ELLs Online test (henceforth "ACCESS test", putting aside the paper-based version of the test, which is not in this sample). The data was an approximately one-percent multistage, stratified, random sample. WIDA's data manager chose fifteen states and then randomly sampled students within states. The original sample contained about 100 students from 10 grades from 15 states, each of whom was observed for three years, for a total of 14,998 students (44,994 observations). The data that was analyzed was less than this, as some cases were dropped as described in the section "Preparation."

The ACCESS Test

The information that follows is summarized from technical reports for the dataset's three test administration years: 2017–2018, 2018–2019, and 2019–2020 (Center for Applied Linguistics, 2019, 2020, 2021). Each year, approximately 1.5 million students in grades one through twelve in at least 38 states and territories in the United States take the ACCESS online test. The students, who are placed into the English language programs where they take the

ACCESS test, are identified as ELLs through the use of English language proficiency screeners and home language surveys (Kim et al., 2018). The ACCESS test is a standards-based, task-based, computer-based, multistage adaptive test of English language proficiency intended to measure social and academic English of grade school ELLs in a school context. The test is task-based in the sense that, following the definition of Norris (2016), questions are situated within contexts, and following others (Bachman, 2007; Brindley, 2013), tasks are used to make inferences about the test taker's language abilities.

Format and Content

The test has four sections corresponding to four language domains: listening, reading, writing, and speaking (Center for Applied Linguistics, 2019, 2020, 2021). Students' scores on the ACCESS test are used for progress tracking, decisions about exiting ELLs from their ELL programs, and accountability. Test items are situated into one or more contexts of social and instructional English and academic English for math, science, language arts, and social studies. The listening section plays recorded stimuli, and the item types include multiple-choice (most common), drag-and-drop (drag an image or text to an area of the screen), and hot spot (click on an area of the screen). The reading section is similar, but the stimuli are written texts rather than recordings. The writing section provides a prompt to which students provide constructed responses. The speaking section involves a virtual test administrator and interlocutor, and a model student who models responses. Stimuli are presented aurally and visually, and students answer by recording spoken responses.

WIDA provides sample items for each tier and domain of the ACCESS test (WIDA, n.d.). Example test items for each domain include selecting pictures to answer questions about details in a recorded story about a child taking a trip (listening), answering questions about

student government election processes given a recorded dialogue and an on-screen chart (listening), selecting items under a certain price given information on a sign at a school fair (reading), translating a paragraph about distances on a baseball field into mathematical equations (reading), describing how to perform a science experiment given a picture of equipment and a data collection table (writing), describing what is happening in a series of pictures of somebody encountering unexpected weather (writing), making predictions about what will happen next in a story about a family who encountered a problem while grocery shopping (speaking), and summarizing points from a recording about a politician's career (speaking). Apparent in this list are aspects of the test already mentioned. The test items are contextualized tasks, and the construct being measured is social and academic English, as seen in the various contexts in which the items are situated, from grocery stores to science labs. The items are integrated, so that items in one domain may require the use of other language domains (e.g., read a passage and then write a response). Pictures of sample practice items are provided in Appendix A.

Scoring

Listening and reading items are machine-scored dichotomously, and a Rasch model is used to calculate students' ability estimates (Center for Applied Linguistics, 2019, 2020, 2021). Writing and speaking items are scored polytomously by the Data Recognition Corporation's trained human raters, and ability estimates are calculated using a Rasch-grouped rating scale. (The scoring model accounts for items grouped by task in the writing and speaking sections.) Ability estimates are then scaled to obtain scale scores via scaling equations so that mean scores are around 300–400 with standard deviations of around 40–50 (see sample descriptives in Table 5, p. 44). Overall composite scale scores are weighted averages of the four sections' scores, where more weight is given to receptive skills: 35% listening, 35% reading, 15% writing, 15%

speaking. All students' scale scores are on a single interval scale and are directly comparable across grade levels, test forms, and test administrations. Some items are repeated across administrations to equate scores across years (i.e., common-item equating). However, scale scores are not comparable between domains (WIDA, 2022).

Reliability estimates are reported by grades, genders, ethnicities, and individualized education plan status. Listening and reading ability score reliabilities generally ranged 0.80 to 0.90 across administrations. Writing Cronbach's alphas were also mostly in the range 0.80 to 0.90, except for the 2019–2020 higher-tier writing forms where alphas ranged 0.56 to 0.70. The low reliability was explained as a function of a change in the section length; writing sections had three or four items in previous years, while the 2019–2020 test had only two items. Speaking Cronbach's alphas generally ranged 0.80 to 0.85. Reliabilities for composite scores were calculated using Cronbach's alpha for the four domain scores, weighted by each section's contribution to the overall score, and these alphas were all around 0.95.

Data

Preparation

Value Modifications. After the data was obtained from WIDA (N = 14,998 students, 44,994 observations), it was prepared for analysis by modifying records, removing records, and calculating and adding new variables (which also included the removal of some records). Values for some students' records were modified. Where the year of testing did not match the academic year, it was changed to the academic year, affecting 17 students' records. Where birth dates and enrollment dates differed within an individual student's records, the first record's value was used for all records, changing three students' birth dates and 19 students' enrollment dates. Where grades did not increase by one each year, they were changed to follow such a sequence, for one

student. If birth date, enrollment date, or L1 were missing for any records, the value for other records was used; this step affected 183 students' birth dates, 3,642 students' enrollment dates, and 1,995 students' L1s. L1 was wholly missing for 9 students, all of whom attended different schools, and these were assigned an "other" L1 that was just their ID number.

Case Deletions. For various reasons, 802 students' records (5.3% of dataset) were wholly dropped from the dataset. This was done when the enrollment date was wholly missing (47 students), the enrollment date was beyond the date of data collection (one student), the L1 had multiple values within a single student's records (69 students), the age of enrollment was less than four years old (656 students), or the date of the first test was before the enrollment date (29 students). The rationale for excluding these cases is that the enrollment date was necessary for calculating age of enrollment and the L1 for calculating L1 density, and that four is a possible cutoff for the beginning of *second* language acquisition, given the previous discussion of age of onset.

Variable Additions. Several variables were added to the dataset either by using variables already present in the dataset to calculate new ones (L1 density, age of enrollment, age at testing, and length of enrollment) or by joining a variable from another external dataset (language similarity). L1 density was the percentage of individuals who spoke a given L1 in a district in a certain year. It was calculated by first summing the number of speakers of each L1 in each district in each year, dividing that number by the total number of students in that district in that year (in the sample), and then multiplying it by 100, so that it could range 0–100. For example, if district A had 90 students in the sample in 2017–2018, of which 63 spoke L1 X and 18 spoke L1 Y (and the remaining nine spoke other L1s), then all speakers of L1 X at that district would have an L1 density of 70 (63/90*100) for that year, and all speakers of L1 Y would have an L1 density

of 20 (18/90*100). Because these are percentages, these numbers would remain the same if everything were scaled linearly, such as multiplying all the numbers by 10, resulting in 900 students in the sample at the district and 630 speakers of L1 X and 180 of L1 Y. This calculation also assumes the relative proportions of each L1 of the students in the sample are representative. Furthermore, the L1 density statistic is only meaningful if the number of students in a district in a year is more than one. The statistic increases in precision with a higher cutoff for the minimum number of students in a district in a year, due to the nature of random sampling. For this reason, a minimum district size of five was used, so that any students who did not attend a district with at least five students every year they were observed were dropped from the dataset.

Age of enrollment was the difference between the birth date and the enrollment date, age at testing was the difference between the birth date and the test date, and length of enrollment was the difference between the enrollment date and the test date. The enrollment date was when a student first enrolled in a school in the United States, after which they entered an ELL program. Because only those who were in an ELL program were included in this dataset, and because school administrators make decisions about ELL program placement based on L2 proficiency exams and home language use, the enrollment date was considered to be the date of onset to L2 English. Therefore, age of enrollment and length of enrollment are proxies for age of onset and length of exposure, respectively.

Language similarity statistics were obtained from the Gateway Languages Database (Gnanadesikan & van Rossum, 2016). Lists of languages were extracted from the ACCESS data and from the Gateway Languages Database. Where a match or close equivalent existed, the similarity of that language to English was associated with that L1 in the ACCESS data. If an individual's L1 was recorded as just "Arabic" or "Chinese" without indication of a specific

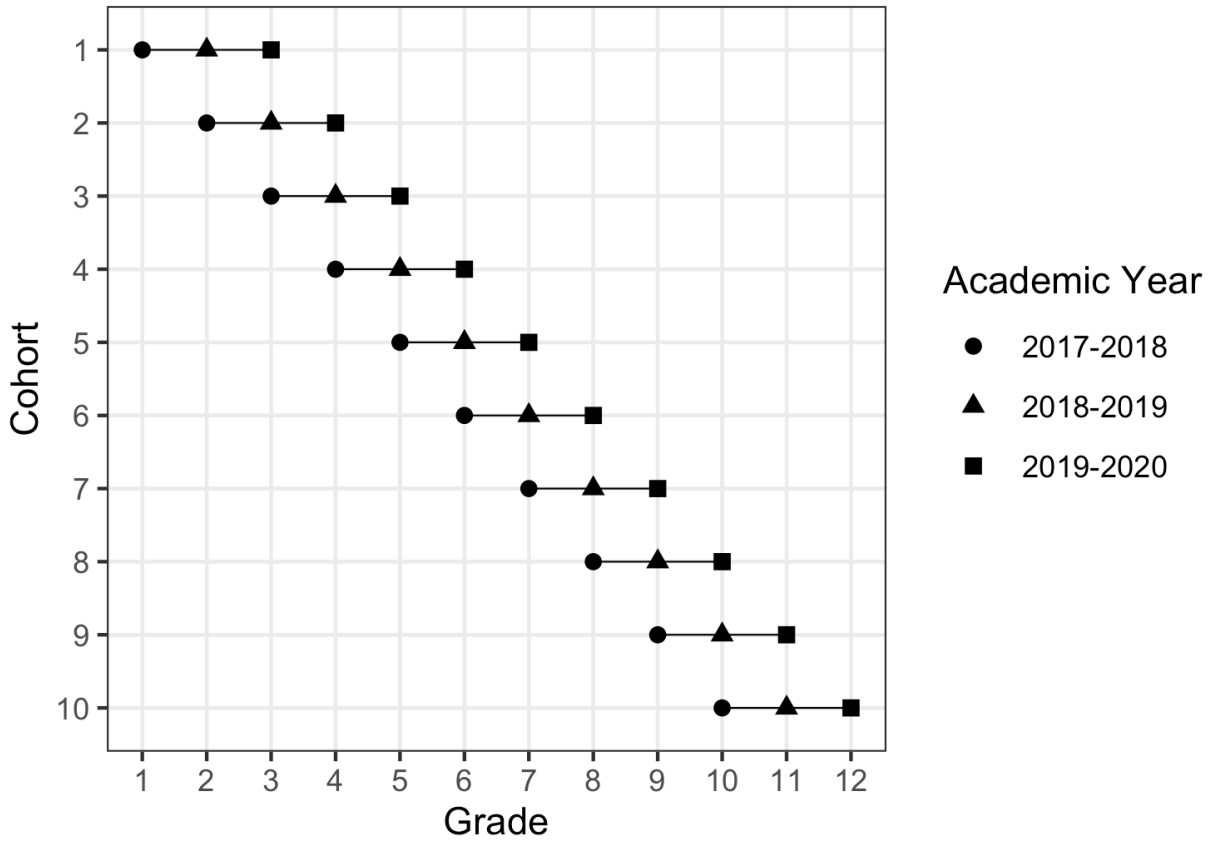
dialect or if it was a dialect of these two not found in the database, they were assigned the language similarity score of the dialect with the most speakers according to the database, Egyptian Arabic and Mandarin Chinese, respectively. The Gateway Languages Database did not have a match for each L1 in the dataset, so only students whose L1s have a match could be used. Applying the two criteria from this section to the dataset (minimum district size of five, L1 in the Gateway Languages Database) further reduced the dataset by about one fourth (3,317 students dropped; 23.4% of dataset after initial case deletions).

Descriptive Statistics

After preparation, the analytic sample with a minimum district size and language similarity statistics had 10,879 students with 32,637 records. The data was for academic years 2017–2018, 2018–2019, and 2019–2020. Following the definition of Ross and Masters (2023), this data was longitudinal because SLA was assessed on multiple occasions at least 10–12 weeks apart. The average gap between tests was 1.00 years (SD 0.04, range of 0.15–1.82). The sample consisted of ten cohorts of students who were in grades 1–10 in the first year, 2–11 in the second year, and 3–12 in the third year. This is an accelerated longitudinal design, a hybrid of cross-sectional and longitudinal designs where subjects are observed for overlapping periods that together form an age range longer than the length of data collection (i.e., 12 years of data collected in three years). The data structure is visualized in Figure 1.

Figure 1

Accelerated Longitudinal Design of ACCESS Dataset



Descriptive statistics for the analytic sample are in Table 5, where variables are denoted as belonging either to the observation or student level. Observation-level variables could vary within students. For example, the length of enrollment increased each year, and L1 density, although it is conceptually a contextual variable, could take on different values every year as the district composition varied. Each student had only one value for each student-level variable; the two student-level variables, age of enrollment and language similarity, did not vary within students. The analyses used all the variables in Table 5 except for age at testing. This variable could not be included in models with the other variables since it was perfectly collinear with age of enrollment and length of enrollment as the sum of these two.

Table 5

Descriptive Statistics for Analytic Sample from ACCESS Dataset

Variable	Mean	SD	Min	Median	Max
Observation Level					
Length of Enrollment	4.82	2.84	0.00	4.42	14.51
Age at Testing	12.67	3.16	6.23	12.61	25.71
Listening Score	386.18	54.43	108.00	391.00	529.00
Reading Score	351.42	40.72	198.00	352.00	492.00
Speaking Score	304.90	51.55	106.00	310.00	476.00
Writing Score	332.82	42.89	111.00	335.00	508.00
L1 Density in District	70.09	31.97	0.11	83.67	100.00
Student Level					
Age of Enrollment	7.85	3.41	4.00	5.89	22.76
Language Similarity	66.85	10.84	24.00	71.00	86.00

In the analytic sample, 62 L1s were represented. The most common of these was Spanish, and its 8,882 speakers made up 82.3% of the sample. The next nine most common L1s in order of frequency were Arabic (411 students, 3.8% of the analytic sample), Somali (193, 1.8%), Vietnamese (148, 1.4%), Mandarin Chinese (130, 1.4%), Swahili (110, 1.0%), French (108, 1.0%), Amharic (86, 0.8%), Portuguese (80, 0.7%), and Urdu (77, 0.7%).

The students and records were distributed across 4,530 schools, 332 districts, and 15 states. Table 6 has statistics for the sizes of these different units. Note that every student had three observations (one per year), and that the minimum number of students per district was five.

Table 6

Counts by Grouping Units in Analytic Sample of ACCESS Dataset

Variable	Mean	SD	Min	Median	Max
Districts per State	22.13	11.26	4	24	42
Schools per State	302.00	166.51	71	285	685
Schools per District	13.64	20.04	1	7	170
Students per State	725.27	197.49	153	765	940
Students per District	34.19	75.72	5	13	913
Students per School	3.71	5.72	1	2	104
Obs. per State	2175.80	592.47	459	2295	2820
Obs. per District	98.30	223.93	5	37	2731
Obs. per School	7.20	12.54	1	4	278
Obs. per Student	3.00	0.00	3	3	3

Research Questions

This study was designed to address seven research questions:

1. What is the effect (regression coefficient) of age of enrollment on L2 proficiency, as measured by ACCESS test scores?
2. What is the effect of length of enrollment on L2 proficiency?
3. What is the effect of the interaction between age of enrollment and length of enrollment on L2 proficiency?
4. What is the effect of language similarity on L2 proficiency?
5. What is the effect of L1 density on L2 proficiency?
6. How does L1 density moderate the effects of age of enrollment and length of enrollment on L2 proficiency?
7. How do the answers to questions one through six vary by language domain?

Each question was answered by fitting multilevel models to the data and interpreting both statistical significance and effect size. Hypotheses of predicted effects corresponding to each of the research questions were offered as follows:

1. Age of enrollment would have a positive effect since older starters were more likely to have had previous instruction in L2 English before enrollment.
2. Length of enrollment would have a positive linear effect since it represented opportunities for exposure to and interaction in the L2, but a negative quadratic effect so that the rate of SLA slowed down with time (Chiswick & Miller, 1992, 2001).
3. Age of enrollment and length of enrollment would have a positive linear interaction due to older starters' rate advantage, but a negative quadratic interaction so that younger starters would eventually catch up to later starters.
4. Language similarity would have a positive effect, but the size of the effect would be small. The hypothesis of a small effect as it was phrased was non-falsifiable, but it flowed from research suggesting a weak relationship between language similarity and general language skills, such as those measured by the ACCESS test.
5. L1 density would have a negative effect on the outcome. Although the L2 would have some nonzero minimum value in a school context, exposure to the L2 would still have varied as a function of peers' L1s.
6. L1 density's interactions:
    a. The interaction of L1 density and age of enrollment would be negative, so that the L1 density effect would be more pronounced among older enrollees, who may have preferred the L1 over the L2 and would use the L1 if the opportunity existed.

b. The interaction of L1 density and length of enrollment would be negative, with the result that length of enrollment's effect would decrease as the proportion of same-L1 peers increased.

7. The question of differences in effects by domain was exploratory, and no specific hypotheses were offered due to the lack of previous research on the topic. This research question was motivated by a study that found L2 use predicted scores in some domains but not others (Genesee & Hamayan, 1980).

Chapter 4: Analysis

The seven research questions for this study were answered by fitting multilevel models to the dataset, predicting ACCESS test scores from the variables of interest. This chapter is divided into two parts: model details and results. The first part discusses issues of variable centering, model fitting, model form, and regression diagnostics. The second part contains tables of fitted model estimates and then discusses the statistical significance and magnitudes of the estimates. The next chapter, Chapter 5, answers this dissertation's research questions and relates the results to the literature on child and adolescent SLA reviewed in Chapter 2.

Model Details

Centering

Centering of variables is a critical issue in multilevel modeling (Brincks et al., 2017; Enders & Tofighi, 2007). Variables may be centered in one of two ways, either by subtracting the mean value of that variable within the group (centering within cluster, or CWC), or by subtracting the overall mean value of that variable across all groups in the entire sample (centering at the grand mean, or CGM). Different research questions require different centering approaches, depending on the level of the effect of interest. CWC is needed to estimate the effects of level-one variables and cross-level interactions, and CGM is needed to estimate the effects of level-two variables while controlling for level-one variables. The reason for this difference is that CWC removes between-group variance to allow for the estimation of the within-group effect. CGM, on the other hand, results in a mixture of within- and between-group variance, so that it cannot be used to estimate within-group effects, but it can be used to control for a level-one variable when estimating the effect of a level-two variable.

An example scenario can illustrate these different centering approaches. In this example, patients are nested within physicians, and physicians are nested within hospitals. This is a three-level model. The outcome is patient satisfaction at level one. Two predictors of interest are patient age at level one and physician experience at level two. Physicians may vary in the average age of their patients, and hospitals may vary in the average level of experience of their physicians. To estimate the unique effect of patient age on patient satisfaction, each patient's age must be centered within physician (CWC), so that zero represents an average age for that physician. Choosing to center by subtracting the mean age of patients in the entire sample (CGM) will result in a variable that is a combination of within-physician, between-physician, and between-hospital variance in age. However, to estimate the unique effect of physician experience on patient satisfaction, CGM must be applied to patients' ages to control for the patient-age effect. In either case, physician experience must be centered within-hospital (CWC) to control for between-hospital differences in average physician experience.

This example reveals the important fact that the unique effects of patient age and physician experience cannot be estimated in a single model because they require the level-one variable (patient age) to be centered in different ways. Different models are needed to answer different research questions.

In the current study, the three levels were observations (level one), students (level two), and states (level three). Two predictors of test scores were at level one (length of enrollment and L1 density), and two were at level two (age of enrollment and language similarity). To estimate the effects of the two level-one variables, these variables needed to be centered within students (CWC), and CWC was also needed to estimate cross-level interactions (age of enrollment with length of enrollment, and age of enrollment with L1 density) and the within-level-one interaction

of length of enrollment with L1 density. Estimation of the effects of the level-two variables required the level-one variables to be centered by their overall means (CGM). In both models, the level-two variables were centered within states (CWC) to control for between-state differences in age of enrollment and language similarity. To help the reader, each table of model coefficients contains columns denoting the level to which each effect belonged and the model where that effect was estimated (either CWC or CGM at level one).

Model-Building Procedure

Before fitting models to address the research questions, null models with only random intercepts for students and states were fit to calculate the intraclass correlations (ICCs) to justify the multilevel modeling approach. The level-two ICCs were calculated using the method of Davis and Scott (1995), where the level-two random intercept variance was divided by the total variance, just as the level-three ICCs were calculated by dividing the level-three random intercept variance by the total variance. The ICCs for four null models are given in Table 7. The nonzero ICCs evidenced that some variance could be accounted for by group effects, so multilevel models were needed to control for dependence among observations.

Table 7

Intraclass Correlations by Level

Outcome	Level-two ICC	Level-three ICC
Listening	.438	.027
Reading	.548	.029
Speaking	.396	.022
Writing	.510	.017

Models were fit in R (version 4.2.1; R Core Team, 2022) with the lme4 package (version 1.1-29; Bates et al., 2015), and *p*-values were obtained using the Satterthwaite method for the degrees of freedom with the lmerTest package (version 3.1-3; Kuznetsova et al., 2017).

Maximum likelihood was used to obtain model estimates, and the BOBYQA optimizer was used to aid in convergence.

Two models (one CWC and one CGM) were fit to each outcome for a total of eight models. These eight models were then fit to two alternate subsamples, for a total of 24 models. Models fit to the analytic sample were of primary interest and are presented in the main body of this dissertation. Descriptive statistics for the two alternate sample constructions are given in Appendix B, and the model estimates for these samples are given in Appendix D and discussed later in this chapter where relevant.

The analytic sample (N = 32,637) was created from the full sample (N = 42,588) by imposing a minimum district size and requiring that each L1 had a language similarity statistic from the Gateway Languages Database. The first three research questions concerning the effects of age of enrollment, length of enrollment, and their interaction could be answered without these constraints. To make the most use of the available data, a set of models was fit to the full sample, and estimates are discussed in the section "Age of Enrollment and Length of Enrollment."

The analytic sample's L1 density statistic was calculated at the district level, which averages over differences between schools. To assess whether the findings were sensitive to the high level at which L1 density was calculated, another smaller dataset was created where each school had a minimum of five students every year. This dataset, termed the school-L1-density sample, had 9,933 observations. Models fit to this dataset used an L1 density statistic calculated at the school level, and these estimates are discussed in the section "L1 Density."

Random slopes were included through a forward testing procedure (Hox et al., 2018). This allowed the inclusion of two random slopes for most models: a random slope of length of enrollment at level two, and a random slope of age of enrollment at level three. These random

slopes allowed individuals to vary in their growth trajectories over time and for age of enrollment's effect to vary across states. In two of the 24 models (the reading CGM model with the analytic and full samples), the inclusion of the length of enrollment random slope was not supported, leaving only the one random slope of age of enrollment at level three in these models.

Model Equations

Table 8 and Table 9 give the model equations for the models fit to the analytic sample in the notation of Hox et al. (2018), and equations for the models fit to the alternate subsamples are in Appendix C. In this notation, γ is used for fixed coefficients, u for random coefficients, and e for level-one residuals. Subscripts denote the level of each term: observation (i), student (j), and school (k). These tables have three columns; in each one, the left column contains the equation terms, the middle column gives the meaning of each term or set of terms, and the right column denotes the level of each term. The difference between the two tables is the centering approach they reflect. Table 8 contains the equations for the CWC model, where level-one predictors were centered at their group means, denoted by the subscript ".jk". Table 9 has the equations for the CGM model, where level-one predictors were centered at the grand mean, denoted by the subscript "…", and the cross-level interactions were omitted. The interactions were omitted with the CGM models since research questions involving cross-level interactions were answered in the CWC models.

Table 8

CWC Model Form

Equation term	Meaning	Level
$\text{Score}_{ijk} =$	Test score	Observation
$\gamma_{000} +$	Intercept	State
$\gamma_{010}\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) +$	Age of enrollment	Student
$\gamma_{020}\left(\text{LanguageSimilarity}_{jk} - \overline{\text{LanguageSimilarity}}_{.k}\right) +$	Language similarity	Student
$\gamma_{100}\left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{.jk}\right) +$ $\gamma_{200}\left(\text{LengthEnrollment}_{ijk}^2 - \overline{\text{LengthEnrollment}}_{.jk}^2\right) +$	Length of enrollment (linear and squared)	Observation
$\gamma_{110}\left(\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) \times \left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{.jk}\right)\right) +$ $\gamma_{210}\left(\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) \times \left(\text{LengthEnrollment}_{ijk}^2 - \overline{\text{LengthEnrollment}}_{.jk}^2\right)\right) +$	Interaction of age of enrollment and length of enrollment (linear and squared)	Cross-level (observation and student)
$\gamma_{300}\left(\text{L1Density}_{ijk} - \overline{\text{L1Density}}_{.jk}\right) +$	L1 density	Observation
$\gamma_{310}\left(\left(\text{L1Density}_{ijk} - \overline{\text{L1Density}}_{.jk}\right) \times \left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right)\right) +$	Interaction of L1 density and age of enrollment	Cross-level (observation and student)
$\gamma_{400}\left(\left(\text{L1Density}_{ijk} - \overline{\text{L1Density}}_{.jk}\right) \times \left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{.jk}\right)\right) +$	Interaction of L1 density and length of enrollment	Observation
$u_{00k} +$	Random intercept	State
$u_{0jk} +$	Random intercept	Student
$u_{01k} +$	Random slope of age of enrollment	State
$u_{1jk} +$	Random slope of length of enrollment	Student
e_{ijk}	Residual	Observation

Table 9

CGM Model Form

Equation term	Meaning	Level
$\text{Score}_{ijk} =$	Test score	Observation
$\gamma_{000} +$	Intercept	State
$\gamma_{010}\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) +$	Age of enrollment	Student
$\gamma_{020}\left(\text{LanguageSimilarity}_{jk} - \overline{\text{LanguageSimilarity}}_{.k}\right) +$	Language similarity	Student
$\gamma_{100}\left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{...}\right) +$ $\gamma_{200}\left(\text{LengthEnrollment}^2_{ijk} - \overline{\text{LengthEnrollment}^2}_{...}\right) +$	Length of enrollment (linear and squared)	Observation
$\gamma_{300}\left(\text{L1Density}_{ijk} - \overline{\text{L1Density}}_{...}\right) +$	L1 density	Observation
$\gamma_{400}\left(\left(\text{L1Density}_{ijk} - \overline{\text{L1Density}}_{...}\right) \times \left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{...}\right)\right) +$	Interaction of L1 density and length of enrollment	Observation
$u_{00k} +$	Random intercept	State
$u_{0jk} +$	Random intercept	Student
$u_{01k} +$	Random slope of age of enrollment	State
$u_{1jk} +$	Random slope of length of enrollment	Student
e_{ijk}	Residual	Observation

Diagnostics

Random effects (u) were assumed to have multivariate normal distributions with other random effects of the same level and be independent of fixed effects (γ) at any level (Snijders & Bosker, 2012). Residuals (e) were assumed to have mean zero, be normally distributed with constant variance, and be independent of fixed effects and other residuals. Over 2,700 diagnostic

plots of model residuals were produced to check these assumptions. None of these plots indicated assumption violations that would threaten the credibility of model results.

Results

Fitted Model Estimates

Model estimates are separated into two tables for space reasons. Table 10 presents estimates from the listening and reading models, and Table 11 presents estimates from the speaking and writing models. Each table has horizontal lines dividing it into three sections: fixed effects, random effects, and R^2 values. The key portion of each table is the fixed effect estimates. The random effects give additional information about the model specifications but are typically left uninterpreted. The R^2 values indicate how much variance in the outcome is explained.

Variables at different levels require different centering approaches for their effects to be estimated. Under each outcome heading, two columns give the estimates from CWC and CGM models. In the section with fixed effects, the two rightmost columns indicate the level of each variable and the model where the unique effect of that variable was estimated, following guidelines on centering in multilevel models (Brincks et al., 2017; Enders & Tofighi, 2007). The age of enrollment and length of enrollment variables can serve as examples in how to interpret the tables. The rightmost column indicates that age of enrollment's effect was estimated with CGM. In the listening CGM model, the effect estimate for age of enrollment was 3.78 and was statistically significant with $p < .001$ (as denoted by three asterisks; see table note). Length of enrollment was a level-one variable, so CWC was needed to estimate its effect. The listening CWC model effect estimate was 36.59, which was also statistically significant. All level-one variables and cross-level interactions had their effects estimated in the CWC model, while the

effects of the two level-two variables (age of enrollment and language similarity) were estimated in the CGM models.

Table 10

Model Estimates for Listening and Reading Scores in Analytic Sample (N = 32,637)

	Listening		Reading			
Variable	CWC	CGM	CWC	CGM	Variable level	Effect estimated in
Fixed Effects						
Intercept	385.92 (2.29)***	385.82 (1.98)***	351.35 (1.77)***	352.08 (2.23)***		
Age of Enrollment	−0.42 (0.21)	3.78 (0.21)***	2.48 (0.27)***	6.91 (0.23)***	Level 2	CGM
Length of Enrollment	36.59 (0.58)***	29.66 (0.40)***	27.31 (0.36)***	20.20 (0.25)***	Level 1	CWC
Length of Enrollment Squared	−2.07 (0.06)***	−1.70 (0.03)***	−1.31 (0.04)***	−0.82 (0.02)***	Level 1	CWC
Age of Enrollment * Length of Enrollment	−1.61 (0.16)***		−0.30 (0.10)**		Cross-Level (1 and 2)	CWC
Age of Enrollment * Length of Enrollment Squared	−0.06 (0.02)**		−0.09 (0.01)***		Cross-Level (1 and 2)	CWC
Language Similarity	−0.14 (0.04)***	0.08 (0.05)	−0.11 (0.03)***	−0.00 (0.04)	Level 2	CGM
L1 Density	0.03 (0.10)	−0.17 (0.02)***	0.06 (0.06)	−0.13 (0.01)***	Level 1	CWC
L1 Density * Age of Enrollment	0.06 (0.03)		0.03 (0.02)		Cross-Level (1 and 2)	CWC
L1 Density * Length of Enrollment	−0.08 (0.14)	0.01 (0.00)*	0.07 (0.09)	−0.00 (0.00)	Level 1	CWC
Random Effects						
Level 2 SD(Intercept)	38.78	30.45	30.88	25.26		
Level 2 SD(Length of Enrollment)	15.32	4.48	9.85			
Level 2 Cor(Intercept, Length of Enrollment)	−0.20	−0.49	0.21			
Level 3 SD(Intercept)	8.68	7.51	6.73	8.58		
Level 3 SD(Age of Enrollment)	0.63	0.65	0.96	0.79		
Level 3 Cor(Intercept, Age of Enrollment)	0.29	−0.00	0.26	−0.25		
Level 1 SD(Residual)	30.76	34.81	18.61	21.37		
R^2 (Fixed Effects Only)	0.09	0.25	0.15	0.40		
R^2 (Fixed and Random Effects)	0.68	0.61	0.79	0.77		

Note. Fixed effect estimates give Estimate (Standard Error) followed by asterisks denoting the significance level: * p < .05, ** p < .01, and *** p < .001.

Table 11

Model Estimates for Speaking and Writing Scores in Analytic Sample (N = 32,637)

	Speaking		Writing			
Variable	CWC	CGM	CWC	CGM	Variable level	Effect estimated in
Fixed Effects						
Intercept	304.58 (1.96)***	304.56 (1.97)***	332.78 (1.44)***	332.99 (1.88)***		
Age of Enrollment	0.38 (0.37)	4.39 (0.32)***	1.75 (0.31)***	5.94 (0.24)***	Level 2	CGM
Length of Enrollment	33.30 (0.58)***	25.06 (0.40)***	30.51 (0.41)***	24.17 (0.30)***	Level 1	CWC
Length of Enrollment Squared	−1.95 (0.06)***	−1.36 (0.03)***	−1.83 (0.04)***	−1.26 (0.02)***	Level 1	CWC
Age of Enrollment * Length of Enrollment	−0.31 (0.16)		−0.42 (0.12)***		Cross-Level (1 and 2)	CWC
Age of Enrollment * Length of Enrollment Squared	−0.16 (0.02)***		−0.13 (0.02)***		Cross-Level (1 and 2)	CWC
Language Similarity	−0.15 (0.04)***	0.11 (0.05)*	−0.08 (0.03)*	−0.03 (0.04)	Level 2	CGM
L1 Density	−0.03 (0.10)	−0.19 (0.02)***	−0.01 (0.07)	−0.08 (0.01)***	Level 1	CWC
L1 Density * Age of Enrollment	0.08 (0.03)**		0.03 (0.02)		Cross-Level (1 and 2)	CWC
L1 Density * Length of Enrollment	0.07 (0.14)	0.01 (0.00)*	−0.07 (0.10)	0.01 (0.00)***	Level 1	CWC
Random Effects						
Level 2 SD(Intercept)	34.60	26.65	32.03	22.88		
Level 2 SD(Length of Enrollment)	12.98	4.67	11.21	4.04		
Level 2 Cor(Intercept, Length of Enrollment)	−0.02	−0.28	−0.18	−0.24		
Level 3 SD(Intercept)	7.41	7.49	5.40	7.20		
Level 3 SD(Age of Enrollment)	1.37	1.15	1.11	0.85		
Level 3 Cor(Intercept, Age of Enrollment)	0.41	0.42	0.07	−0.19		
Level 1 SD(Residual)	32.55	35.19	21.77	24.52		
R^2 (Fixed Effects Only)	0.08	0.22	0.11	0.34		
R^2 (Fixed and Random Effects)	0.60	0.56	0.74	0.70		

Note. Fixed effect estimates give Estimate (Standard Error) followed by asterisks denoting the significance level: * p < .05, ** p < .01, and *** p < .001.

The four outcomes showed general agreement in their patterns of statistical significance and the sign (positive or negative) of the effect estimates where they were statistically significant. The sizes of the coefficients cannot be compared between outcomes since scale scores are only comparable within domains, not across them (WIDA, 2022). For the most part, there were positive effects of age of enrollment and length of enrollment, and several other variables had negative effects: length of enrollment squared, the interaction between age of enrollment and length of enrollment, and the interaction between age of enrollment and length of enrollment squared. The exceptions to these were all in the speaking models, where the interaction of age of enrollment and length of enrollment was not statistically significant, language similarity had a positive effect, and there was a positive interaction between L1 density and age of enrollment. Outside of the speaking model, language similarity and L1 density and its interactions were not statistically significant.

The following discussion is organized by effect, either individual effects (e.g., language similarity) or groups of effects (e.g., L1 density and its interactions). The models fit to other samples are discussed in the relevant sections: the full sample in the section "Age of Enrollment and Length of Enrollment" and the sample with L1 density calculated at the school level in "L1 Density." First, however, the issue of effect magnitudes must be discussed in order to evaluate whether variables' effects were not only statistically significant, but whether their predicted effects on test scores were nontrivial.

Effect Magnitudes

The models were fit to ACCESS test scale scores. The advantages of these scores are that they are on an interval scale and that all grades are on a single continuum. The disadvantages are that their scales are arbitrary and that scores are not comparable across domains (WIDA, 2022).

Additionally, with a large sample size comes enough power to detect effects of little practical significance. In order to judge whether changes in variables are associated with meaningful changes in test scores, discussions of model estimates rely on proficiency levels. In contrast to scale scores, proficiency levels are directly comparable across domains, but they are grade-specific and on an ordinal scale. Proficiency levels range one to six and are derived from scale scores by applying domain-specific, grade-specific cutoffs (Cook & MacGregor, 2017). Proficiency level cutoffs increase across grades, so that a higher scale score is needed to achieve the same proficiency level in higher grades, and that the same scale score is associated with lower proficiency level as grades increase. For example, a scale score of 335 on the reading test is the cutoff for a proficiency level three for grade four but proficiency level two for grade eight, while a score of 366 is needed for proficiency level three for grade eight. Proficiency levels are on an ordinal scale; the between-grade differences between the same proficiency levels decrease with grades, and the within-grade differences between successive proficiency levels are not constant.

Proficiency levels remain important to educators and administrators since these scores are used to compare scores between domains and assess whether students' L2 English proficiency levels are sufficient for them exit from ELL programs. Therefore, it is helpful to use proficiency level cutoffs as a measure of effect size when interpreting the coefficients from models using scale scores. Using the tables provided in Cook and MacGregor (2017), it is possible to calculate median differences between cutoffs between and within grades. The median difference between cutoffs for the same proficiency level between grades is seven scale score points, and the median difference between cutoffs for successive proficiency levels within grades is 32 points. In other words, a student needs to gain seven points on average every year to maintain their proficiency

level between years, and a student needs to gain 32 points on average within a year to increase one proficiency level. Again, proficiency levels are on an ordinal scale, and their cutoff curves differ by domain and vary across grades. Despite these limitations, median differences between cutoffs are useful in interpreting coefficient magnitudes.

Age of Enrollment and Length of Enrollment

This section begins with discussions of the main effects of age of enrollment and of length of enrollment, followed by a discussion of their interaction terms. Age of enrollment was a student-level variable, so its effect was estimated in CGM models. Both length of enrollment variables (linear and squared), varied within students as observation-level variables, so their effects were estimated in CWC models. The cross-level interactions between age of enrollment and length of enrollment were likewise estimated in CWC models. All five of these effects (three main effects and two interactions) were statistically significant for all four domains, except for the interaction of age of enrollment and length of enrollment in the speaking CWC model. Age of enrollment and length of enrollment had positive effect estimates, while length of enrollment squared and the interactions between age of enrollment and the two length of enrollment variables were all negative.

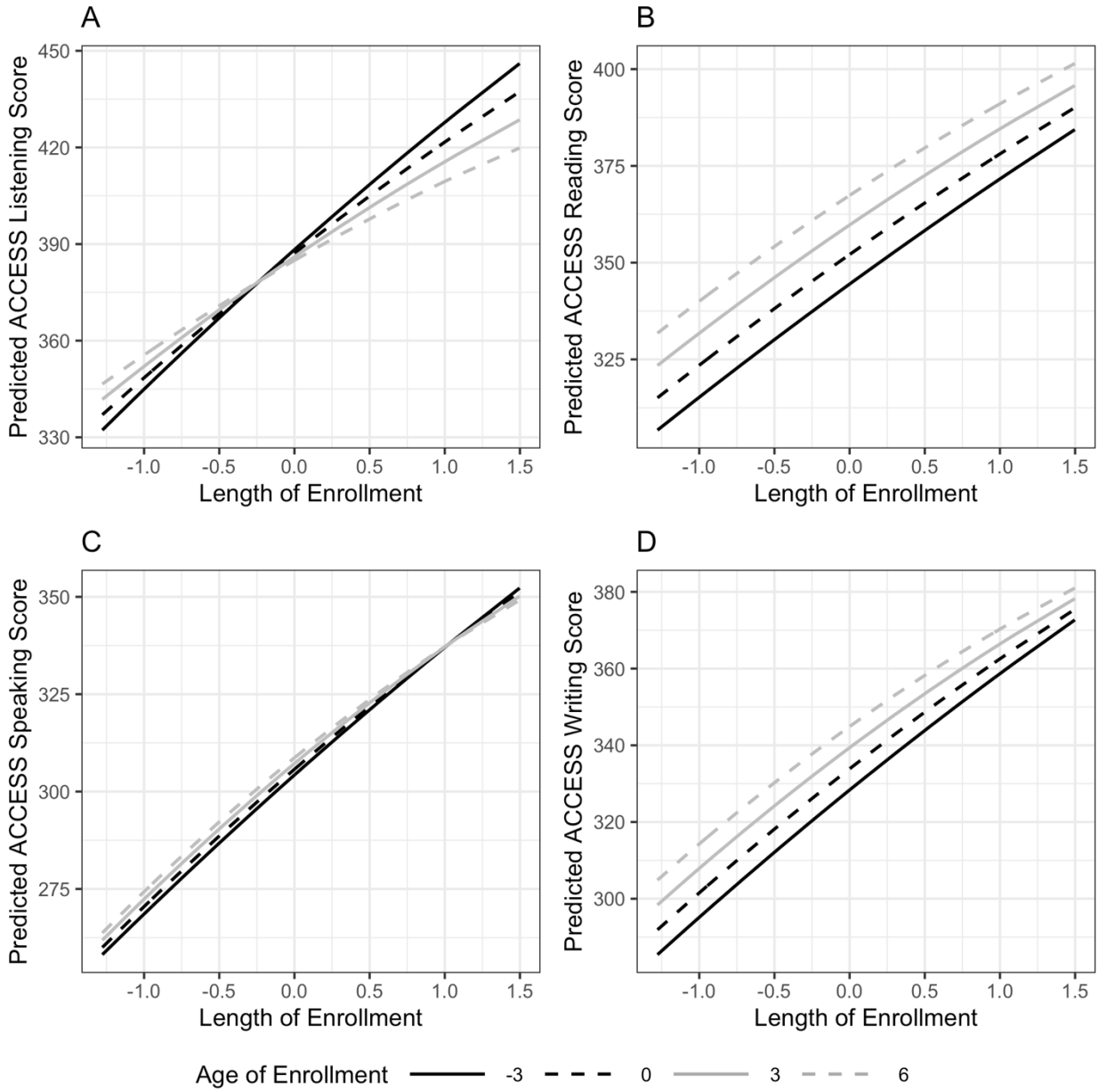
An increase of one year in age of enrollment was associated with about a five-point increase in scale scores (coefficients from CGM models: 3.78, 6.91, 4.39, 5.94). Since the median within-grade between-proficiency level cutoff difference is 32, that means, holding all else constant, an individual whose age of enrollment was six years greater than another was expected to score about one proficiency level higher.

The estimates for length of enrollment are especially difficult to interpret due to interactions and polynomial terms. Marginal effects plots can be used to visualize their effects.

Marginal effects are expected changes in the outcome associated with changes in one or more variables of interest while holding others constant. Figure 2 gives the marginal effect of length of enrollment from the four CWC models. Marginal effects were calculated and plotted with the ggeffects R package (version 1.1.2; Lüdecke, 2018). These estimates also assume that the values of all other fixed and random effects are zero, meaning these predictions hold for the average observation of the average individual in the average state. Because these variables were centered, their zero points are not zeros on their original scales. The zero point for each variable depends on the level at which it was centered. Length of enrollment was centered by student, so values of −1, 0, and 1 roughly correspond to the first, second, and third observations, respectively, with some variability since tests are not administered exactly one year apart. Age of enrollment was centered by state, so a value of zero represents the state average for age of enrollment, which averaged approximately eight. Four age of enrollment values are plotted to visualize the interaction of age of enrollment and length of enrollment. These four centered values of −3, 0, 3, and 6 roughly correspond to age of enrollment values of 5, 8, 11, and 14, respectively.

Figure 2

Predicted Scores by Age of Enrollment and Length of Enrollment



Because plots use the estimates from the CWC models, the absolute intercept differences cannot be interpreted as the age of enrollment effect. In these plots, it is only appropriate to interpret the slopes and relative differences between the four lines: length of enrollment (linear and squared) and its interactions with age of enrollment.

In Panel A, which shows predicted listening scores, the lines' overall slopes are positive, meaning that scores are expected to increase over time. The lines are not straight, however, since the effect of length of enrollment squared was negative; their slopes decrease from left to right. The interaction between age of enrollment and length of enrollment and the interaction between age of enrollment and length of enrollment squared were both negative, meaning that those enrolling at a later age were expected to increase in scores at a lower rate; the gray dashed line's (age of enrollment of six above the state average) overall slope is lower than that of the black solid line (age of enrollment of three below the state average).

The marginal effects of length of enrollment on the other three domain scores are visualized in Panel B (reading), Panel C (speaking), and Panel D (writing) of Figure 2. Again, ignoring the intercept differences between the four age of enrollment lines, the slopes of the sets of lines show the same patterns seen in Panel A. Holding all else constant, test scores were expected to increase with length of enrollment, the rate of increase was expected to slow over time, and a higher age of enrollment was associated with a slower rate of growth in test scores.

To assess exactly how much slower later starters were expected to progress and to understand the impact of the interaction effects, the magnitudes of the interaction terms' coefficients estimates can be compared to those of the main effects. This is done by dividing pairs of coefficients from the CWC models, and the resulting statistic is a measure of how much age of enrollment would need to change to double or cancel out length of enrollment's effect, depending on whether they are being added or subtracted. Table 12 gives ratios of each main effect's coefficient to its corresponding interaction's coefficient for each outcome's CWC model.

Table 12

Ratio of the Coefficients of Main Effects to their Interactions with Age of Enrollment

Variable	Listening	Reading	Speaking	Writing
Length of Enrollment	−22.7	−91.0	−107.5[a]	−73.1
Length of Enrollment Squared	35.3	14.8	11.9	14.1

[a]The interaction of age of enrollment with length of enrollment was not statistically significant in the speaking CWC model.

For example, in the listening CWC model, the ratio of length of enrollment (36.59) to the interaction of age of enrollment with length of enrollment (−1.61) is −22.7. Holding all else constant,[1] an increase of 22.7 in age of enrollment from the state average was expected to decrease the linear slope of length of enrollment to zero. An increase of this magnitude is far outside the age of enrollment range observed in this sample, where the maximum deviation from eight was about 14 years (22.76; see Table 5, p. 44), and a more typical age for those completing grade 12 is 18 years old. One fourth of 22.7 is 5.7, so an individual with an age of enrollment around 13.7 was expected to have a rate of growth about one-fourth slower than the average individual. However, with such limited time remaining before their expected exit from the sampling frame, such a change in slope would have had minimal effects in the period of observation. The ratios of length of enrollment to its interactions with age of enrollment for the other domain outcomes are even larger (−91.0, −107.5, −73.1), making the impact of the interaction on growth rates negligible.

As for the length of enrollment squared, its effect should be quantified when age of enrollment is at the state average first. Setting up the length of enrollment effects as a set of expressions and taking their derivatives yields predictions for when length of enrollment's

[1] "Holding all else constant," while a strong assumption to make in most contexts, is impossible in this context where another model term was calculated by squaring length of enrollment.

overall slope would equal zero, when a student's growth was expected to plateau. For example, for the listening CWC model, this expression would be

$$36.59 * \text{LengthOfEnrollment} - 2.07 * \text{LengthOfEnrollment}^2$$

and the derivative crosses the axis of length of enrollment at 8.8. This means that the average individual was expected to show no more growth after 8.8 years past their mean length of enrollment, which is around the typical maximum age in this population. The plateau points for the other outcomes were 10.4 years (reading), 8.5 years (speaking), and 8.3 years (writing), which are likewise at points at or beyond grade 12.

Then, to understand the impact of the interaction of age of enrollment with length of enrollment, their coefficients can be added to the expressions. Doing so for an individual whose age of enrollment was six years above the state average (so, around 14 years old) and then taking derivatives would yield expected plateau points of 5.5 years (listening), 6.9 years (reading), 5.4 years (speaking), and 5.8 years (writing) beyond the average length of enrollment. These again point to times not expected within the population of grade school ELLs. This was anticipated by the large coefficient ratios in the second row of Table 12.

What these calculations demonstrate is that although age of enrollment's interactions with length of enrollment (linear and squared) were statistically significant, their relative magnitudes were such that length of enrollment's slope was expected to change little as age of enrollment changed. This too is shown in Figure 2. Again, while the absolute ordering of the lines calculated from the CWC models do not represent the effect of age of enrollment, the relative slopes do represent the interaction effects. The range of the lines decreases from left to right and the lines corresponding to the higher ages of enrollment have the lowest overall slopes, most markedly in the listening model (Panel A) since it has the smallest coefficient ratio for the

linear effect (Table 12). However, the difference of the differences between successive lines is only about 10 points for the years visualized in these plots.

An additional set of eight models with fewer predictors was fit to the full sample (N = 42,588). The full sample did not exclude any students whose L1 did not have a language similarity statistic in the Gateway Languages Database, and there was no minimum district size imposed. The fixed effects structure included age of enrollment, length of enrollment (linear and squared), and the interaction of the age and length variables. Tables of model estimates are in Appendix D. For these five effects, the only difference in statistical significance was that the interaction of age of enrollment and length of enrollment was statistically significant in the speaking CWC model, where it was not in the model using the analytic sample. The coefficients were very similar, however (−0.31 in model with analytic sample; −0.33 in model with full sample). A series of post-hoc Wald tests were run to compare coefficients from the two sets of models, dividing the difference of the pairs of coefficients by the square root of the sum of their squared standard errors ($z = \frac{\widehat{\beta}_1+\widehat{\beta}_2}{\sqrt{SE_1+SE_2}}$). The resulting test statistics indicate the likelihood of observing the two coefficient estimates in each pair had they been drawn from a single distribution. All test statistics were decidedly statistically non-significant, with the greatest *z*-value at 1.19 without even correcting for 20 multiple comparisons. The inference to be drawn from these tests is that the criteria applied in the creation of the analytic sample to include language similarity and L1 density as predictors were unlikely to have substantially changed the relationships between the variables.

Language Similarity

Language similarity was a student-level variable, so its effect was estimated in the CGM models. The coefficient estimate was not statistically significant in the listening, reading, and

writing models. In the speaking model, the coefficient estimate was 0.11 and statistically significant. The maximum difference between two individuals' similarity score was 62 (range 24–86; see Table 5). Holding all else constant, the expected difference in the scores for two students, one each at the minimum and maximum values of language similarity, would have been about 6.9 points on the speaking test, where the individual speaking the L1 more similar to L2 English would have had the higher score. This equates to approximately one fifth of the median difference between proficiency levels, and a couple points more than the expected change associated with a change of one in age of enrollment (4.39; Table 11).

Because over 80% of the students in the sample were speakers of the same L1 (Spanish), additional CGM models were fit to test whether the effect estimates for L1 similarity were sensitive to the uneven representation of different L1s, where most or all the L1 Spanish students were excluded from the dataset (model estimates not provided). Language similarity's effect estimates and patterns of statistical significance remained stable. Furthermore, the diagnostics for the models fit to the analytic sample did not suggest that the mean or variance of residuals varied with language similarity.

L1 Density

Although at a conceptual level, students were nested within different linguistic contexts, the L1 density variable varied year to year within students and was therefore a level-one variable. This means its effect was estimated in CWC models, as were its cross-level interaction with age of enrollment and its within-level-one interaction with length of enrollment. The main effect of L1 density was not statistically significant for any of the four domain CWC models, and only one of its eight interaction terms was statistically significant. In the speaking model, the interaction of L1 density and age of enrollment was estimated at 0.08. This means that

individuals enrolling at an age greater than the state average (approximately eight) were predicted to have higher scores as L1 density increases, while individuals enrolling at an earlier age were expected to excel in lower L1 density contexts. The age of enrollment coefficient in the speaking CGM model (4.39) was 56 times larger than the interaction of L1 density with age of enrollment in the CWC model, meaning that an increase of 56 percentage points in L1 density (e.g., from 30% to 86%) was associated with a doubling of the age of enrollment effect. A decrease in L1 density of that magnitude was also predicted to be associated with a negation of the age of enrollment effect.

An additional set of models was fit where the L1 density statistic was calculated at the school level rather than at the district level. The purpose of these models was to test the sensitivity of how the L1 density variable was calculated. This sample, called the school-L1-density sample, had a minimum school size of five students per year, and this drastically reduced the sample size to 9,933 observations. Other than the difference in the level at which L1 density was calculated, there was only one difference between the models fit to the school-L1-density sample and those fit to the analytic sample; the school-L1-density reading CGM model included a random slope for length of enrollment at the student level, making its random effects structure identical to those of the other seven models. Tables of model estimates are in Appendix D. The effects of interest were L1 density and its interactions with age of enrollment and length of enrollment in the CWC models, a total of 12 effects. There were three differences; the speaking model did not have a statistically significant interaction of L1 density and age of enrollment, and the listening and reading models both had statistically significant main effects of L1 density. The false positive rate with 12 additional tests is 46% (Family-wise error rate $= 1 - (1 - \alpha)^m = 1 - .95^{12} = .46$), so some differences due to random chance are to be expected. The L1 density

coefficient estimates were larger in magnitude as well: 0.18 in the listening model and 0.11 in the reading model, compared to 0.03 and 0.06, respectively. In these two models, the predicted change in ACCESS scores associated with a change of 50 percentage points in school L1 density would be 9.1 and 5.3 points, respectively, or 28% and 17% of a proficiency level.

To check whether L1 density's effect would be statistically significant if its interactions with age of enrollment and length of enrollment were not included in the model, an additional set of four CWC models were fit without these interactions (model estimates not provided). However, L1 density's effect remained statistically non-significant ($ps > .50$).

Domain Differences

Because scale scores are not comparable across domains (WIDA, 2022), coefficients from models fit to different domain scores cannot be compared. Therefore, to assess whether these models differed in any of their estimates, just the statistical significance and sign of effects were compared. Table 13 provides the sign of coefficient estimates for each domain where statistically significant, drawing from either the CWC or CGM model where appropriate. This presentation of the data makes apparent that the listening, reading, and writing models showed the same patterns, and that the speaking models differed for three effects. The interaction between age of enrollment and length of enrollment was not statistically significant, and there were positive effects found for the interaction of L1 density and age of enrollment, and for language similarity. These effects were discussed above in the relevant sections.

Table 13

Coefficient Estimate Signs Where Statistically Significant by Domain

Variable	Listening	Reading	Speaking	Writing
Variables from CWC Models				
Length of Enrollment	+	+	+	+
Length of Enrollment Squared	–	–	–	–
Age of Enrollment * Length of Enrollment	–	–		–
Age of Enrollment * Length of Enrollment Squared	–	–	–	–
L1 Density				
L1 Density * Age of Enrollment			+	
L1 Density * Length of Enrollment				
Variables from CGM Models				
Age of Enrollment	+	+	+	+
Language Similarity			+	

Note. Table cells are blank where coefficient estimates were not statistically significant.

Summary of Results

The models fit to the four outcomes were mostly consistent in their effects. Age of enrollment had a small, positive effect. Length of enrollment had a positive effect that attenuated over time. Age of enrollment and length of enrollment had statistically significant interactions that were mostly negligible in their effect sizes. Language similarity was statistically significant only in the speaking model and had a small, positive coefficient estimate. L1 density was not statistically significant in any of the models with the analytic sample, and its only significant interaction was with age of enrollment in the speaking model, which was positive.

Chapter 5: Discussion

Following the quantification of the relationships between the variables of interest and ACCESS test scores in Chapter 4, this chapter relates this study's substantive findings to the literature on child and adolescent SLA presented in Chapter 2, and it answers the seven research questions posed in Chapter 3. Table 14 contains a summary of the research questions, hypothesized relationships with the outcomes, and observed relationships. Each research question is discussed in its own section.

This discussion raises the possibility of false positives at several points. The effects of nine variables and interactions were estimated for four outcomes in three sample constructions, for a total of 108 statistical tests, not including post hoc comparisons of coefficients between models. There exists a real chance of a spurious relationship wherever these tests agreed or disagreed. The frequentist approach to statistics, as employed in this dissertation, assumes repeated sampling and testing, so it is necessary for novel studies like this one to be replicated again and again to accumulate evidence for or against hypotheses. Nevertheless, although the current study is only one study, its findings can be integrated with the existing literature to make conjectures about the forms of relationships between variables. Where the findings are inconsistent, such as where patterns of statistical significance differ for a single outcome or for models fit to an alternate subsample, conclusions are even more tentative following the likelihood of false positives in multiple comparisons. The way forward is more research, more studies exploring the effects of age, exposure, language similarity, and L1 density in child and adolescent ELLs.

Table 14

Summary of Research Questions, Hypotheses, and Results

Research question	Hypothesized relationship	Observed relationship	As expected?
1. Age of enrollment	Positive effect.	Positive effect.	Yes
2. Length of enrollment	Positive linear effect. Negative quadratic effect.	Positive linear effect. Negative quadratic effect.	Yes
3. Interaction of age of enrollment and length of enrollment	Positive interaction with linear effect. Negative interaction with quadratic effect.	Negative interaction with linear effect. Negative interaction with quadratic effect.	Partially
4. Language similarity	Positive effect.	Positive effect in speaking model only.	Partially
5. L1 density	Negative effect.	None.	No
6. Moderating effect of L1 density on age of enrollment and length of enrollment	Negative interaction with age of enrollment. Negative interaction with length of enrollment.	Positive interaction with age of enrollment in speaking model only. No interaction with length of enrollment.	No
7. Domain differences	(No hypotheses offered.)	Mostly consistent. Stronger rate disadvantage in listening model. Several differences in speaking model.	Not applicable

Research Question 1: Age of Enrollment

Age of enrollment had a small, positive association with L2 proficiency, as measured by ACCESS test scores. The proposed mechanism for this increase in expected scores with a later age of enrollment is that, before enrollment in a school in the United States, at least some of these ELLs received exposure to English language instruction in school. The result is that the later individuals enroll in a United States school, the higher their expected level of L2 proficiency. The increase in scores with age of enrollment was small, and the magnitudes of the marginal effects of age of enrollment and length of enrollment are compared in the next section.

Research Question 2: Length of Enrollment

The model results showed a substantial, positive relationship between length of enrollment and ACCESS test scores, but the strength of this relationship decreased over time. The implication is that ELLs were indeed advancing in their L2 English proficiency over time during their enrollment in school, and that the ACCESS test was capturing this increase. In other words, ELL education was effective, and the ACCESS test measured the results of that education.

The progress that occurred while enrolled in a United States school (length of enrollment effect) was greater than the progress that occurred prior to enrollment (age of enrollment effect). The ratio of the linear length of enrollment slope to the age of enrollment slope was around six, though this ratio decreased over time due to the negative quadratic slope of length of enrollment. This supports an earlier-is-better model of child and adolescent SLA down to age four (the minimum in this sample). Furthermore, as discussed in the next section, not only was age of enrollment's effect much smaller than that of length of enrollment, ELLs enrolling at a later age progressed at an equal or slower rate.

The fact that age of enrollment showed a weaker relationship with L2 proficiency than did length of enrollment was somewhat to be expected. While some students likely had L2 English instruction in their countries of origin, the quality of this education surely varied, and some may have had little if any formal education at all (Drake, 2017). Averaging across these populations yields an effect much smaller than that of years of enrollment in United States schools with specialized ELL programs, with the result that age of enrollment's marginal effect was a fraction of that of length of enrollment.

Research Question 3: Interaction of Age of Enrollment and Length of Enrollment

Contrary to the hypotheses, a rate disadvantage was found for later enrollees. The form of the rate disadvantage varied across the domains; only the quadratic interaction was significant in the speaking model, and the effect magnitudes were negligible in all but the listening model. Whether a nontrivial rate disadvantage was found for ACCESS test scores is less clear than what was not found: a rate advantage.

Previous literature supports a rate advantage for later starters in the short term (Golberg et al., 2008; Krashen et al., 1979), and for earlier starters in the long term (Jia & Aaronson, 2003; Jia & Fuse, 2007; Krashen et al., 1979). Models were designed so that these short- and long-term advantages would have been captured by allowing age of enrollment to interact with length of enrollment's linear and quadratic terms, respectively. The expectation was that the interaction with the linear term would be positive to reflect a short-term rate advantage for later starters, but that the interaction with the quadratic term would be negative so that earlier starters would show a long-term rate advantage. This hypothesis was not supported by the results, as the earlier starters showed a rate advantage across the board; later starters were never expected to progress at a rate faster than earlier starters.

Assuming the results of the current study are trustworthy, and great care was taken to that end, two possibilities remain. First, the later starters' rate advantage, if it exists, is so short-lived in a second language context (e.g., a United States school) that it was not observed in the ACCESS data. Second, there may be a rate advantage for later starters, but it was counteracted by a preference for the L1.

Regarding the possibility that later starters' rate advantage is limited to the very early stages of SLA, previous research attests to this phenomenon. Golberg et al. (2008) found a rate advantage for later starters whose length of enrollment ranged 0.8–2.8 years, and only for one of two outcomes (receptive vocabulary but not productive vocabulary). Krashen et al. (1979) provided summary tables of studies comparing younger and older children's rates of learning (p. 577). Among the four studies in second language contexts, the longest length of exposure was three years, and two of the four had lengths of exposure of only one year or less (Ekstrand, 1976; Ervin-Tripp, 1974; Fathman, 1975; Snow & Hoefnagel-Höhle, 1978). Others likewise maintain that the rate advantage for later starters in a second language context lasts only a year (Singleton, 2003; Singleton & Ryan, 2004). In contrast, the mean length of enrollment in the current study was much longer at 4.8 years, and its maximum was 14.5 years (see Table 5, p. 44). Furthermore, test scores were collected on an annual basis, so there was little opportunity to observe a rate advantage confined to the first year.

Another possibility is that later starters do enjoy a rate advantage for some amount of time, but it is negated by a preference for the L1. This follows from the research of Jia and Aaronson (2003), who found that later arrivals preferred using the L1 with friends and for media consumption. The consequence in the current context would be that, even though later enrollees

learned at a faster rate, their preference for the L1 decreased their actual exposure to the L2, canceling out or reversing their rate advantage.

Whichever possibility is true, the results suggest that later starters do not enjoy a rate advantage that allows them to "catch up" to or surpass earlier starters for some time. Rather, ELLs enrolling at a later age require specialized attention to ensure that they do not lag behind their peers in their L2 proficiency.

Research Question 4: Language Similarity

No evidence was found for an association between language similarity and listening, reading, and writing scores. However, language similarity was found to have a small, positive relationship with speaking scores, with an expected difference of about one fifth of a proficiency level across the whole range of language similarity.

For the three domains where language similarity was uncorrelated with the outcome, this agrees with research that language similarity plays a greater role in early, basic SLA (Jeon & Yamashita, 2014; Melby-Lervåg & Lervåg, 2011; Ross, 2001). For example, ELLs who are already literate in L1s with Latin scripts may show an advantage over their counterparts who speak nonalphabetic L1s in rudimentary skills like letter recognition and handwriting. However, these skills would not be as important for higher level tasks like reading passages for main ideas or structuring arguments in an essay.

The model results did suggest that language similarity may have been related to scores on the speaking domain. Of course, it is possible that this result was a false positive, and the small effect size means that even if this were a true positive, the practical significance of the finding was small. If there was truly a relationship between language similarity and ACCESS speaking scores, it may signal that the L1 played a (small) role in higher level L2 tasks by influencing

intelligibility. Language sounds are one of the variables used in calculating the language similarity statistics in the Gateway Languages Database (Gnanadesikan & van Rossum, 2016). Intelligibility is defined as how much of one's speech is understood by listeners and is affected by factors such as pronunciation of certain phonemes and prosodic stress (Munro, 2011). Raters of the ACCESS speaking items take delivery into account (WIDA, 2020). Putting these points together reveals a possible mechanism through which language similarity can affect speaking scores. The language similarity score was partially determined by the similarity of sounds between languages, which affected L2 pronunciation under the assumption of L1 transfer, which affected intelligibility, which affected clearness of delivery, which partially determined the score assigned to speaking items. The many steps and composite measures in this causal chain would have attenuated relationships. A stronger relationship (assuming one actually exists) would be expected if, for example, a quantitative measure of L1-L2 sound similarity were used to predict speaking delivery, without regard to other aspects of language similarity or of the speaking task requirements.

A counterargument to interpreting the language similarity variable as a measurement of the distance between the L1 and L2 is that it was confounded with other variables. Like many variables in observational studies, language similarity is inseparable from other individual characteristics. Immigrant speakers of different L1s vary in many ways, including reasons for immigration, educational backgrounds, and challenges and discrimination experienced after immigration (The APA Presidential Task Force on Immigration, 2013). Some individuals move to the United States with advanced degrees to engage in specialized professions as doctors or scientists, while others with little formal education move in search of "low-skilled" labor. At a national level, some countries experience conflict and unrest that prompt individuals to emigrate

regardless of class. A consequence of the complexities of global migration is that ELLs enrolling in schools in the United States have preexisting differences.

These differences are confounded with L1 backgrounds if the L1 is treated as a categorical variable, as in studies reviewed in Chapter 2 (Chen et al., 2012; Friesen and Krauth, 2011; Paradis 2005, 2011). In this dissertation, because L1 backgrounds were treated as a continuous variable, these backgrounds would only be expected to be confounded with L1 if there were a correlation between language similarity and other variables influencing SLA. In other words, the differences in learners' starting points are only a concern if the similarity of their L1s to English systematically varied with variables like education or discrimination. Examples of this would be if speakers of languages more similar to English tended to come from backgrounds with higher (or lower) educational achievement. While an interesting possibility, this idea remains a hypothesis. It could be tested with individual-level data on family histories and experiences in the United States.

Putting this possibility aside and assuming that language similarity was not confounded with other effects, the results of this study suggest that language similarity is uncorrelated with L2 proficiency, or that it may have a weak, limited correlation. In other words, the good news is that no matter what an individual's L1 is, they can become just as proficient in L2 English as any other ELL.

Research Question 5: L1 Density

No evidence was found for a relationship between L1 density and ACCESS test scores in the models fit to the analytic sample. In the models fit to the school-L1-density sample, L1 density's effect was statistically significant in the listening and reading models, where its estimated effect was small and positive (see Appendix D).

Four possibilities exist for why no effect was found in the main set of models. These are that L1 density had (a) no effect, (b) a net zero effect, (c) an effect that depended on individual differences, or (d) an effect but required measurement at a lower level. These four hypotheses are discussed in turn.

First, L1 density may have had no relationship with ACCESS test scores. This could be because L1 density did not affect contact with the L2. This would signal a break from the mediation model proposed by Vervoort et al. (2012), where L1 density affects contact with the L2, and contact with the L2 affects L2 proficiency. In a school context where ELLs are limited in their autonomy and cannot self-select into wholly L1 environments, they must use the L2 in social and academic situations. Contact with the L2 therefore did not vary as a function of peers' L1s, leaving no variance for L1 density to explain. L1 density was thus unrelated to L2 proficiency, as measured by the ACCESS test. The L2 remained valuable and necessary to the ELL.

Second, L1 density may have had both positive and negative effects on ACCESS test scores, which canceled out to result in a net zero effect. L1 density would have had a positive effect on L2 proficiency arising from same-L1 academic peer support (Ahn & Jepsen, 2015; Carhill-Poza, 2015). The negative effect of L1 density would have arisen from decreased contact with the L2, which is assumed to vary under this hypothesis (Vervoort et al., 2012). L1 density may have had at least these two indirect paths to L2 proficiency: a positive effect through peer support and a negative effect through L2 contact. If the coefficients on these two paths were of equal magnitudes and opposite signs, their sum effect would be zero, and this would have been reflected in the null effect of L1 density on L2 proficiency.

Third, the effect of L1 density on ACCESS test scores may have been heavily moderated by individual differences. This hypothesis states that L1 density was not the sole, or even most important, predictor of L2 contact. What mattered more was the individual's response to the L1 density context: the interaction between individual differences and L1 density. Without controlling for these individual differences, L1 density was not able to explain any variance in L2 proficiency. These individual differences may include variables like motivation, where students who were more highly motivated to learn the L2 did so without regard to their peers' L1s. Less motivated students who perceived less value in acquiring the L2 may have preferred using the L1 with same-L1 peers. In other words, the L1 density effect varied across students, but there is a lack of evidence for this hypothesis since the models did not support a random slope of L1 density by students.

Fourth, L1 density may have had an effect on L2 proficiency in child and adolescent ELLs, but detection of the effect would have required L1 density to be measured at a lower level. The main analyses in this dissertation used an L1 density statistic calculated at the district level, and L1 density's coefficients were not statistically significant (see Table 10 and Table 11). Additional analyses found that when L1 density was calculated at the school level, one step closer to the individual, the effect was statistically significant and positive in the listening and reading models (see Table D3). These may have been false positives given the number of tests, but if they were true positives, they suggest that a more fine-grained measure of L1 density would have been needed to detect its effect.

The results from the current study cannot provide evidence for any one of these hypotheses over the others, except for the disagreement between the third hypothesis and the models' random effects structure. However, future research can distinguish among them by

collecting more data at the student level in the form of questionnaires about their L2 contact, social networks, and motivation. This would answer the questions posed in each hypothesis by determining whether L2 contact has variance that can be explained by L1 density (hypotheses one, two, and three) and whether individuals vary in their degrees of L2 contact and how these can be explained by individual differences (hypotheses three and four).

For the purposes of this study, however, the tentative conclusion is that there is no evidence for a link between L1 density and L2 proficiency in the context of ELL instruction in the United States and the ACCESS test. This means that concerns and fears about L1 enclaves in the ELL classroom are unjustified (see discussion in Billings & Walqui, 2021). While it is a possibility that the L1 is a positive resource in child and adolescent SLA (Carhill-Poza, 2015), it must be recognized that the observed net effect of L1 density on ACCESS test scores was nearly zero, implying that L1 density is a neutral factor in SLA.

Research Question 6: Moderating Effects of L1 Density

No evidence was found for a moderating effect of L1 density on length of enrollment (Table 10 and Table 11), but there was one significant interaction of L1 density and age of enrollment in the speaking model (Table 11). However, none of L1 density's interaction terms were statistically significant in the models fit to the school-L1-density sample (Table D3 and Table D4). The failure to find anything in a second set of models, in addition to the many tests, favor the possibility that the one statistically significant interaction was a false positive.

Weak to no evidence was found against the stability of age of enrollment and length of enrollment effects across L1 density contexts. These results contrast with research on the L1 density effect in adult populations that found a negative interaction with age of arrival (see Table 3, p. 26; Chiswick & Miller, 1992; Struck, 2020) and a positive interaction with length of

residence (Chiswick & Miller, 1992). The hypotheses of the current study were negative interactions with both variables, for the reasons that later enrollees would have preferred the L1, and that the exposure associated with length of enrollment would have decreased as L1 density increased. The underlying rationale of both reasons is a decrease in L2 contact associated with an increase in L1 density. However, as discussed in the previous section, these two variables may not have covaried, or the covariance between the two may have been moderated by other variables. If that were true, and if the relationships between L2 contact, age of enrollment, and length of enrollment were correctly hypothesized, then it would follow that L1 density would not have moderated the effects of age of enrollment and length of enrollment on ACCESS test scores.

Research Question 7: Domain Differences

The pattern of results was similar across the models fit to the four domains of the ACCESS tests (see Table 13), but some differences were found regarding the rate disadvantage and role of language similarity.

The rate disadvantage was markedly stronger in the listening model, so that the rate of increase in test scores slowed as age of enrollment increased (see Figure 2 and Table 12). This would agree with research suggesting an earlier sensitive period for the acquisition of phonology (e.g., Krashen et al., 1979), but this explanation is insufficient because the speaking section of the ACCESS test also requires L2 phonology through prompts (i.e., listening) and spoken responses. It may be possible then that some other aspect of the listening section is more susceptible to age-related decline, such as the length of the recordings or the item types, but this is only speculation.

Only the speaking model had a statistically significant effect of language similarity. While the effect was small, it suggests an ongoing relevance of the L1 in the later stages of SLA as measured in a task-based assessment.

The speaking model also showed a positive interaction between L1 density and age of enrollment. However, as discussed in the previous section, the multiple tests and inconsistent results across the district-level and school-level calculations of L1 density indicate this result may have been a false positive.

Chapter 6: Conclusion

The objective of this dissertation was to understand individual and contextual variables in child and adolescent SLA. This was accomplished by fitting statistical models that predicted four domains of ACCESS test scores as measures of L2 English proficiency from variables of age of enrollment, length of enrollment, language similarity, and L1 density. The key features of this dissertation study were its large sample size of over 10,000 individuals, the use of a standardized and well-researched measurement of L2 proficiency, the longitudinal nature of the data, a continuous measure of language similarity, the measurement of the L1 density effect in children and adolescents, and the comprehensiveness of the statistical analyses.

The major findings of this study are now summarized. An earlier age of enrollment was beneficial for SLA, as no benefits were found to be associated with a later age. The rate of progress in ACCESS test scores was comparable or worse as age of enrollment increased. Speaking an L1 more similar to L2 English may have had a small benefit for speaking scores, but little evidence was found that speakers of different L1s would become differentially proficient in the L2. Another variable with little predictive value was L1 density; little to no evidence was found that the prevalence of one's L1 among their peers was related to ACCESS test scores. Some small differences existed between predicted growth trajectories for the four domains, but results were largely the same.

In other words, the most important findings in this study were the not-findings, things that were not found. The lack of a rate advantage for later starters challenges the widely held belief that the rate of SLA increases with age. The weak to null effects of language similarity and L1 density signal that these variables are of little consequence in predicting L2 proficiency in a school setting.

Limitations and Future Directions

While every effort was made to ensure that analyses were done correctly and inferences made appropriately, this study has its limitations. The task of research is never done, so here suggestions are given on how future research may build upon the work begun in this study. The limitations arise from the granularity of data collection and the observational nature of the data.

One limitation of this study was the coarseness of some of its variables. As discussed in Chapter 5, there may be a rate advantage for older starters in the very early stages of SLA. To measure this, future research could collect data more frequently to test short-term differences in rate. Another limited variable was L1 density, which was calculated at the district level in the main analyses and at the school level in supplementary analyses. These are still removed from the daily experience of the ELL in particular classrooms and social circles. The links between L1 density and L2 proficiency could be better understood by collecting detailed data from individual students. Analyses of such data would illuminate the relationship between L1 density and L2 contact (who interacts with whom and why), as well as disentangle any positive and negative paths between L1 density and L2 proficiency, such as same-L1 academic support (positive) and decreased exposure to the L2 (negative).

Another limitation of the current study was due to the confounds present in the observational data. For example, the model estimates for age of enrollment did not only represent the effect of starting SLA at different ages. Rather, this variable was inseparable from other variables like previous instruction in L2 English, and reasons for immigration at different ages with associated variables of identity and motivation. Nevertheless, the results of this study can be used to make claims about what is expected in L2 proficiency progress in this context. These results therefore remain useful to educators and administrators for understanding the ACCESS

test score trajectories of ELLs in the United States, and this has been the very objective of this dissertation.

To continue researching child and adolescent ELLs' individual variables in context, researchers may consider making use of existing secondary datasets for two reasons. First, secondary datasets are often of a size far larger than researchers have the time or funding to collect themselves. There are other options to increase sample sizes like collaborating with research labs or collecting data online, but it remains a fact that much of SLA research is severely underpowered (Brysbaert, 2021). While not all research questions can be answered with secondary datasets, they are a valuable and underutilized source of data for SLA research. Second, the use of secondary datasets forces a dialogue between researchers and practitioners. To obtain the ACCESS data, the author first had to make the case to WIDA that this project was worth their time and would benefit ELLs. Then, while making sense of the results, the author spoke with ELL teachers who administer the ACCESS test, to better understand what variables matter and why in child and adolescent SLA. The purpose of research and modeling is to first understand how variables relate and then apply this understanding to make improvements. The conversations that happen when working with secondary datasets help ensure that research does not end with simply understanding, but rather puts findings into the hands of administrators and educators with the power to help real people.

Appendix A

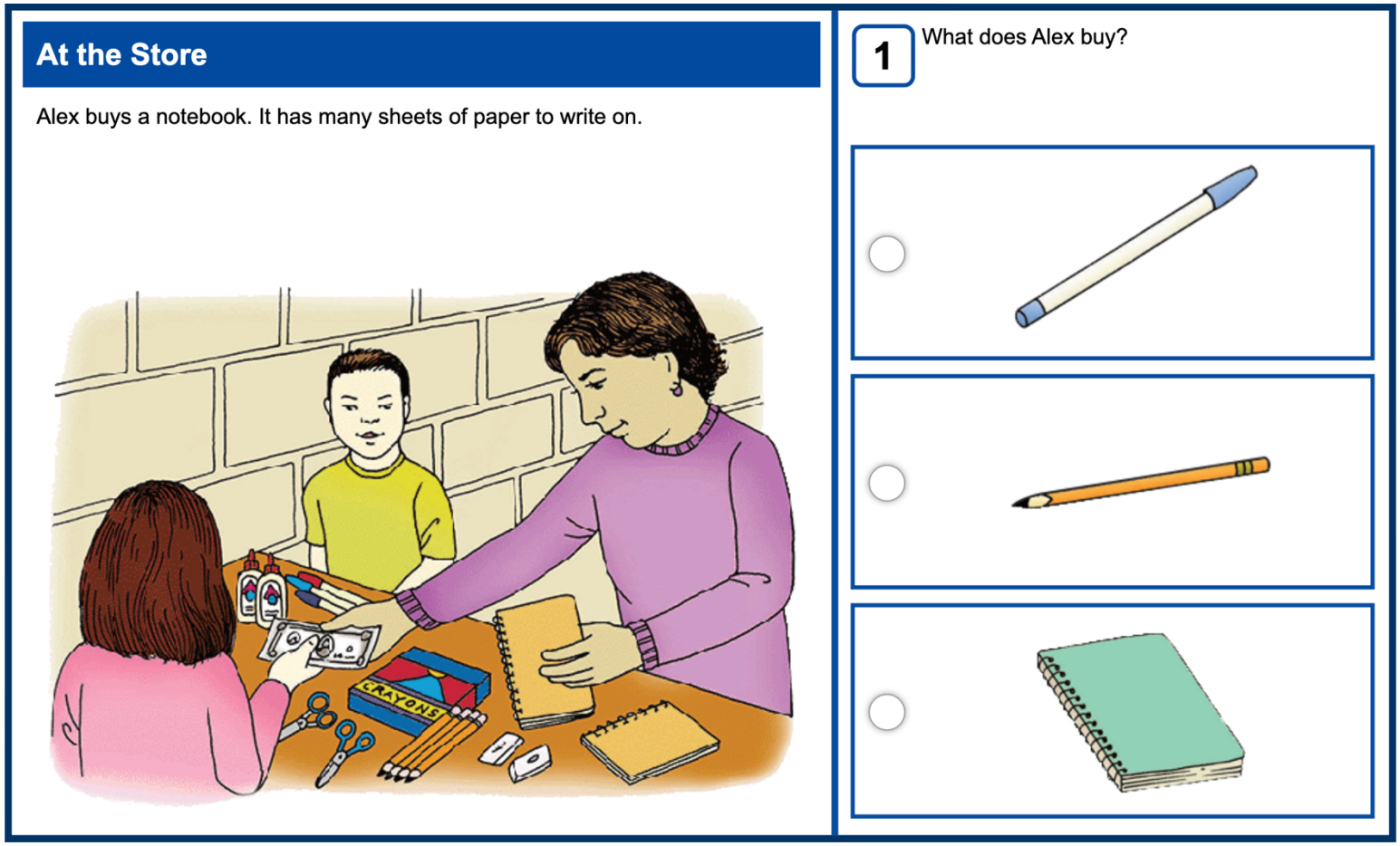
ACCESS Test Practice Items

This appendix contains images and text for a sample of five practice items for the ACCESS test, one for each of the five grade level groups. Where any recordings accompanied a screen, the recording was transcribed and given below the image of the screen. One item is provided for each of the listening, speaking, and writing domains, and two for the reading domain, for a total of five items. These were retrieved from https://wida.wisc.edu/assess/access/preparing-students/practice.

Grade 1 Reading: School Store – Tier B

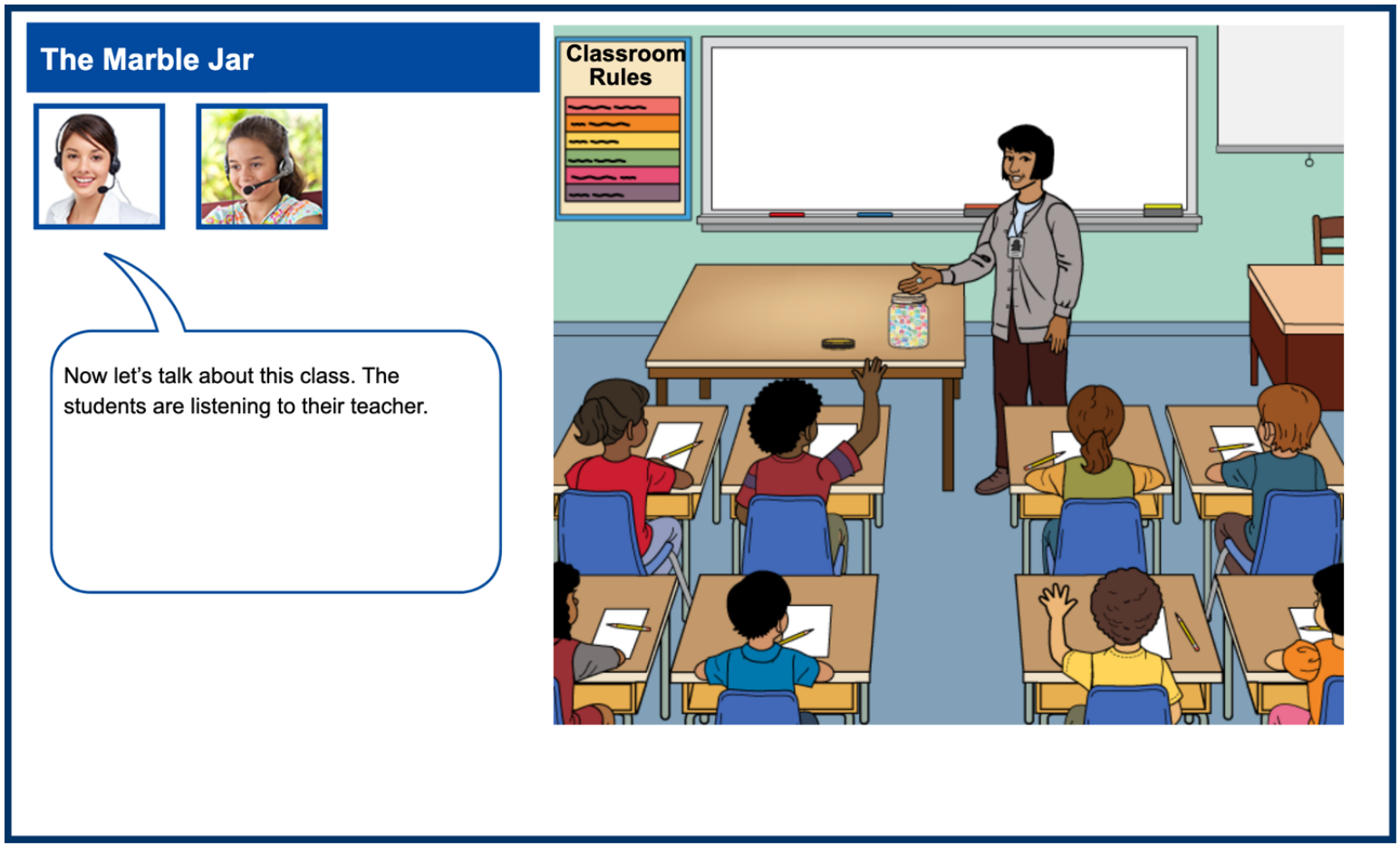
Screen 1:

Screen 2:


At the Store
Alex buys a notebook. It has many sheets of paper to write on.
1
What does Alex buy?
CRAYONS

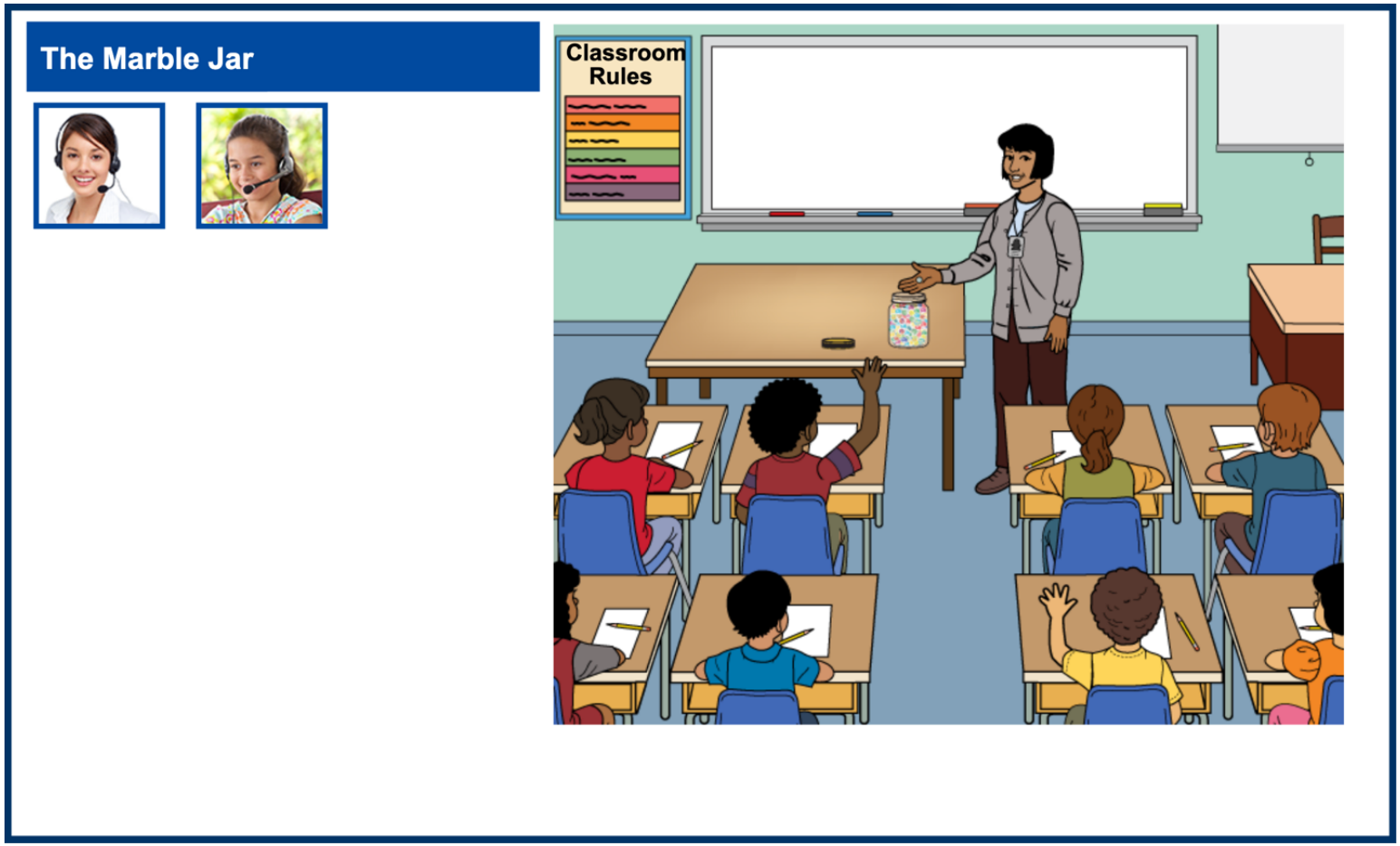
Grades 2–3 Speaking: The Marble Jar – Tier A

Screen 1:



Virtual Test Administrator: Now let's talk about this class. The students are listening to their teacher.

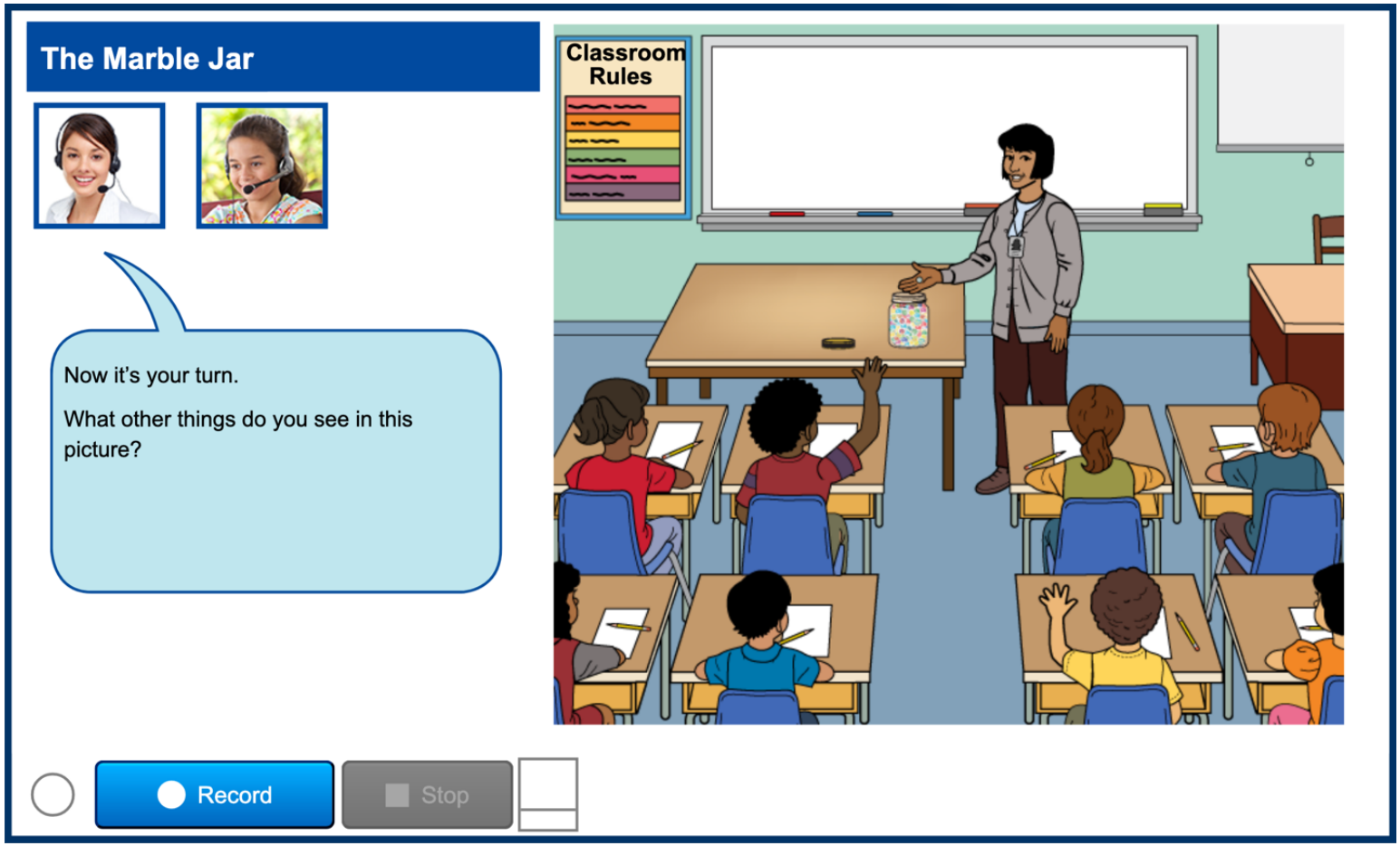
Screen 2:



Virtual Test Administrator: Nina, what are two things you see in this picture?

Virtual Interlocutor: A teacher and desks.

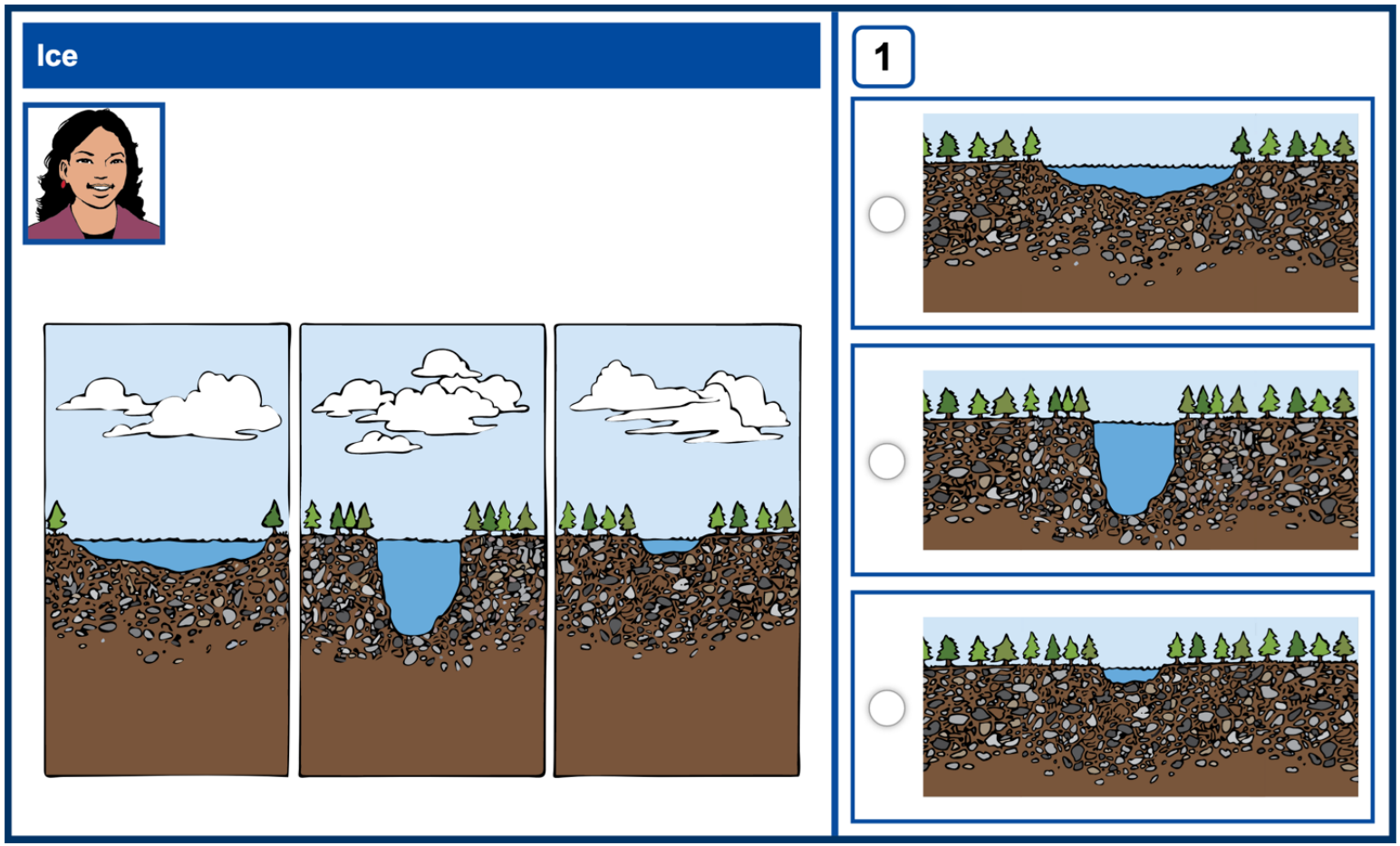
Screen 3:



Virtual Test Administrator: Now it's your turn. What other things do you see in this picture?

Grades 4–5 Listening: Ice – Tier B

Screen 1:



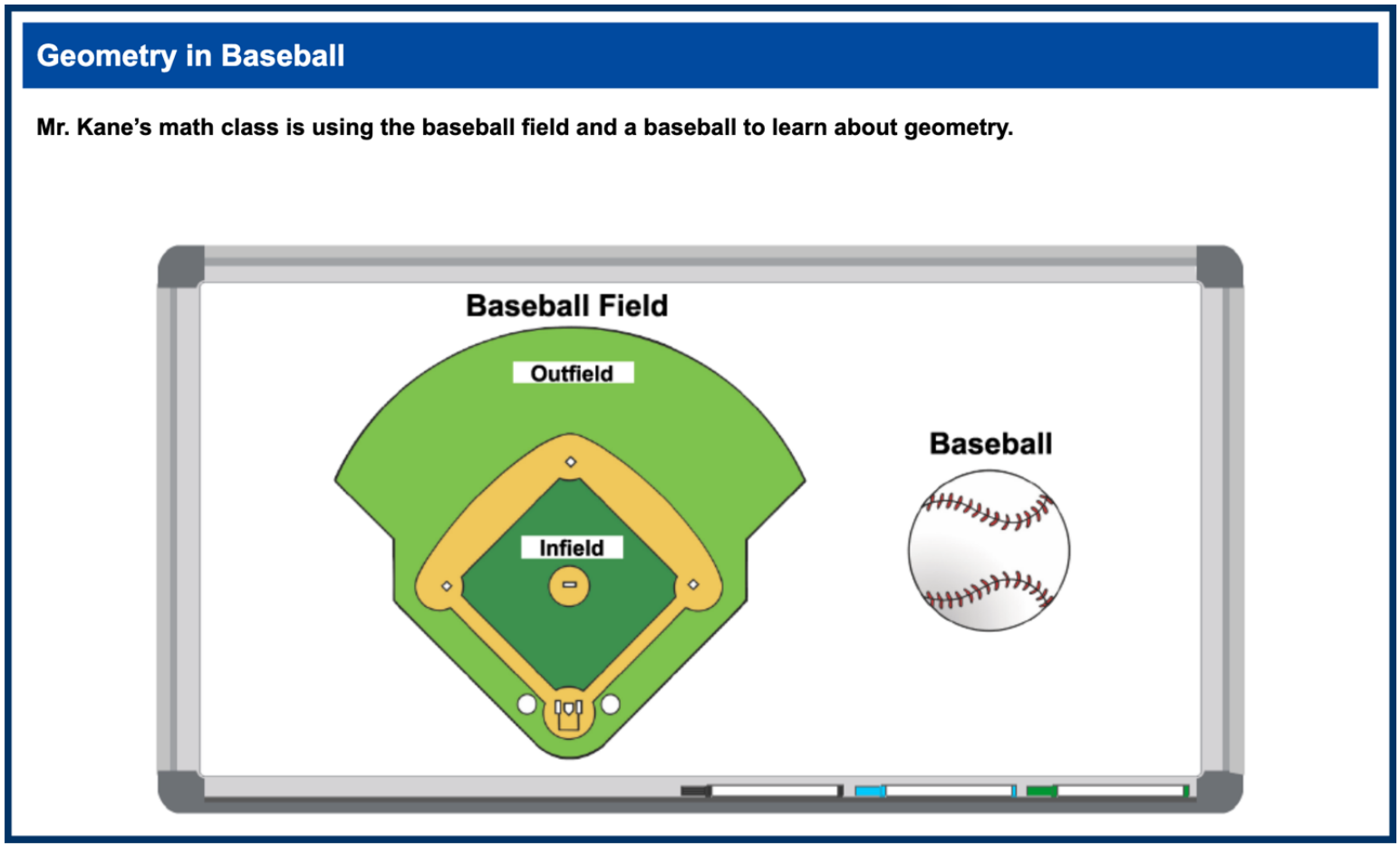
Virtual Test Administrator: Ms. Ortiz is talking to a student about how ice forms in different areas of the world.

Ms. Ortiz: Look at the picture of three different lakes. When the temperature is at zero degrees Celsius, the water will begin to freeze. Lakes that are small and shallow, or not very deep, freeze before lakes that are large and deep.

Virtual Test Administrator: From what you heard, which lake will freeze first when the temperature is at zero degrees Celsius?

Grades 6–8 Reading: Geometry in Baseball – Tier C

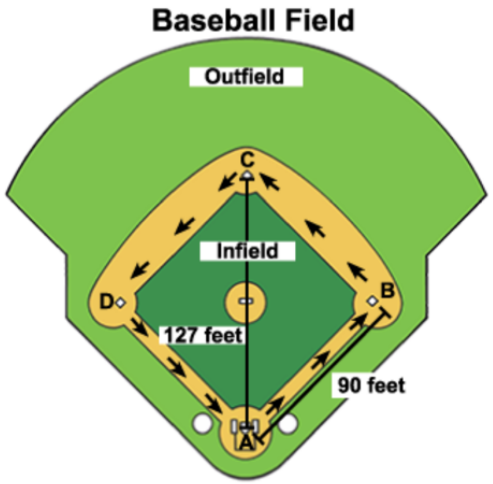
Screen 1:



Screen 2:

Geometry in Baseball

The small arrows in the picture show the path a player runs in a baseball game. To find how far a player runs, the students first measure the distance from point A to point B and find that it is 90 feet. They also find that the distance between points A, B, C, and D is always the same. To find the total distance a player runs, the students can either add all four sides of the infield together or multiply one side by 4.

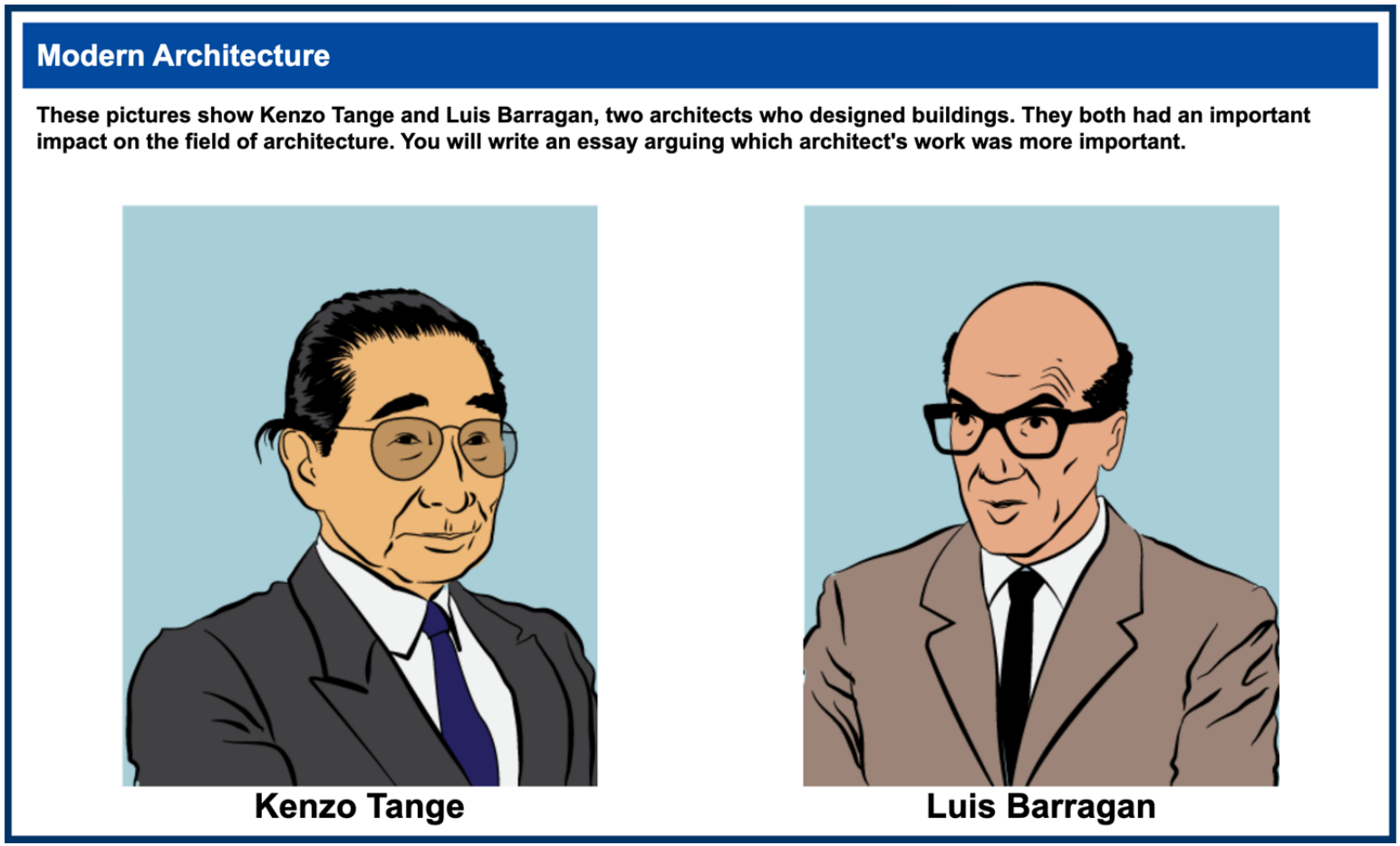


1 Which equation shows the distance a player runs?

- Perimeter = 90 feet x 4
- Perimeter = 127 feet x 4
- Perimeter = 90 feet + 4 feet
- Perimeter = 90 feet + 90 feet + 127 feet

Grades 9–12 Writing: Modern Architecture – Tier B/C

Screen 1:



Screen 2:

Modern Architecture

Here are notes about Kenzo Tange.

Kenzo Tange (1913–2005)

Education

- Architecture degree from Tokyo Imperial University
- Completed additional studies in city planning

Work Experience

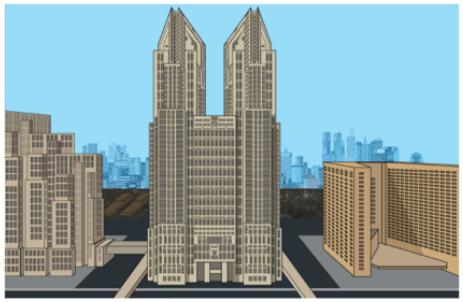
- Designed many public buildings in Japan and around the world
- Led the design of a museum and park used by many people
- Worked on a plan to help the city of Tokyo grow and change

Design Inspiration and Ideas

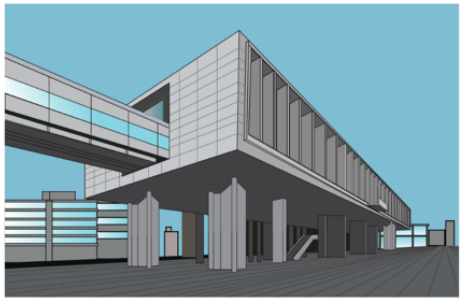
- Designed buildings to allow for expansion without rebuilding
- Combined traditional Japanese styles with modern styles

Awards and Accomplishments

- Won the Pritzker Architecture Prize in 1987
- Lectured at many universities
- Wrote influential books about architecture and urban planning



A government office building designed by Tange



A museum designed by Tange

Screen 3:

Modern Architecture

Here are notes about Luis Barragan.

Luis Barragan (1902–1988)

Education

- Engineering degree from the Free School of Engineers in Guadalajara, Mexico
- Self-taught as an architect and greatly inspired by traveling in Europe and Morocco

Work Experience

- Designed houses and housing complexes
- Emphasized the design of outdoor spaces to enjoy at home (gardens and fountains)
- Designed the furniture for his most important houses

Design Inspiration and Ideas

- Combined modern design features with traditional Mexican architecture styles
- Connected his work to the calm of the natural world
- Was well-known for his use of light and shadow, bold colors, and simple geometric shapes

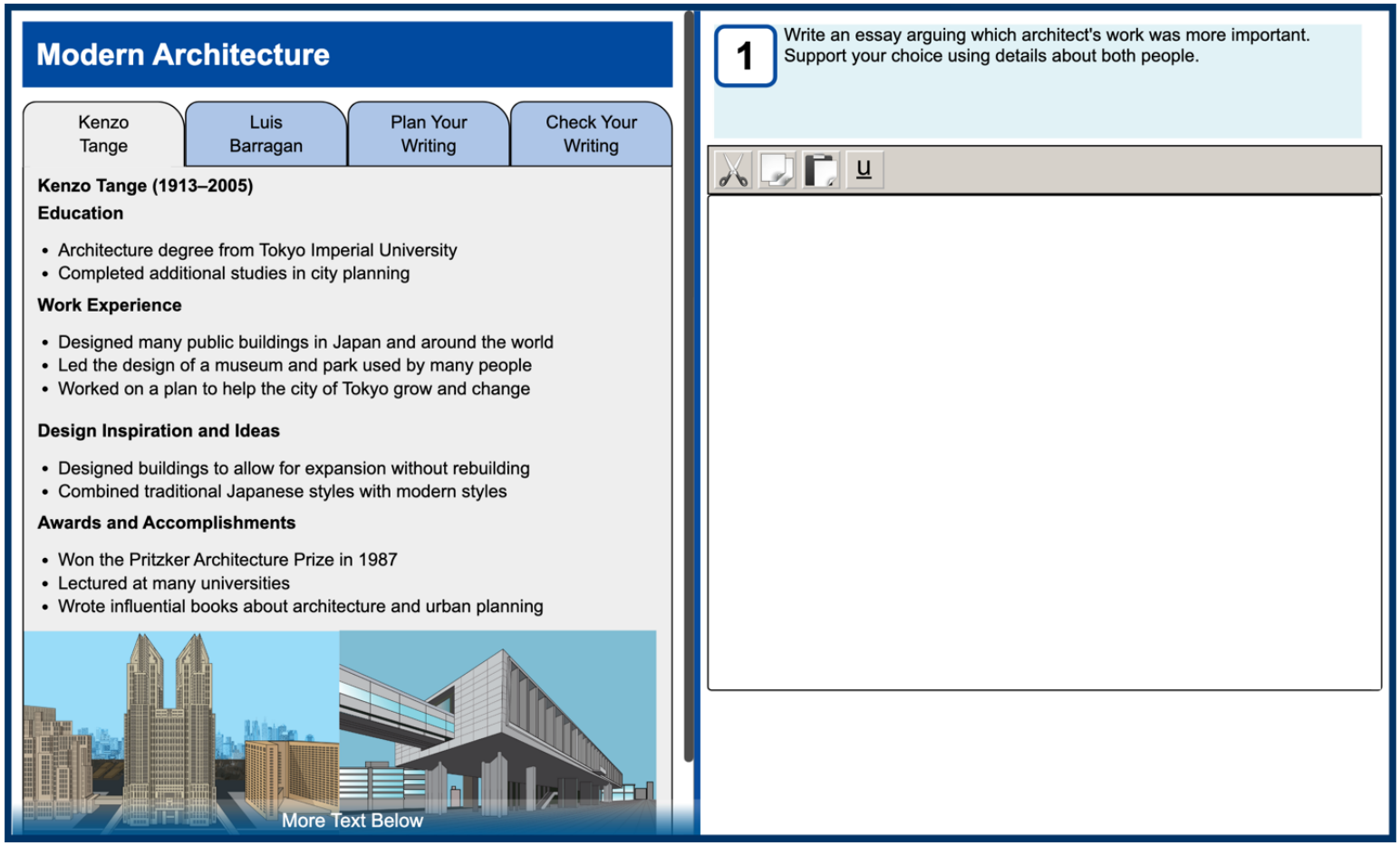
Awards and Accomplishments

- Won the Pritzker Architecture Prize in 1980
- His house and studio in Mexico City are now a museum

Pool and fountain in Barragan's style

A room in Barragan's house

Screen 4:


Modern Architecture
Kenzo Tange
Luis Barragan
Plan Your Writing
Check Your Writing
Kenzo Tange (1913–2005)
Education
• Architecture degree from Tokyo Imperial University
• Completed additional studies in city planning
Work Experience
• Designed many public buildings in Japan and around the world
• Led the design of a museum and park used by many people
• Worked on a plan to help the city of Tokyo grow and change
Design Inspiration and Ideas
• Designed buildings to allow for expansion without rebuilding
• Combined traditional Japanese styles with modern styles
Awards and Accomplishments
• Won the Pritzker Architecture Prize in 1987
• Lectured at many universities
• Wrote influential books about architecture and urban planning
More Text Below
1
Write an essay arguing which architect's work was more important. Support your choice using details about both people.
U

Appendix B

Descriptive Statistics for Alternate Subsamples of ACCESS Dataset

Tables B1 and B2 contain descriptive statistics for the full sample and school-L1-density sample, respectively, and descriptive statistics of the analytic sample are in Table 5 (p. 44). Counts by grouping units in each subsample are in Tables B3 and B4, which correspond to Table 6 in the main text (p. 45).

Of special note are the lines "Students per District" and "Students per School" in Tables B3 and B4. While the analytic sample had a minimum of five students per district in a year, the full sample (Table B3) had a minimum of one since no floor on group size was imposed. The school-L1-density sample (Table B4) restricted the group size at an even lower level, with a minimum of five students per school per year.

Table B1

Descriptive Statistics for Full Sample of ACCESS Dataset

Variable	Mean	SD	Min	Median	Max
Observation Level					
Length of Enrollment	4.76	2.81	0.00	4.40	14.51
Age at Testing	12.69	3.14	6.23	12.62	25.71
Listening Score	386.80	54.11	108.00	391.00	529.00
Reading Score	352.31	40.77	198.00	354.00	492.00
Speaking Score	305.76	51.45	106.00	312.00	476.00
Writing Score	333.92	42.93	111.00	337.00	508.00
L1 Density in District	65.46	34.55	0.11	81.94	100.00
Student Level					
Age of Enrollment	7.93	3.41	4.00	6.03	22.76
Language Similarity[a]	66.92	10.83	24.00	71.00	86.00

[a]In the full sample, only 37,137 observations have values for language similarity.

Table B2

Descriptive Statistics for School-L1-Density Sample of ACCESS Dataset

Variable	Mean	SD	Min	Median	Max
Observation Level					
Length of Enrollment	4.64	2.89	0.01	3.78	13.86
Age at Testing	13.40	3.42	6.26	13.81	21.66
Listening Score	386.07	52.82	108.00	389.00	529.00
Reading Score	356.45	40.58	222.00	357.00	483.00
Speaking Score	306.17	51.36	106.00	312.00	455.00
Writing Score	338.16	42.42	122.00	342.00	508.00
L1 Density in School	75.66	31.28	1.02	90.00	100.00
Student Level					
Age of Enrollment	8.77	3.85	4.00	7.42	18.71
Language Similarity	66.33	11.14	24.00	71.00	86.00

Table B3

Counts by Grouping Units in Full Sample of ACCESS Dataset

Variable	Mean	SD	Min	Median	Max
Districts per State	78.67	54.49	16	74	176
Schools per State	427.07	206.52	92	453	882
Schools per District	5.43	12.37	1	2	173
Students per State	946.40	54.59	826	976	998
Students per District	12.91	46.60	1	2	948
Students per School	3.44	5.89	1	2	105
Obs. per State	2839.20	163.77	2478	2928	2994
Obs. per District	36.09	135.78	1	6	2822
Obs. per School	6.65	12.50	1	3	281
Obs. per Student	3.00	0.00	3	3	3

Table B4

Counts by Grouping Units in School-L1-Density Sample of ACCESS Dataset

Variable	Mean	SD	Min	Median	Max
Districts per State	7.27	4.62	1	6	17
Schools per State	21.33	12.47	1	18	45
Schools per District	2.94	5.12	1	1	45
Students per State	220.73	186.95	5	162	672
Students per District	30.84	77.29	5	11	578
Students per School	13.65	13.47	5	9	97
Obs. per State	662.20	560.86	15	486	2016
Obs. per District	91.13	229.41	15	33	1707
Obs. per School	31.04	32.41	5	22	266
Obs. per Student	3.00	0.00	3	3	3

Appendix C

Equations for Models fit to Alternate Subsamples

Tables C1 and C2 contain the model equations for the CWC and CGM models fit to the full sample. These models differ from those fit to the analytic sample in that they lack terms for language similarity, L1 density, and L1 density's interactions. The equations of the models fit to the school-L1-density sample are identical to those fit to the analytic sample, which were presented in Table 8 and Table 9 (pp. 53–54), so these are not repeated.

Table C1

CWC Model Form for Models Fit to Full Sample

Equation term	Meaning	Level
$\text{Score}_{ijk} =$	Test score	Observation
$\gamma_{000} +$	Intercept	State
$\gamma_{010}\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) +$	Age of enrollment	Student
$\gamma_{100}\left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{.jk}\right) +$ $\gamma_{200}\left(\text{LengthEnrollment}^2_{ijk} - \overline{\text{LengthEnrollment}}^2_{.jk}\right) +$	Length of enrollment (linear and squared)	Observation
$\gamma_{110}\left(\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) \times \left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{.jk}\right)\right) +$ $\gamma_{210}\left(\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) \times \left(\text{LengthEnrollment}^2_{ijk} - \overline{\text{LengthEnrollment}}^2_{.jk}\right)\right) +$	Interaction of age of enrollment and length of enrollment (linear and squared)	Cross-level (observation and student)
$u_{00k} +$	Random intercept	State
$u_{0jk} +$	Random intercept	Student
$u_{01k} +$	Random slope of age of enrollment	State
$u_{1jk} +$	Random slope of length of enrollment	Student
e_{ijk}	Residual	Observation

Table C2

CGM Model Form for Models Fit to Full Sample

Equation term	Meaning	Level
$\text{Score}_{ijk} =$	Test score	Observation
$\gamma_{000} +$	Intercept	State
$\gamma_{010}\left(\text{AgeEnrollment}_{jk} - \overline{\text{AgeEnrollment}}_{.k}\right) +$	Age of enrollment	Student
$\gamma_{100}\left(\text{LengthEnrollment}_{ijk} - \overline{\text{LengthEnrollment}}_{...}\right) +$ $\gamma_{200}\left(\text{LengthEnrollment}^2_{ijk} - \overline{\text{LengthEnrollment}}^2_{...}\right) +$	Length of enrollment (linear and squared)	Observation
$u_{00k} +$	Random intercept	State
$u_{0jk} +$	Random intercept	Student
$u_{01k} +$	Random slope of age of enrollment	State
$u_{1jk} +$	Random slope of length of enrollment	Student
e_{ijk}	Residual	Observation

Appendix D

Estimates for Models Fit to Alternate Subsamples

Tables D1 and D2 give model estimates for the models fit to the full sample (N = 42,588), and Tables D3 and D4 provide estimates for models fit to the school-L1-density sample (N = 9,933). The purpose of these additional models was to assess whether sample definition and variable creation influenced the inferences drawn from the models fit to the analytic sample. The full sample was used to estimate the effects of age of enrollment and length of enrollment with all the available data, and the school-L1-density sample was used to check whether the L1 density effect varied depending on the level at which it was calculated.

The form of the full sample models was simpler than the form of the models fit to the analytic sample. It omitted language similarity, L1 density, and the interactions containing L1 density, leaving only age of enrollment, length of enrollment (linear and squared), and their interactions (see Appendix C for equations). The fixed effect structure of the models fit to the school-L1-density sample was identical to those of the models fit to the analytic sample. Of course, while the structure was the same, the L1 density variable was calculated at the school rather than district level. The random effects structure differs only for the reading CGM model, where a random slope was included for length of enrollment.

The results of these models are discussed in Chapter 4 where relevant: the full sample in the section “Age of Enrollment and Length of Enrollment” and the school-L1-density sample in the section “L1 Density.”

Table D1

Model Estimates for Listening and Reading Scores in Full Sample (N = 42,588)

	Listening		Reading			
Variable	CWC	CGM	CWC	CGM	Variable level	Effect estimated in
Fixed Effects						
Intercept	386.75 (1.97)***	386.48 (1.94)***	352.32 (1.58)***	352.51 (2.21)***		
Age of Enrollment	−0.09 (0.26)	4.01 (0.29)***	2.69 (0.28)***	7.14 (0.27)***	Level 2	CGM
Length of Enrollment	35.89 (0.50)***	28.57 (0.34)***	27.04 (0.31)***	19.92 (0.22)***	Level 1	CWC
Length of Enrollment Squared	−2.07 (0.05)***	−1.65 (0.03)***	−1.30 (0.03)***	−0.81 (0.02)***	Level 1	CWC
Age of Enrollment * Length of Enrollment	−1.55 (0.14)***		−0.38 (0.09)***		Cross-Level (1 and 2)	CWC
Age of Enrollment * Length of Enrollment Squared	−0.07 (0.02)***		−0.08 (0.01)***		Cross-Level (1 and 2)	CWC
Random Effects						
Level 2 SD(Intercept)	38.95	31.18	30.97	25.90		
Level 2 SD(Length of Enrollment)	15.40	4.58	9.70			
Level 2 Cor(Intercept, Length of Enrollment)	−0.19	−0.51	0.22			
Level 3 SD(Intercept)	7.51	7.40	6.04	8.49		
Level 3 SD(Age of Enrollment)	0.90	1.02	1.02	1.00		
Level 3 Cor(Intercept, Age of Enrollment)	0.12	0.13	0.29	−0.06		
Level 1 SD(Residual)	30.50	34.60	18.61	21.29		
R^2 (Fixed Effects Only)	0.08	0.23	0.15	0.39		
R^2 (Fixed and Random Effects)	0.68	0.62	0.79	0.77		

Note. Fixed effect estimates give Estimate (Standard Error) followed by asterisks denoting the significance level: * p < .05, ** p < .01, and *** p < .001.

Table D2

Model Estimates for Speaking and Writing Scores in Full Sample (N = 42,588)

	Speaking		Writing			
Variable	CWC	CGM	CWC	CGM	Variable level	Effect estimated in
Fixed Effects						
Intercept	305.69 (1.66)***	305.57 (2.15)***	333.96 (1.40)***	333.76 (1.90)***		
Age of Enrollment	0.70 (0.35)	4.69 (0.33)***	2.00 (0.31)***	6.10 (0.28)***	Level 2	CGM
Length of Enrollment	32.66 (0.50)***	24.10 (0.34)***	29.87 (0.36)***	23.22 (0.26)***	Level 1	CWC
Length of Enrollment Squared	−1.94 (0.05)***	−1.31 (0.03)***	−1.82 (0.04)***	−1.21 (0.02)***	Level 1	CWC
Age of Enrollment * Length of Enrollment	−0.33 (0.14)*		−0.60 (0.10)***		Cross-Level (1 and 2)	CWC
Age of Enrollment * Length of Enrollment Squared	−0.16 (0.02)***		−0.11 (0.01)***		Cross-Level (1 and 2)	CWC
Random Effects						
Level 2 SD(Intercept)	35.01	27.47	32.04	23.15		
Level 2 SD(Length of Enrollment)	13.12	4.89	11.10	4.22		
Level 2 Cor(Intercept, Length of Enrollment)	0.01	−0.30	−0.16	−0.25		
Level 3 SD(Intercept)	6.30	8.25	5.31	7.30		
Level 3 SD(Age of Enrollment)	1.31	1.22	1.17	1.03		
Level 3 Cor(Intercept, Age of Enrollment)	0.34	0.52	0.22	0.09		
Level 1 SD(Residual)	32.22	34.91	21.83	24.50		
R^2 (Fixed Effects Only)	0.08	0.21	0.11	0.32		
R^2 (Fixed and Random Effects)	0.61	0.57	0.74	0.70		

Note. Fixed effect estimates give Estimate (Standard Error) followed by asterisks denoting the significance level: * p < .05, ** p < .01, and *** p < .001.

Table D3

Model Estimates for Listening and Reading Scores in School-L1-Density Sample (N = 9,933)

	Listening		Reading			
Variable	CWC	CGM	CWC	CGM	Variable level	Effect estimated in
Fixed Effects						
Intercept	384.96 (3.19)***	383.23 (2.26)***	358.71 (2.29)***	356.14 (2.10)***		
Age of Enrollment	−0.49 (0.33)	3.63 (0.30)***	2.18 (0.47)***	6.94 (0.34)***	Level 2	CGM
Length of Enrollment	33.76 (1.01)***	26.44 (0.68)***	26.08 (0.64)***	19.01 (0.43)***	Level 1	CWC
Length of Enrollment Squared	−1.90 (0.12)***	−1.44 (0.05)***	−1.36 (0.07)***	−0.72 (0.04)***	Level 1	CWC
Age of Enrollment * Length of Enrollment	−1.84 (0.25)***		−0.23 (0.16)		Cross-Level (1 and 2)	CWC
Age of Enrollment * Length of Enrollment Squared	−0.04 (0.03)		−0.13 (0.02)***		Cross-Level (1 and 2)	CWC
Language Similarity	−0.20 (0.07)**	−0.13 (0.08)	−0.13 (0.06)*	−0.04 (0.06)	Level 2	CGM
L1 Density	0.18 (0.08)*	−0.08 (0.03)**	0.11 (0.05)*	−0.08 (0.02)***	Level 1	CWC
L1 Density * Age of Enrollment	−0.03 (0.02)		−0.03 (0.02)		Cross-Level (1 and 2)	CWC
L1 Density * Length of Enrollment	0.10 (0.12)	0.02 (0.01)*	0.06 (0.08)	0.01 (0.00)	Level 1	CWC
Random Effects						
Level 2 SD(Intercept)	37.07	30.32	30.36	25.14		
Level 2 SD(Length of Enrollment)	13.67	3.87	9.90	0.58		
Level 2 Cor(Intercept, Length of Enrollment)	−0.14	−0.41	0.15	0.24		
Level 3 SD(Intercept)	11.45	7.82	8.13	7.48		
Level 3 SD(Age of Enrollment)	0.91	0.68	1.60	1.09		
Level 3 Cor(Intercept, Age of Enrollment)	0.35	−0.14	−0.36	−0.49		
Level 1 SD(Residual)	30.62	34.17	18.33	21.19		
R^2 (Fixed Effects Only)	0.09	0.23	0.14	0.41		
R^2 (Fixed and Random Effects)	0.67	0.60	0.80	0.77		

Note. Fixed effect estimates give Estimate (Standard Error) followed by asterisks denoting the significance level: * p < .05, ** p < .01, and *** p < .001.

Table D4

Model Estimates for Speaking and Writing Scores in School-L1-Density Sample (N = 9,933)

	Speaking		Writing			
Variable	CWC	CGM	CWC	CGM	Variable level	Effect estimated in
Fixed Effects						
Intercept	305.73 (2.97)***	304.01 (2.55)***	340.94 (2.40)***	338.44 (1.98)***		
Age of Enrollment	0.19 (0.57)	4.17 (0.44)***	1.49 (0.49)**	5.88 (0.36)***	Level 2	CGM
Length of Enrollment	29.88 (1.04)***	22.40 (0.68)***	29.97 (0.74)***	20.99 (0.50)***	Level 1	CWC
Length of Enrollment Squared	−1.64 (0.12)***	−1.14 (0.06)***	−1.96 (0.08)***	−0.97 (0.04)***	Level 1	CWC
Age of Enrollment * Length of Enrollment	−0.83 (0.25)**		−0.84 (0.18)***		Cross-Level (1 and 2)	CWC
Age of Enrollment * Length of Enrollment Squared	−0.09 (0.03)*		−0.14 (0.02)***		Cross-Level (1 and 2)	CWC
Language Similarity	−0.24 (0.07)***	−0.13 (0.08)	−0.12 (0.06)*	−0.04 (0.06)	Level 2	CGM
L1 Density	0.16 (0.09)	−0.10 (0.03)***	0.07 (0.06)	−0.07 (0.02)**	Level 1	CWC
L1 Density * Age of Enrollment	−0.02 (0.03)		−0.01 (0.02)		Cross-Level (1 and 2)	CWC
L1 Density * Length of Enrollment	−0.20 (0.13)	0.02 (0.01)*	0.03 (0.09)	0.01 (0.01)*	Level 1	CWC
Random Effects						
Level 2 SD(Intercept)	33.61	26.73	30.93	23.37		
Level 2 SD(Length of Enrollment)	12.37	4.27	10.95	2.92		
Level 2 Cor(Intercept, Length of Enrollment)	0.08	−0.33	−0.11	−0.26		
Level 3 SD(Intercept)	10.63	9.04	8.54	7.01		
Level 3 SD(Age of Enrollment)	1.94	1.40	1.64	1.15		
Level 3 Cor(Intercept, Age of Enrollment)	0.33	0.49	−0.42	−0.54		
Level 1 SD(Residual)	32.68	35.16	21.58	24.73		
R^2 (Fixed Effects Only)	0.07	0.20	0.11	0.33		
R^2 (Fixed and Random Effects)	0.60	0.55	0.74	0.69		

Note. Fixed effect estimates give Estimate (Standard Error) followed by asterisks denoting the significance level: * p < .05, ** p < .01, and *** p < .001.

References

Ahn, T., & Jepsen, C. (2015). The effect of sharing a mother tongue with peers: Evidence from North Carolina middle schools. *IZA Journal of Migration, 4*(5), 1–21. https://doi.org/10.1186/s40176-015-0030-2

The APA Presidential Task Force on Immigration. (2013). Crossroads: The psychology of immigration in the new century. *Journal of Latina/o Psychology, 1*(3), 133–148. https://doi.org/10.1037/lat0000001

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). University of Ottawa Press. https://doi.org/10.2307/j.ctt1ckpccf.9

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https:/doi.org/10.18637/jss.v067.i01

Billings, E., & Walqui, A. (2021). *Topic brief 5: Dispelling the myth of "English only": Understanding the importance of the first language in second language learning*. New York State Education Department. http://www.nysed.gov/bilingual-ed/topic-brief-5-dispelling-myth-english-only-understanding-importance-first-language

Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research, 52*(2), 149–163. https://doi.org/10.1080/00273171.2016.1256753

Brindley, G. (2013). Task-based assessment. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 5526–5532). Blackwell. https://doi.org/10.1002/9781405198431.wbeal1141

Brysbaert, M. (2021). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition, 24*(5), 813–818. https://doi.org/10.1017/S1366728920000437

Burrows, C. (1993). *Assessment guidelines for the Certificate in Spoken and Written English: Stage 3: English for study*. NSW Adult Migrant English Service.

Carhill-Poza, A. (2015). Opportunities and outcomes: The role of peers in developing the oral academic English proficiency of adolescent English learners. *The Modern Language Journal, 99*(4), 678–695. https://doi.org/10.1111/modl.12271

Center for Applied Linguistics. (2019). *Annual technical report for ACCESS for ELLs 2.0 online English language proficiency test series 402, 2017–2018 administration* (Annual Technical Report 14A).

Center for Applied Linguistics. (2020). *Annual technical report for ACCESS for ELLs online English language proficiency test series 403, 2018–2019 administration* (Annual Technical Report 15A).

Center for Applied Linguistics. (2021). *Annual technical report for ACCESS for ELLs online English language proficiency test series 501, 2019–2020 administration* (Annual Technical Report 16A).

Chen, X., Ramirez, G., Luo, Y. C., Geva, E., & Ku, Y.-M. (2012). Comparing vocabulary development in Spanish- and Chinese-Speaking ELLs: The effects of metalinguistic and

sociocultural factors. *Reading and Writing, 25*, 1991–2020. https://doi.org/10.1007/s11145-011-9318-7

Chiswick, B. R., & Miller, P. W. (1992). Language in the labor market: The immigrant experience in Canada and the United States. In B. R. Chiswick (Ed.), *Immigration, language and ethnic issues: Canada and the United States* (pp. 229–96). American Enterprise Institute.

Chiswick, B. R., & Miller, P. W. (1995). The endogeneity between language and earnings: International analyses. *Journal of Labor Economics, 13*(2), 246–288. https://www.jstor.org/stable/2535104

Chiswick, B. R., & Miller, P. W. (2001). A model of destination-language acquisition: Application to male immigrants in Canada. *Demography, 38*(3), 391–409. https://doi.org/10.1353/dem.2001.0025

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development, 26*(1), 1–11. https://doi.org/10.1080/14790710508668395

Clark, M., Jackson, S., Kim, S., O'Rourke, P., Aghajanian-Stewart, K., & Ross, S. (2016). *Empirical evaluation of language difficulty categories* (Technical Report). College Park, MD: University of Maryland Center for Advanced Study of Language.

Cook, H. G., & MacGregor, D. (2017). *The ACCESS for ELLs 2.0 2016 standard-setting study* (Technical Report). Madison, WI: Board of Regents of the University of Wisconsin System.

Creagh, S., Kettle, M., Alford, J., Comber, B., & Shield, P. (2019). How long does it take to achieve academically in a second language? Comparing the trajectories of EAL students

and first language peers in Queensland schools. *Australian Journal of Language and Literacy, 42*(3), 145–155.

Cysouw, M. (2013). Predicting language-learning difficulty. In L. Borin & A. Saxena (Eds.), *Approaches to Measuring Linguistic Differences* (pp. 35–58). De Gruyter Mouton. https://doi.org/10.1515/9783110305258.57

Davis, P., & Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology, 21*(2), 99–106.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–533. https://doi.org/10.1017/S0272263100004022

Drake, K. (2017). Competing purposes of education: The case of underschooled immigrant students. *Journal of Educational Change, 18*, 337–363. https://doi.org/10.1007/s10833-017-9302-3

Ekstrand, L. 1976. Age and length of residence as variables related to the adjustment of migrant children, with special reference to second language learning. In G. Nickel (Ed.), *Proceedings of the Fourth International Congress of Applied Linguistics* (Vol. 3, pp. 179–198). Stuttgart: Hochschulverlag.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Ervin-Tripp, S. M. (1974). Is second language learning like the first. *TESOL Quarterly, 8*(2), 111–127. https://doi.org/10.2307/3585535

Fathman, A. (1975). The relationship between age and second language productive ability. *Language Learning, 25*(2), 245–253. https://doi.org/10.1111/j.1467-1770.1975.tb00244.x

Friesen, J., & Krauth, B. (2011). Ethnic enclaves in the classroom. *Labour Economics, 18*(5), 656–663. https://doi.org/10.1016/j.labeco.2011.01.005

Genesee, F., & Hamayan, E. (1980). Individual differences in second language learning. *Applied Psycholinguistics, 1*(1), 95–110. https://doi.org/10.1017/S0142716400000758

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in U.S. schools: An overview of research findings. *Journal of Education for Students Placed at Risk, 10*(4), 363–385. https://doi.org/10.1207/s15327671espr1004_2

Gnanadesikan, A., & van Rossum, J. (2016). *Gateway languages database* (Technical Report). College Park, MD: University of Maryland Center for Advanced Study of Language.

Golberg, H., Paradis, J., & Crago, M. (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics, 29*(1), 41–65. https://doi.org/10.1017/S014271640808003X

Granena, G., & Long, M. H. (2013). Introduction and overview. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. ix–xv). John Benjamins. https://doi.org/10.1075/lllt.35.002int

Hart-Gonzalez, L. and Lindemann, S. (1993). *Expected achievement in speaking proficiency, 1993* (mimeo). School of Language Studies, Foreign Services Institute, Department of State.

Hill, L., Betts, J., Hopkins, M., Lavadenz, M., Bachofer, K., Hayes, J., Lee, A., Murillo, M. A., Vahdani, T., & Zau, A. C. (2019). *Academic progress for English learners: The role of*

school language environment and course placement in grades 6–12. Public Policy Institute of California. https://eric.ed.gov/?id=ED593702

Hopp, H. (2011). Internal and external factors in the child L2 development of the German determiner phrase. *Linguistic Approaches to Bilingualism, 1*(3), 238–264. https://doi.org/10.1075/lab.1.3.02hop

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. https://doi.org/10.1111/lang.12034

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics, 24*(1), 131–161. https://doi.org/10.1017/S0142716403000079

Jia, G., & Fuse, A. (2007). Acquisition of English grammatical morphology by native Mandarin-speaking children and adolescents: Age-related differences. *Journal of Speech, Language, and Hearing Research, 50*(5), 1280–1299. https://doi.org/10.1044/1092-4388(2007/090)

Kim, A. A., Molle, D., Kemp, J., & Cook, H. G. (2018). *Examination of identification and placement decisions made for K–12 English learners* (WCER Working Paper No. 2018-12).

Kohn, J. (1997). Using English outside the classroom. In A. Burns, & S. Hood (Eds.), *Teachers' voices 2: Teaching disparate learner groups* (pp. 98–107). National Centre for English Language Teaching and Research.

Krashen, S. D., Long, M. H., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. *TESOL Quarterly, 13*(4), 573–582. https://doi.org/10.2307/3586451

Kuznetsova, A., Brockhoff, P. B, & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Long, M. H. (1990). Maturational constraints on second language development. *Studies in Second Language Acquisition, 12*(3), 251–285. https://doi.org/10.1017/S0272263100009165

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software, 3*(26), 772. https://doi.org/10.21105/joss.00772

Martin, S. (1998). *New life, new language: The history of the Adult Migrant English Program*. National Centre for English Language Teaching and Research.

Masters, M. C. (2014). *Self-assessment in second language testing: An updated statistical meta-analysis*. Unpublished manuscript.

Masters, M. C. (2023). Pathways to proficiency. In S. J. Ross & M. C. Masters (Eds.), *Longitudinal studies of second language learning: Quantitative methods and outcomes* (pp. 84–108). Routledge.

Meade, G., & Dijkstra, A. (2017). Mechanisms underlying word learning in second language acquisition. In M. Libben, M. Goral, & G. Libben (Eds.), *Bilingualism: A framework for understanding the mental lexicon* (pp. 49–71). John Benjamins. https://doi.org/10.1075/bpa.6.03mea

Meisel, J. M. (2009). Second language acquisition in early childhood. *Zeitschrift für Sprachwissenschaft, 28*(1), 5–34. https://doi.org/10.1515/ZFSW.2009.002

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading, 34*(1), 114–135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

Muñoz, C. (2006). The effects of age on foreign language learning: The BAF project. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 1–40). Multilingual Matters. https://doi.org/10.21832/9781853598937-003

Munro, M. J. (2011). Intelligibility: Buzzword or buzzworthy? In J. Levis & K. LeVelle (Eds.), *Proceedings of the 2nd pronunciation in second language learning and teaching conference* (pp. 7–16). Iowa State University.

National Center for Education Statistics. (2022). *English learners in public schools*. Condition of Education. https://nces.ed.gov/programs/coe/indicator/cgf.

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230–244. https://doi.org/10.1017/S0267190516000027

Odlin, T. (2003). Cross-linguistic influence. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 436–486). Blackwell. https://doi.org/10.1002/9780470756492.ch15

Paradis, J. (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with specific language impairment. *Language, Speech, and Hearing Services in Schools, 36*(3), 172–187. https://doi.org/10.1044/0161-1461(2005/019)

Paradis, J. (2007). Early bilingual and multilingual acquisition. In P. Auer & L. Wei (Eds.), *Handbook of multilingualism and multilingual communication* (pp. 15–44). De Gruyter Mouton. https://doi.org/10.1515/9783110198553.1.15

Paradis, J. (2011). Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism, 1*(3), 213–237. https://doi.org/10.1075/lab.1.3.01par

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191–215. https://doi.org/10.1006/jpho.2001.0134

Pope, N. G. (2016). The marginal effect of K-12 English language development programs: Evidence from Los Angeles Schools. *Economics of Education Review, 53*, 311–328. http://doi.org/10.1016/j.econedurev.2016.04.009

R Core Team (2022). *R: A language and environment for statistical computing* [software]. R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Raijman, R. (2013). Linguistic assimilation of first-generation Jewish South African immigrants in Israel. *Journal of International Migration and Integration, 14*, 615–636. https://doi.org/10.1007/s12134-012-0257-1

Raijman, R., Semyonov, M., & Geffen, R. (2015). Language Proficiency among Post-1990 Immigrants in Israel. *Journal of Ethnic and Migration Studies, 41*(8), 1347–1371. https://doi.org/10.1080/1369183X.2014.982523

Ross, S. J. (1997 May). *CSWE outcomes: 15 issues of description and inference* (Paper presentation). National Working Forum on Assessment and Reporting in English

Language and Literacy Programs: Trends and Issues, Macquarie University, Sydney, Australia.

Ross, S. J. (2001). Individual differences and learning outcomes in the Certificates in Spoken and Written English. In G. Brindley (Ed.), *Studies in immigrant English language Assessment* (Vol. 1, pp. 191–214). National Centre for English Language Teaching and Research.

Ross, S. J., & Masters, M. C. (2023). Introduction: Longitudinal research designs. In S. J. Ross & M. C. Masters (Eds.), *Longitudinal studies of second language learning: Quantitative methods and outcomes* (pp. 10–25). Routledge.

Schumann, J. H. (1978). The acculturation model for second language acquisition. In Gingras, R. (Ed.), *Second language acquisition and foreign language teaching* (pp. 27–50). Center for Applied Linguistics.

Singleton, D. (2003). Critical period or general age factor(s)? In M. P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 3–22). Multilingual Matters. https://doi.org/10.21832/9781853596407-002

Singleton, D., & Ryan, L. (2004). *Language acquisition: The age factor* (2nd ed.). Multilingual Matters. https://doi.org/10.21832/9781853597596

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Snow, C. E., & Hoefnagel-Höhle, M. (1978). Age differences in second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 333–344). Newbury House.

Struck, J. (2020). *Individual variables in context: Migrant second language competency achievement*. Unpublished manuscript.

Szuber, A. (2007). Native Polish-speaking adolescent immigrants' exposure to and use of English. *The International Journal of Bilingual Education and Bilingualism, 10*(1), 26–57. https://doi.org/10.2167/beb316.0

van Tubergen, F., & Wierenga, M. (2011). The language acquisition of male immigrants in a multilingual destination: Turks and Moroccans in Belgium. *Journal of Ethnic and Migration Studies, 37*(7), 1039–1057. https://doi.org/10.1080/1369183X.2011.572476

Vervoort, M., Dagevos, J., & Flap, H. (2012). Ethnic concentration in the neighbourhood and majority and minority language: A study of first and second-generation immigrants. *Social Science Research, 41*, 555–569. https://doi.org/10.1016/j.ssresearch.2012.01.002

WIDA (2020). *WIDA speaking scoring scale.* https://wida.wisc.edu/sites/default/files/resource/WIDA-Speaking-Scoring-Scale-Gr-1-12.pdf

WIDA (2022). *ACCESS for ELLS interpretive guide for score reports*. https://wida.wisc.edu/sites/default/files/resource/Interpretive-Guide.pdf

WIDA (n.d.). *ACCESS Test Practice and Sample Items*. https://wida.wisc.edu/assess/access/preparing-students/practice