ABSTRACT

Title of Dissertation Proposal:	A MULTILEVEL ANALYSIS OF LANGUAGE PROFICIENCY GROWTH
	Beth A. Mackey, Doctor of Philosophy, 2023
Dissertation directed by:	Dr. Kira Gor, School of Languages, Literatures, and Cultures

The U.S. Military Services employ thousands of servicemen and women in language-related positions that are critical to the nation's national security. These positions require personnel with high-level capability in various languages and dialects (Asch & Winkler, 2013). A complex accession and training system that begins at local recruiting stations across the nation leads to worldwide placement of language professionals who serve multiyear tours in the U.S. Air Force, Army, Navy and Marine Corps. High levels of cognitive ability, as measured by two cognitive aptitude batteries, one general (ASVAB) and one language (DLAB), are required for selection into these positions. Following significant investments in basic levels of training, the jobs themselves demand high level skills, and the service members find themselves constantly challenged to grow their skills. Traditional research on the effectiveness of the accession and training processes focuses on learning outcomes, rather than growth. This research used a longitudinal design to investigate how general aptitude,

language aptitude, non-cognitive and language distance measures impact language proficiency growth. Hierarchical linear models and hierarchical generalized linear models were used and the significant findings were similar. The study found that overall, while language test scores followed a drop-and-recover pattern, there was very little growth overall. Three aptitude subtests, one from ASVAB (Mechanical Comprehension) and two DLAB subtests (Part 3 and Part 4) were found to constrain initial growth in the listening modality. Language distance was found to constrain initial and subsequent growth in listening and reading.

A MULTILEVEL ANALYSIS OF LANGUAGE PROFICIENCY GROWTH

by

Beth A. Mackey

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:

Dr. Kira Gor, Chair
Dr. Donald Bolger, Co-Chair
Dr. Robert DeKeyser, Committee member
Dr. Catherine Doughty, Committee member
Dr. Jared Novick, Dean's representative

Acknowledgements

I would like to thank Dr. Kira Gor and Dr. Donald Bolger, who stepped in to serve as my Chair and Co-Chair midway through my dissertation. I would also like to thank Dr. Robert DeKeyser, who also joined my committee after my proposal defense, and to Dr. Catherine Doughty, for serving on my committee and providing such constructive feedback. Last but certainly not least, thank you to Dr. Jared Novick for serving as the Dean's Representative. You each brought expertise that strengthened my research, and I am grateful for your questions and input.

This dissertation would not have been possible without the original support of Dr. Steve Ross. He was not only my original advisor, but he encouraged me to return to the program after I took a several year gap due to demands of a full-time job. I am forever indebted.

The Defense Language Institute supported my research efforts and I want to thank Dr. Seumas Rogan, Mr. Kalman Weinfeld and Dr. Pradyumna Amatya for compiling the original dataset and assisting with the DLI IRB protocol. The opinions in the paper are mine and do not reflect the opinions of the DLI.

I have been lucky to have a number of people in my corner throughout this somewhat lengthy endeavor. First, from my original cohort at UMD, who all managed to get to this point much quicker than I: Dr. Tom Wagener, Dr. Megan Masters, and Dr. Elizabeth Rhoades, thank you all for your interest, encouragement, and technical expertise. I'd also like to thank Dr. Laura Stapleton, who despite her demanding schedule as acting Dean, took time to sit down with me last year to work through the challenges of a three-level model. From my work life, I want to especially thank Dr. Rachel Brooks,

Dr. Heeyeon Dennison, and Dr. Pardee Lowe, Jr. who over the years have become such good friends. You have willingly shared your expertise, ideas, and resilience techniques with me, especially down the home stretch. My friends have been an endless source of ideas to pursue and have willingly shared their experience in the language field. I especially want to thank Cheryl Houser and Sheryl Gordon.

Last but not least, I want to thank my family, especially my husband Ben. Without you, none of this would have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AFQT	Armed Forces Qualifying Test
ASVAB	Armed Services Vocational Aptitude Battery
DLAB	Defense Language Aptitude Battery
DLIFLC	Defense Language Institute Foreign Language Center (also abbreviated DLI)
DLPT	Defense Language Proficiency Test
DOD	Department of Defense
ILR	Interagency Language Roundtable
OPI	Oral Proficiency Interview

Chapter 1: The Problem

The US Department of Defense (DoD) considers foreign language skills to comprise an "enduring critical competency" (DoDD 5160.41E, p. 2). The acquisition of these foreign language skills is the responsibility of Defense Language Institute Foreign Language Center (DLIFLC) in Monterey, California (Army Regulation 350-20/OPNAVINST 1550.13/AFI 35-4004/MCO 1550.4E). After completing their language training, DLIFLC graduates serve in variety of key positions in the United States and around the world as language professionals, defined by policy as:

> "a person who is certified in a foreign language proficiency of at least skill level 2 (as identified in the Federal Government Interagency Language Roundtable (ILR) Skill Level Descriptions (*ILR*, 1985) in two of the three modalities (listening, reading, and speaking) in one or more foreign languages, and who requires that foreign language to perform his or her primary function" (DoDD 5160.41E, p. 17).

A subset of these professionals, those assigned to the intelligence community, are by policy required to be at least ILR skill level 3 in listening and reading (DoDI 5160.70). ILR level 3 is considered to be the "general professional" level, as skills at this level include "reading at a normal range of speed with almost complete comprehension a variety of authentic prose material on unfamiliar subjects" as well as "…[a]lmost always able to interpret material correctly, relate ideas and 'read between the lines'…" (ILR, 1985).

Reaching even ILR Level 2 is not easy, and the military has developed a stringent process through which it recruits and selects individuals to join the language program. Since the 1970's, recruits seeking to join the language profession have been required to meet high standards on two cognitive aptitude batteries, one a general

aptitude measure (the Armed Services Vocational Aptitude Battery) and the other a language aptitude measure (the Defense Language Aptitude Battery). Language training is intense, and a typical student graduates from DLIFLC at ILR Level 2 in reading and listening. Success as a language professional on the job, however, especially in the Intelligence Community, and as noted above requires even higher levels of language ability. According to a study conducted by Rand Corporation (Asch & Winkler, 2013) and sponsored by the Office of the Director of National Intelligence:

> Language professionals play a pivotal role in U.S. national security. The adversaries of the United States communicate their plans and actions in many different languages and dialects. Consequently, U.S. intelligence operations require personnel with high-level capability in these languages and dialects (p. xi).

In order to perform at these higher levels, language professionals must at the very least maintain the levels they reach at DLIFLC and/or grow their language skills after they graduate.

The DoD monitors the language program's accession, training and attrition of the language workforce. Most research has focused on predicting successful graduation from DLIFLC. This is understandable, given the high cost of language training. As Welsh & Kucinkas (1990) explained in their report, estimated annual cost savings of $80m ($181m in today's dollars) could be attributed to even small improvements (validity of .02 (r^2)) in the selection of recruits into training programs. However, it is short-sighted to only consider recruiting and selection costs, as how trainees continue to grow and improve in their skills on the job is also important for

government planners. This study investigated how the selection measures are related to both graduation outcomes and growth.

This chapter briefly describes the military language professional pipeline to provide context and reviews the literature related to the aptitude and foreign language measures used by the US Government for selection and training.

Military pipeline

Entry into the language profession in the US military services begins through successful performance on a measure of general aptitude, the Armed Services Vocational Aptitude Battery (ASVAB), usually summarized by one of several composite scores that draw from ten subtests. Recruits who earn high scores on the ASVAB are offered positions in more demanding skill fields in the military services, such as those in the nuclear and intelligence areas. The ASVAB has served as the entrance exam for enlistment since 1976 (Segall & Moreno, 1999). The original purpose of the test was to predict occupational success in the military, and its purpose was later extended to offer career exploration services to high school students (ASVAB Program, n.d.). Today's ASVAB is computer-adaptive and takes, on average, approximately one and half hours to complete. The ASVAB is designed to measure aptitude in four domains: Verbal, Math, Science and Technical, and Spatial. Recruits interested in a language career in the military must qualify not only on the ASVAB, but also on a language aptitude test, the Defense Language Aptitude Battery (DLAB). DLAB was developed in the 1970's to predict language training outcome measures, specifically for the DLIFLC academic context (Peterson & Al-Haik, 1976).

Today DLAB scores are still used to place students into language training, and higher language aptitude scores are required for the "difficult" languages.

Language Training

Once accepted in the military language job family, recruits attend basic military training. Upon successful completion of boot camp, they begin language training at DLIFLC in Monterey, California. These full-time intensive programs now run between 36 and 64 weeks, depending on how difficult the language is to learn for a native speaker of English. The U.S. national security interests drive the selection of languages taught at DLIFLC in any given year; training is currently offered in fourteen languages (Defense Language Institute, n.d.). The academic program is intensive, comprising seven hours of classroom time each day plus homework. Class size is small, usually between six and eight students per section. In addition to keeping up with their coursework, students also continue to receive military training to maintain their physical fitness and meet other basic expectations of the services. While course grades are important, the goal of training is to reach high levels of overall language proficiency in listening, reading, and speaking, as measured by the Defense Language Proficiency Testing system. Once on-the-job, these professionals continue to be assessed annually in listening and reading. Speaking proficiency is not routinely assessed post-DLI and therefore was not included in this research.

Across the US Department of Defense, language proficiency in listening and reading is measured by Defense Language Proficiency Tests (DLPT), now in their fifth generation (tests are known as the DLPT5). DLPT scores are mapped onto the Interagency Language Roundtable (ILR) Skill Level Descriptions (1985), a functional

scale ranging from Level 0 to Level 5, with sublevels known as "plus levels" from Levels 0 to 4. Graduates of the DLIFLC programs are expected to reach a minimum of ILR level 2 in listening, level 2 in reading and level 1+ in speaking (commonly abbreviated 2/2/1+). Listening and reading scores are measured by the DLPT; speaking scores are measured by the Oral Proficiency Interview (DoD Instruction 5160.71, 2019).

To put these expectations in context, these levels exceed the language proficiency levels typically found among university graduates. Tschirner (2016) reported that French and Spanish undergraduate students' reading and listening fourth year scores ranged between ACTFL Intermediate High (IH) and Advanced Low (AL) (approximately ILR Level 1+/2); only Spanish university reading scores reached the threshold of ILR 2. Russian and Japanese scores after four years were even lower, ranging from Novice High in Reading to Novice Low for Russian Listening (Tschirner, 2016). The comparison with university graduates applies to speaking as well, where university graduates appear to have difficulty reaching ILR Level 2. Glisan et al. (2013) reported that only 58% of the teacher candidates met the minimum oral proficiency levels (IH or AL, depending on the language) established by the ACTFL/NCATE Program Standards for the Preparation of Foreign Language Teachers. These statistics show that the DLI intensive language program results in proficiency levels that meet or exceed those of four-year college programs.

Working in the language

Following graduation from DLIFLC and advanced technical training, military language professionals enter the government workforce, where even higher levels of

language and cultural proficiency are needed to be successful. As Coakley (2016) explained in a chapter about language training at DLI, graduates who were '2/2 linguists' (as they are known) have been successful on their jobs in the past, but more recently the military services have sought both higher proficiency and more critically thinking service members. "Skilled language professionals bring enormous value to the defense enterprise; when it comes to nuanced understanding of intent, the skills of a single linguist testing at 3/3/2 … can surpass the work performance of any number of 2/2/1+ linguists." (Coakley, pp. 201-2).

Summary

The recruitment and training required to become a military linguist is time-consuming and expensive. It is no wonder, therefore, that the DoD is interested in ensuring that only the most qualified candidates are selected to attend language training. At the time of its original publication in 1976, DLAB was found to be a reliable and valid predictor of academic success in terms of grade point average, as well as a predictor of academic attrition, at the intensive language program at DLIFLC (Petersen & Al-Haik, 1976), even after accounting for general aptitude as measured by ASVAB. These two outcome criteria, graduation and attrition, are important to the US military. Selecting service members who are likely to be successful in military language training is key to maintaining the staffing of critical language positions across the DoD.

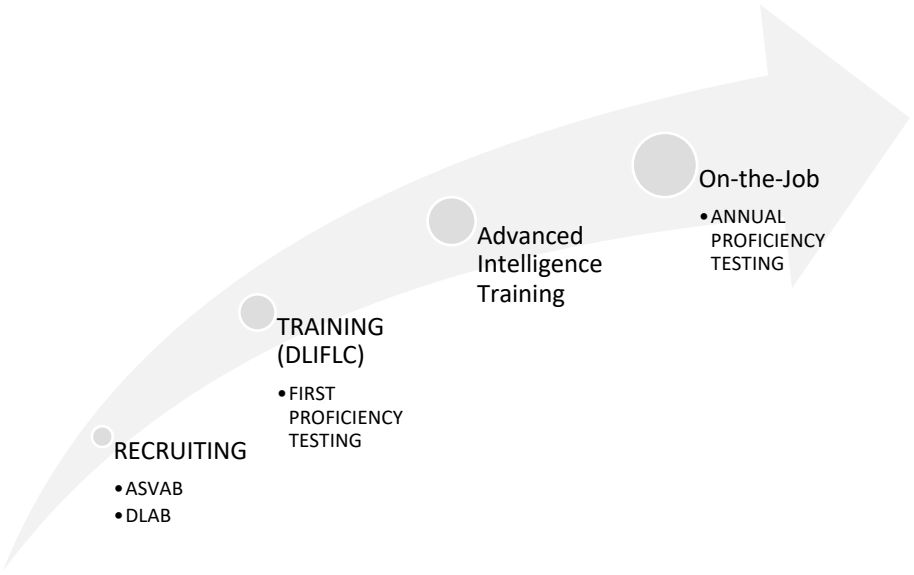
DoD continues to employ the DLAB to screen and assign, and DLIFLC has tracked and reported statistics for decades to internal government stakeholders regarding student graduation rates. Languages studied at DLIFLC are grouped into

four categories according to language difficulty, with the easiest languages assigned to Category (Cat) I and the hardest into Cat IV. Language difficulty will be addressed in more detail below. Little information is available to the public regarding the success of the DLAB in predicting graduation, but one early published example perhaps illustrates why DLI has continued to rely on language aptitude scores for selection. Lett et al., (2003) reported the following statistics regarding DLI graduation success at ILR 2/2/1+: a student who placed into a Cat I language with a DLAB score up to 94 was 76% likely to meet the DLI graduation standard; a student in a Cat I language with a score of 115 or higher was 96% likely to graduate. Similar comparisons were reported for Cat III languages, and for the most difficult languages, Cat IV, students who met the minimum DLAB requirement for this category were only 68% likely to meet the graduation standard but were 87% likely to graduate if their DLAB scores were 30 points higher than the minimum. Similar specific breakouts by language category for subsequent years have not been published.

In summary, the U.S. military services each recruit and train new enlistees each year to fill critical positions in military intelligence as the process is illustrated below:

Figure 1

Military accession and training process



As the Defense Language Transformation Roadmap explained in 2005,

> Establishing a new "global footprint" for DoD, and transitioning to a more expeditionary force, will bring increased requirements for language and regional knowledge to work with new coalition partners… This new approach to warfighting in the 21st century will require forces that have foreign language capabilities beyond those generally available in today's force. *P. 10*

As explained above, selection and training are not the end of the pipeline – they are just the beginning. Understanding how to select recruits who will not only be successful in language training, but who continue to grow their language skills beyond the classroom, is also key to the nation's defense. This research considered how current selection batteries could be leveraged to best select those individuals who would be successful in the long term.

Chapter 2: Literature Review

This chapter will discuss the literature on two components of cognitive ability, general aptitude and language aptitude. It will then review the literature on language growth and the impact of language distance on learning difficulty. The chapter begins with a summary of the research on ASVAB and DLAB and discusses the relationship between the two. It then moves on to review the current literature on proficiency growth and concludes with a section discussing language difficulty.

The Armed Services Vocational Aptitude Battery

Since 1976 the Armed Services Vocational Aptitude Battery has been used by the Department of Defense for both screening and classification (ASVAB history of military testing, n.d.). Recruits must meet minimum score qualifications to enter an individual service (Air Force, Army, Marines, Navy), and higher scores qualify recruits for a variety of different career fields. The full battery has historically contained ten subtests which cover four domains: Science/Technical, Math, Verbal and Spatial (see Table 1).

Table 1
Armed Services Vocational Aptitude Battery subtests (Culver, n.d)

Test	Description	Domain
General Science (GS)	Knowledge of physical and biological sciences	Science/Technical
Arithmetic Reasoning (AR)	Ability to solve arithmetic word problems	Math
Word Knowledge (WK)	Ability to select the correct meaning of words presented in context and to identify the best synonym for a given word	Verbal
Paragraph Comprehension (PC)	Ability to obtain information from written passages	Verbal
Mathematics Knowledge (MK)	Knowledge of high school mathematics principles	Math
Electronics Information (EI)	Knowledge of electricity and electronics	Science/Technical
Auto Information (AI)*	Knowledge of automobile technology	Science/Technical
Shop Information (SI)*	Knowledge of tools and shop terminology and practices	Science/Technical
Mechanical Comprehension (MC)	Knowledge of mechanical and physical principles	Science/Technical
Assembling Objects (AO)	Ability to determine how an object will look when its parts are put together	Spatial

*AI and SI are combined into AS in the computer adaptive test, so there are nine subtest scores in the studies below.

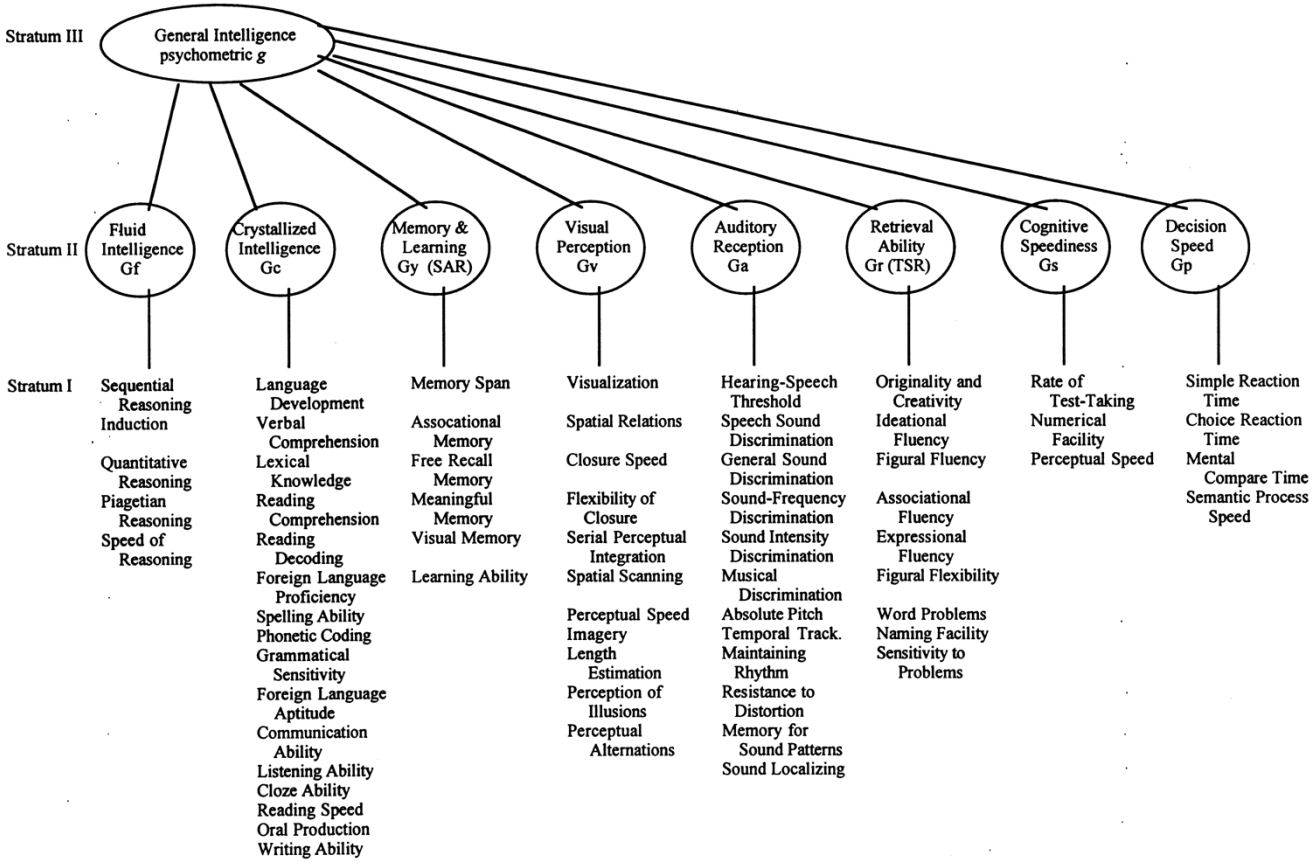
Each of the military services has a different set of minimum requirements to be considered eligible for a language career field, but a qualifying score on one or more ASVAB subtests (as well as minimum DLAB scores) are required by all four services. All four services rely on different combinations of ASVAB subtests to make these decisions, but the specific measures used do overlap: the services share an interest in a Verbal score (based on WK and PC) and one or both math subtests (AR and/or MK) (Schmitz et al., 2009).

Several different composite scores have been developed over the decades, and the two most commonly reported in language related studies are the Armed Forces Qualifying Test (AFQT) and the General Technical (GT) scores. AFQT is a composite drawn from arithmetic reasoning (AR), math knowledge (MK), word knowledge (WK) and paragraph comprehension (PC). A Verbal Score (VE) is formed from an optimally weighted composite of unrounded WK and PC standard scores. This Verbal Score, in turn, is double weighted in the computation of AFQT scores: AFQT = AR + MK + 2(VE) (Ostrow, 2002). The GT composite is formed by equally weighted VE and AR scores (Culver, n.d.). Over the decades since the introduction of ASVAB, subscores and composite scores have been used for selection and research.

The cognitive skills measured by instruments such as ASVAB are usually described as a hierarchy of abilities. Carroll (1993, as cited in Roberts et al., 2000) modeled intelligence in a three-level, hierarchical structure of human cognitive abilities, positing eight abilities at the second strata which enable the top level, psychometric *g* (see Figure 2 below). A crystallized general intelligence factor at the second level influences verbal abilities, including many specific foreign language abilities, while fluid intelligence drives abilities in quantitative reasoning and speed of reasoning.

Figure 1

Carroll's three-stratum model of the structure of human cognitive abilities (picture from Roberts et al., 2000, p. 86)



Factor analyses of ASVAB results typically find four factors (Verbal, Clerical, Mathematical, Technical) that are correlated, and researchers agree that the battery measures a significant amount of general cognitive ability (Ree & Earles, 1990; Welsh et al., 1990). Roberts et al. (2000) examined ASVAB in the context of fluid intelligence (Gf) and crystallized intelligence (Gc) theory, finding that ASVAB is biased towards Gc (p. 87). More recently, Martin et al. (2020) focused their attention on adding additional measures of fluid intelligence to the ASVAB to improve selection by broadening the underlying construct of intelligence currently measured by the battery.

Numerous studies over the years in the literature have shown ASVAB to be predictive of training outcomes, first term attrition, and job performance in the military. Welsh et al. (1990) investigated three types of training outcomes, final grades, training attrition, and time to completion. They summarized analyses from six previous studies, finding that aptitude indices (a variety of subtest and composite scores) seemed to make the largest unique contribution to prediction of success in training, but other variables made more of a contribution to other criteria, such as job performance or first-term attrition (p. 39). Ree & Earles (1990, 1991) found that general ability was effective in predicting technical training (1990, p. 10 and 1991, p. 330). Hunter (1986) reported the "massive evidence" that general cognitive ability predicted training success and job performance across a range of manual and mental jobs (p. 340) and he dismissed other findings that supported differential aptitude. His meta-analysis drew upon civilian and military validation studies. Cognitive ability was defined as "usually measured by summing across tests of several specific

aptitudes, usually verbal aptitude, quantitative aptitude, and sometimes technical aptitude" (p. 341). Hunter found that for military job families, the average validity of general cognitive ability predicting training success were high, ranging from $r = .58$ for clerical jobs to $r = .67$ for those in the electronic field. His findings showed that the more complex the job, the better cognitive ability is in predicting performance ratings. For example, a meta-analysis of the U.S. Employment Services data showed that for general job families, general cognitive ability predicted a high complexity job with a correlation of $r = .58$ as compared to a low complexity job where the average correlation is $r = .40$. The lowest category in industry families had a validity of $r = .23$. (p. 344).

Other large scale studies have found that both general cognitive ability and specific cognitive ability measures predict training success and job performance, both in the United States and in Great Britain. Lang et al. (2010) found that general mental ability predicted job performance, though narrower cognitive measures also played an important role. In a UK meta-analysis, Bertua et al. (2005) found that more complex jobs demonstrated higher operational validities between cognitive measures and both outcome measures (training and job performance). These general studies, though large in scale and across many military occupations, did not include language professionals in their sample focal populations.

In smaller, more focused studies, ASVAB has been shown to predict training outcomes and attrition for language training programs (Peterson & Al Haik, 1976; Silva & White, 1993; Wagener, 2016; Watson et al., 2012), but very few studies have investigated job success, or even how long during a career ASVAB remains a valid

predictor of language performance. There is a need to look beyond training success to determine how well cognitive ability predicts language growth. As Schmidt (2012) noted, “the more different lines of validity evidence supporting test use and interpretation, the stronger is the foundation for an inference of validity” (p. 7).

Defense Language Aptitude Battery

The idea that there is something we can specifically call *language aptitude*, that is, a set of cognitive abilities that are different from general cognitive abilities, is generally accepted (Doughty, 2019; Li, 2015, 2016; Skehan, 2019; Smith and Stansfield, 2016). Much of the research in the field of second language acquisition has been focused on how language aptitude predicts foreign language learning, and the primary measurement instrument in the general literature has been the Modern Language Aptitude Test (Carroll & Sapon,1959).

The Modern Language Aptitude Test (MLAT) assesses phonetic coding ability, grammatical sensitivity, rote learning ability and inductive language learning ability (Doughty, 2019). In two large meta-analyses drawing primarily from MLAT-based studies (14 out of 17 primary studies in 2015, 57 out of 66 studies in 2016), Li (2015, 2016) analyzed the construct validity of language aptitude and found that general aptitude and language aptitude are different constructs, though they are correlated. Li (2016) concluded, “given the overlap between the two constructs, it might be necessary to consider both aptitude and intelligence when making pedagogical decisions related to selection, guidance, and placement” (p. 827). He reported a strong association of aptitude and L2 proficiency, with 25% of the variance in learning outcomes accounted for by language aptitude (p. 829). As expected, given

MLAT's origins, aptitude was more strongly correlated with proficiency in studies of instructed SLA with high school learners rather than with university learners (Li, 2015).

It may well be that language aptitude predicts differentially at different stages of learning. The High-Level Language Aptitude Battery, Hi-LAB, was designed with this in mind to complement, rather than replace MLAT. It was also designed to look not just at prediction, but also to accommodate aptitude-by-treatment interaction. Findings from early Hi-LAB studies (Linck et al., 2013; Doughty, 2019) showed that composite scores predicted outcomes for a population of high level adult learners across a variety of languages. With results suggesting that composites' predictive validity varied by outcome (listening, reading, or speaking) and language difficulty, further research was warranted on how aptitude measures could be used to predict success across the lifecycle of a language professional and not just during training, or as a predictor of training outcomes.

Abilities that predict learning success may well vary from those that predict growth or ultimate attainment. Welsh & Kucinkas (1990) found that as compared to specific aptitude area composites, the more general AFQT composite had lower validity coefficients for job performance than for training school grades, concluding that "it may be that general trainability becomes a less important factor later in time, and specific abilities and/or experience assume greater importance the longer a recruit stays in a particular job. That the more specific aptitude area composites predict such criteria better than the AFQT does is interesting and needs to be explored in future

validity research." (pp. 42-43) This research included the ASVAB and DLAB subtest scores, rather than overall composite scores, to explore growth.

Within the military, the more specific cognitive aptitude of interest, language aptitude, is measured by the Defense Language Aptitude Battery (DLAB), which was developed to maximize predictive validity as a selection device and to differentially predict training outcomes by language (Peterson & Al-Haik, 1976). Pulling together research on language aptitude conducted in the 1950's and 1960's, the test was designed specifically for DLIFLC with practicality and validity concerns in mind. Given the military testing system and goals, the test had to be practical and efficient, while maintaining the ability to predict learning outcomes for all languages taught at the Institute. The test takes approximately an hour and a half to administer. It includes four parts: biographical data, spoken stress, deductive rule application and inductive pattern application (Lett et al., 2003).

The original DLAB design drew upon Carroll's research on language aptitude, and it has not changed its composition since its original publication in 1976. The first part includes several biographical questions designed to elicit information about previous language learning and academic background. The three remaining parts were developed from exploratory factor analysis of items from two experimental batteries, the Al-Haik Foreign Language Auditory Aptitude Test (AFLAAT) and Horne's Assessment of Basic Linguistic abilities (HABLA). AFLAAT assessed the ability to distinguish foreign language sounds, associate sounds with symbols, and apply ever-complex grammar rules to new utterances. HABLA required the examinee

to generalize new linguistic forms in an artificial language from pictures (Peterson & Al Haik, 1976).

The developers' goals were to improve upon the predictive power of the MLAT, provide for differential prediction of success by language, and examine other predictor variables (Peterson & Al-Haik, 1976). Drawing on research available at the time, Peterson and Al-Haik developed a final battery that was heavily influenced by the practical needs of the military, yet achieved "maximum predictive validity". Their analysis resulted in three factors that then drove item selection and resulted in a four-part test. After an introductory part on language biography, DLAB Part 2 involves recognizing stress patterns. Part 3 requires test takers to translate a printed phrase according to a set of rules from an artificial language and match it to a spoken utterance. The items assess grammatical rules such as noun and adjective agreement, formation of possessive phrases, and sentence structure. The fourth and final part of the DLAB combines the rules provided in the earlier parts and asks the examinee to apply them at the same time (Bunting et al., 2011).

The zero order correlation of the resulting battery with the final Grade Point Average (GPA) of 879 students was found to be .43. DLAB reliability (KR-21) was reported as .89 (Peterson & Al-Haik, 1976, p. 378), and no reliability estimates were provided for GPA. There was no correction for range restriction, which means that the correlation may well have been higher, if data from a full range of scores had been considered. The reliability of the DLIFLC GPA is unknown, but the reliability of grades is likely to be less than that of a high-stakes, standardized test (Westrick, 2017). Rather than GPA, subsequent researchers (Bush, 1987; Jackson et al., 2011;

Silva & White, 1993) used standardized language proficiency tests as the outcome measure (the DLPT), although at least one (Wagener, 2016) also used GPA. The current study used proficiency scores as the outcome measure, not only due to its reliability, but also because they had the advantage of being available over time.

The DLAB has been in constant use since its publication, and it serves as an effective screen for entry into the military language field. The success rate has varied over time, but those scoring over 85, historically the minimum score to attend language training, was reported as approximately 50% (Schmitz et al., 2009, p. 11). The minimum DLAB score has changed over the years as the services balanced the need to fill language positions with the evidence that higher DLAB scores increase success. Despite the efforts to increase the likelihood of success at DLIFLC (by requiring higher DLAB scores, adding time to training courses), descriptive analyses show that graduation rates still vary based upon language difficulty, with 80% reaching minimum DLPT goals in Cat I languages, yet only 60% in Cat IV languages (Schmitz et al., 2009). Findings such as these suggested the need for further analysis of how aptitude measures are used, not only for selection into the career field itself, but also for language assignment.

There has been interest in updating the DLAB to bring the test into the 21st century. In the 2010's, the Center for the Advanced Study of Language (CASL) developed an updated DLAB, the DLAB2 (Bunting et al., 2011). DLAB2 was administered to more than 1300 service members, and the data were analyzed in a series of logistic regression models. The dependent variables in the final analyses were course completion and attrition, rather than ILR levels. In their final analyses,

the authors recommended keeping either DLAB Part 3 or DLAB Part 2 and 4, four ASVAB subtests (ASVAB-AR, ASVAB-MK, ASVAB-PC and ASVAB-WK that contribute to the AFQT composite), a new working memory measure, an explicit induction measure and a number of non-cognitive measures (especially prior language experience) (Bunting et al., 2011). We return to these analyses in a later discussion on the relationship between general aptitude and language aptitude.

Aptitude research studies in the literature generally fall into one of two categories, predictive or interactional, and the predictive research is focused on understanding the influence of aptitude on learning. Interactional studies involve relationships between individual differences and aptitude, often using an aptitude-by-treatment approach to determine how aptitude mediates learning in different contexts. (Li, 2019) The focus of the current research is on aptitude's role in predicting outcomes, specifically addressing the role aptitude plays in predicting language growth beyond graduation.

For the U.S. military, how aptitude, both general and language, predicts training outcomes is useful information, but it should also be of interest to look beyond the classroom. Sustainment, or better still, improvement, of language skill acquired at DLI is key to job success in a language profession, yet very few studies have taken a longitudinal approach. One early study was the Language Skill Change Project (Bush, 1987), though it was limited in scope to four languages and a three-year period. The project was sponsored by the US Army and was designed to follow DLI graduates from graduation through their first enlistment to better understand the rate and nature of post-DLIFLC skill change, which factors influenced skill change,

and how language proficiency and job performance were related. The Language Skill Change Project (LSCP) collected data on 1900 Army soldiers who studied German, Korean, Russian, and Spanish at DLIFLC in 1986 and 1987 and followed them for an additional three years. Research variables were collected for the study, including demographic, affective, and cognitive measures. The outcome measures were language proficiency (as measured by the DLPT of the day, the third generation or DLPT III), academic data from both language training and advanced intelligence training, and supervisor performance ratings. In contrast to the original DLAB study, the LSCP included both GPA as well as language proficiency test scores as training outcome measures.

Due to the large number of data points, the demographic and affective predictor variables were condensed; 13 of the measures were consolidated following principal component analysis into three: total years of education in another foreign language, motivation at the start of training and motivation during training at DLIFLC. Multiple regression was then used to analyze which measures predicted two outcomes: training attrition and proficiency. Variables were entered in blocks based upon practical considerations, such as which variables were already available, and which could be obtained for low cost. For example, ASVAB and DLAB were entered first, because they were collected by the services at the recruiting stage. The squared multiple correlations ranged from .15 to .48 (O'Mara et al., 1994, Vol II, p. 13), confirming that the measures analyzed predicted initial proficiency outcomes, though at quite varying rates. Attrition was less predictable than proficiency.

Predictability varied considerably by language and criterion, but ability – defined as the ASVAB-GT composite, DLAB, and other abilities - consistently and significantly predicted outcomes at DLIFLC, although the change in the prediction model was quite small and varied by language. ASVAB-GT was stronger for the easier languages (German and Spanish) and DLAB for the harder (Korean and Russian). Non-cognitive and affective measures collected specifically for this study also showed additional incremental validity, with increases (also small) once again varying by predictor and criterion. Despite these significant findings, even the final models accounted for only an average of 27.1% of the variance in attrition and proficiency, meaning that a significant amount of variance in graduation outcomes remained unaccounted for (p. 16).

General aptitude (as measured by the ASVAB-GT composite) predicted training outcomes and attrition in all languages, and language aptitude contributed to the prediction equations for listening, reading, speaking proficiencies as well as attrition, above and beyond what was predicted by general aptitude alone, in all but two cases (Listening-Spanish and Speaking-Korean). The study's scope was limited, though, as data was drawn only from one military service (Army) and for only four languages (German, Korean, Russian, Spanish). To improve upon these earlier studies, data for the current study was collected from all four military services and for languages taught over the last decade at DLIFLC.

After three subsequent years of data collection, the Language Skill Change Project modeled language growth post-graduation using MANOVA. It was expected that if foreign language skills were to deteriorate, the largest loss would occur in the

first year, and that speaking proficiency, a productive skill, would suffer more than listening or reading, the receptive skills. The data confirmed a sharp drop at end of advanced intelligence training and a gradual increase thereafter (except for Spanish that showed a general upward trend), although many service members failed to ever regain the level of proficiency they had upon graduation. The greatest skill loss was found in the more difficult languages (Russian and Korean) (O'Mara et al., Vol IV, 1994). The generalization of the findings is limited, however, due to the small sample size in each language in each subsequent year due to attrition, which dropped from almost 2,000 at the start of the project to 441 soldiers who tested four years later. Another shortcoming of the LSCP is that the relationship of the predictor variables to outcomes was only analyzed for training outcomes, not for subsequent growth.

Almost two decades after the initial publication of the DLAB, Silva and White (1993) visited the question of incremental validity, looking at the additional improvement in prediction afforded by DLAB over ASVAB. Silva and White's study followed a decade of large studies of differential aptitude that were published in the 1980's and 1990's (Schmidt, et al., 1988; Ree & Earles, 1991). These studies, based on ASVAB subtests and composites, found that tests of specific cognitive ability made only marginal improvements in the prediction of training outcomes and job performance. The language career field was not considered in these studies.

With all military service members taking the ASVAB by the early 1990's, Silva and White were able to examine whether a specific cognitive aptitude, in this case language aptitude, as measured by the DLAB, contributed over and above the ASVAB subtests, or by a general cognitive ability factor drawn from the ASVAB.

They investigated the following predictors: the ten ASVAB subtests, a *g* factor score based on ASVAB subtests, and the DLAB. Unlike the Peterson and Al-Haik (1976) study which used GPA, their criteria were academic attrition and a standardized proficiency test, the DLPT. Silva and White (1993) found that DLAB did provide incremental validity in the prediction of both proficiency and attrition for each language difficulty category, though the gains were small: depending on skill modality, correlation coefficient gains ranged from .01 to .13 (Silva & White 1993, p. 89). The strongest predictions were found when general cognitive aptitude was modeled as a general intelligence factor, rather than with the ten individual ASVAB subtests. The authors concluded that language ability is distinct from general intelligence and suggested that DLAB taps into abilities not measured by tests of general ability. Unlike the LSCP findings, (which were not published until 1994), Silva and White (1993) found that the incremental gains in listening and reading were much lower than those for speaking.

The implication drawn by the authors was that the differential aptitude measured by DLAB tapped strategies "to extract and organize the semantic, syntactic and phonemic structure of language, constituting a specific kind of crystallized ability with predictive power beyond that of *g*" (Silva & White 1993, p. 91). Though significant for the services, their study was also limited to training outcomes, and they did not extend the research timeline to look at growth.

Beginning in the mid-2000's, a team of researchers sponsored by the U.S. Special Forces published studies on predictors of language proficiency with the goal of improving selection and training assignments. The student pipeline for Special

Operations Forces (SOF) is similar to the one described above, but rather than attending the DLIFLC, these military language students receive their training at the United States Army John F. Kennedy Special Warfare Center and School in North Carolina. Program emphasis is slightly different, with graduation outcomes expected to be ILR Level 1/1+ or 1+/1 or higher on the two-skill Oral Proficiency Interview, which assesses listening and speaking. Once in the field, SOF linguists continue to take the two-skill OPI, and the DLPT in listening and reading. The researchers' findings, summarized below, consistently support the use of DLAB to predict language outcomes (Surface et al., 2005; SWA 2009; Watson et al., 2012).

In an early example of multilevel modeling in the foreign language field, Surface et al. (2005) found that ASVAB (as measured by the AFQT composite) and DLAB (as measured by the standardized overall score) had a significant, positive relationship with SOF learning outcomes. Language students were nested by class and by instructor in a three-level model. The results showed that individual differences, specifically ASVAB-AFQT and DLAB, were significant predictors of learning outcomes, accounting for 13-24% of the variance in scores. At the next levels of nesting, proficiency scores did vary by class (level-2) and by instructor (level-3), and language difficulty was the only significant predictor of between-class variance.

Research on SOF linguist outcomes continued in the late 2000's, with studies continuing to refine analysis of predictors to improve selection. SWA (2009) found that of four predictors available to the study, DLAB, ASVAB-AFQT, Army General Technical (GT) and Wonderlic Personnel Test™, DLAB was the best predictor of the

two-skill, Oral Proficiency Test outcomes in speaking and listening. This study's conclusions, while confirmatory of earlier language aptitude research, were limited by a small and narrow sample size and number of languages.

A more comprehensive multilevel analysis a few years later. Watson et al. (2012) found the DLAB overall score was predictive of which SOF trainees were most likely to attain ILR 2 speaking proficiency following initial training, as well as the maximum proficiency level linguists attained over the duration of their career (p. 4). Of interest to the present study, the likelihood of reaching ILR Level 2 was related to language difficulty. Accounting for language, the DLAB score explained 4.3% of the differences in listening, 7.3% of reading, and 8.3% in maximum speaking outcomes. The conclusions are somewhat limited in their generalizability to the overall military linguist population, however, since on average, the highest levels obtained were in the range of ILR Level 2. The current study drew from a broader population, although the range was also limited, given that with the exception of Spanish tests in this dataset, the DLPT only measures up through ILR Level 3.

There have been several other studies that took a longitudinal approach. Wagener (2016) extended existing research on DLIFLC graduates to examine the effects of context on outcomes over time. He found that DLAB predicted grade point averages for the foreign language classroom context, even after including other general cognitive measures into a model. Looking at proficiency score outcomes over four years, rather than GPA, DLAB continued to be a predictor of listening and reading, along with two quantitative measures (ASVAB-AR and ASVAB-MK), but the magnitude of their coefficients decreased over the four cycles as test scores rose.

Findings such as these support efforts such as the Hi-LAB (Jackson et al., 2011; Doughty, 2019), which posits that different aptitudes are needed to predict ultimate attainment, rather than learning outcomes.

Confirming earlier studies, Wagener (2016) found that DLAB remained a significant predictor of GPA after accounting for ASVAB for any language taught at DLIFLC, replicating earlier studies (Peterson & Al Haik, 1976; Silva & White, 1993). He also found differences by difficulty category in the relationships among individual difference measures and GPA, this despite efforts by DLIFLC to lengthen courses in the "harder" languages and restrict these languages to only those with the highest language aptitude scores. Unlike other studies, Wagener (2016) used four of the ASVAB subtests, rather than composite scores, for his measurement of general cognitive aptitudes. He concluded that his findings supported differential aptitude theory.

Language Growth

While the question about how foreign language learning develops over time should be of interest to the government, few published studies exist. This is surprising, given that DLIFLC graduates serve at least for three or more years following graduation, and the language demands on the job are many. This section reviews the literature on how language proficiency changes over time, and what variables predict change. As mentioned above, prediction models have been primarily focused on initial learning outcomes, rather than on development. This is not just a pattern in the military community. Ortega and Iberri-Shea (2005) point out the need to broaden the timeline: "Indeed, it can be argued that many, if not all, fundamental

problems about L2 learning that SLA researchers investigate are in part problems about "time," and that any claims about "learning" (or development, progress, improvement, change, gains, and so on) can be most meaningfully interpreted only within a full longitudinal perspective." (p. 26) Despite intensive efforts to improve learning outcomes at DLIFLC, graduates still fall short of the proficiency levels demanded on the job, and they need further opportunities to improve their language skills.

While there is a need for such sustained and even improved language proficiency, the results show that in fact, gains are not common. The earliest longitudinal results of military language students appear to be from the Language Skill Change Project (O'Mara et al., 1994) mentioned above, which reported findings that after graduation, patterns varied by language, but mostly showed a drop in proficiency, followed by a slow increase. Not all graduates were able to recover their graduation levels, and soldiers in the more difficult languages had a higher incidence of skill loss (p. 2).

More recent studies, some of which were mentioned above, found similar drop-and-recover patterns. Surface et al. (2004) found that for SOF members, listening proficiency dipped and recovered. Their analysis considered several predictor variables, including language difficulty, education level and general cognitive ability. Initial proficiency outcomes were negatively impacted by language difficulty, and growth was constrained in the more difficult languages. While education level and general cognitive ability were predictive of initial proficiency outcomes, they did not predict growth, which was an unexpected finding. Their

analysis also found that language difficulty negatively impacted initial proficiency and constrained growth. Subsequent research below suggested that general cognitive measures do, at least to some extent, predict growth in other studies and therefore remained of interest in the current study.

To examine the language proficiency change of DLI graduates, Shearer (2013) used survival analysis. He found a probability of 25% that those graduating at ILR Level 2 in listening would drop at least a sublevel to Level 1+, with reading trajectories slower to attrite. Graduates with higher exiting ILR proficiencies were less likely to attrite, suggesting that DLIFLC's goal to increase proficiency levels for graduates in their basic program would have a positive impact. Updating this research with current data to model growth and investigate the relationship of aptitude measures to graduation outcomes and growth following graduation motivated this research. Data for this research included graduates up through 2018, so the likelihood of higher test scores at graduation was greater, given the increased emphasis on higher standards since 9/11.

Other studies have shown that patterns of growth differ by skill modality (listening and reading). Bloomfield et al. (2014) found different patterns of growth over four test sessions in both listening and reading. While models showed a correlation between the graduation outcome (intercept) and pattern of change (slope), the direction of the correlation was opposite: in listening, those with lower initial scores had a faster rate of improvement; in reading, those with higher initial scores improved more quickly. Separate analyses by language or by language category were

not conducted in this study, nor were predictor variables related to aptitude or language difficulty included in the analysis.

To investigate how aptitude relates to language proficiency outcomes over time, B. Mackey (2014) used latent growth curve modeling (LGM), which considers group level information (as expressed by factor means) as well as individual differences (the variances). LGM can be used to describe growth at the group and individual level, as well as to see which variables influence growth. As Duncan and Duncan (2004) explain, the strengths of LGM include: "an ability to test the adequacy of the hypothesized growth form… to correct for measurement error in observed indicators… and to develop from the data a common developmental trajectory, thus ruling out cohort effects" (p. 8). LGM allows for change to be linear or curvilinear and B. Mackey (2014) confirmed findings from earlier studies that a non-linear growth model best fit the data. The estimated proficiency level at the time of graduation from DLIFLC, the mean intercept, was just over ILR Level 2 for listening and close to ILR Level 2+ for reading. The mean growth (slope) in both listening and reading was positive, though not significant, indicating very little change over time. Significant variances indicated meaningful inter-individual differences in graduation levels, growth and the effect of predictor variables. Language difficulty category had a significant, negative impact on graduation and no significant relationship to growth, but her model did not account for clustering by language. Regarding the predictive power of aptitude scores, she found that ASVAB-AFQT had a significant positive relationship on graduation outcomes, but contrary to expectation, no significant relationship to growth for both listening and reading. The DLAB, in contrast, was

unexpectedly not a significant predictor of listening score upon graduation, but it did explain a small amount of between-person variation in change for those who did show growth. Unlike the listening data, however, for reading proficiency, DLAB was a significant predictor of graduation outcomes, and it had a positive significant relationship to growth. The findings related to aptitude were not all as predicted and were possibly due to the pooling of the languages in this study. A multilevel approach would be a more appropriate method to analyze data in which test scores are nested by individual and by language. Given more recent research, such as the role ASVAB-MK appears to play in predicting growth (Wagener, 2016), it would also be more useful to break out the ASVAB-AFQT score to its component parts to further investigate the role of specific cognitive aptitudes, rather than a composite, especially in a framework that accounts for differences that might be attributed to language difficulty.

Language Difficulty

There is no doubt in the literature that some languages are more difficult to learn than others. As Cysouw (2013) explains, "larger differences between languages are correlated with larger difficulty, though not all differences are equally important." (p. 52) What is equally true is that there are a variety of approaches to operationalize the construct of language difficulty. As mentioned previously, DLIFLC uses a four-category scale of language difficulty, with languages easiest for a native speaker of English coded as Cat I, and the hardest languages as Cat IV. According to Clark et al. (2016a), the DLI categories were based upon experiences at the Foreign Service Institute (FSI), the U.S. State Department's language school. While not a perfect

overlap of language-by-category, for the most part the two lists are similar. DLI has adopted two policies to mitigate the impact of language difficulty and improve the likelihood of on-time, on-standard graduation rates: first, only those with the highest scores on the DLAB are enrolled into Cat IV languages; second, training time now varies from 36 weeks for the easiest languages and up to 64 weeks for the hardest. Despite these policies, language difficulty continues to depress graduation rates (Bermudez-Mendez, 2020; Schmitz, 2009; Wagener, 2016; Wang, 2004) and increase the likelihood of being recycled or re-languaged (Schmitz, 2009).

There are indications that even within a DLI language difficulty category there is a range of difficulty. While Masters (2018) found "coherence" within one category (Arabic, Chinese and Korean, all Cat IV languages), Wagener (2016) found that the strength of the individual differences measures in predicting outcomes varied not only overall, but also within each language difficulty category. This finding was interpreted as support for claims regarding language learning difficulty (Child, 1998; Lowe, 1998) that individual difference measures related to general aptitude and language aptitude differentially predict outcomes, even within the same language difficulty category.

Other studies have looked beyond graduation and continued to find that language difficulty depresses outcomes. While B. Mackey (2014) did not find a significant relationship of language difficulty to slope in a latent growth curve analysis of language proficiency, Clark et al. (2016) used event history analysis and found that the odds of reaching ILR level 3 over time were higher for some languages than for others within the same category. Other longitudinal analyses showed that

within each of the three largest categories (I, III and IV), there were languages that were either easier or harder than the average. And when they compared all the languages in one analysis, they found the estimated difficulties of languages in different categories were virtually completely overlapping (p. 7). These are interesting findings, as they suggest that the current language difficulty categories impact not only language learning outcomes at DLIFLC, but that they also impact language growth. The findings also point to a need to better define language difficulty.

Outside of the government context, language difficulty has also been found to moderate learning outcomes in academic settings (Elder & Davies, 1998; Lee & Kim, 2010; Snow, 1998; Verhoeven et al., 2019; Zhang, 2019). In the literature, differences in L1-L2 are more often described descriptively in terms of distance between two languages, rather than as related to learning difficulty, and researchers have suggested a variety of methods for categorization of the differences. The approaches range from the philosophical (Mackey, W.F., 1971) to the statistical (Gamalloa, et al., 2017). Neither extreme is helpful for the present research, though they raise interesting questions that could drive a potential new categorization system: what is the context [second language learning or foreign language learning]? How does the learning occur [direct or indirect]? What outcomes are measured [listening, reading, speaking, or writing]? What is the approach [synchronic or diachronic]? Are the measures linguistic or behavioral? Who are the learners?

There are few other categorization schemes of distance in the literature. Child (1998) described a multidimensional matrix, with a three-scale distance measure

(1, 2 or 3) in each of three categories: orthography as a representation of the spoken language; grammatical system as framework for communication; and semantic system as cultural outlook. His matrix has not been used in research. In a similar approach, Ross (2000) operationalized a scale of language distance based upon orthography (alphabetic, syllabic, ideograph), canonical word order (SVO, SOV, OSV, etc.) and typological grouping (e.g., Germanic, Slavic, Altaic, Sino-Tibetan, etc.). Both Child and Ross assigned their ratings with languages closest to English (e.g., German) given the highest proximity value and the most distant languages (e.g., Cantonese) the lowest. Ross (2000) found little to no effect of language distance on speaking and writing outcomes.

Other researchers have reduced language distance to only one or two categories, which then restricts the type of analyses that can be used to a categorical, rather than continuous measure. In their meta-analysis of cross-linguistic transfer of oral language (decoding, phonological awareness and listening comprehension), Melby-Lervåg and Lervåg (2011) used a measure of distance based on writing system (alphabetic or idiographic) and found that whether both L1 and L2 were alphabetic script did moderate the magnitude of the L1-L2 correlation in decoding skills, but not listening comprehension. Jeon and Yamashita (2014) used two measures of distance, the alphabetic/non-alphabetic distinction, as well as Indo-European or not, as moderator variables in their meta-analysis of L2 reading comprehension and ten correlates. Contrary to their hypotheses, neither measure had a significant moderating effect on correlations, though a deeper study of age and specific L1 features did suggest an influence on

decoding skills. They did, however, find a stronger correlation for L1 reading on L1-L2 language distance for those languages in the Indo-European family than for those which were in a different language family. Jeon and Yamashita (2014) suggested that morphosyntactic knowledge varies even within a language family, so this could explain why they did not find a moderator effect in the subskill analysis but did in the higher level process of reading.

Rather than a typological approach, Chiswick and Miller (2004) used FSI outcome data to develop a continuous language distance measure based upon language test scores that they then applied to their analysis of English proficiency among immigrants in Canada. Their "linguistic distance" measure ranged from .33 (Afrikaans, Swedish) to 1.0 (Japanese, Korean). Holding other variables constant (level of education, age, duration in country) they found that the larger the distance between the immigrant's L1 and English, the lower their English proficiency.

There are other possible ways to describe language difficulty. Cysouw (2013) used two measures of language learning difficulty, the Chiswick and Miller (2004) measure (hereafter CM), as well as a revised FSI measure (one with seven, rather than three categories), to compare other possible measures of language distance: geographic, genealogical classification, typological similarity, orthographic similarity, and size of the orthographic system. Geographical distance was significantly correlated with both FSI-levels and CM (p. 40); genealogical differences, operationalized with only two levels, showed that closely related languages were, as expected, easier to learn. Cysouw (2013) assessed the relative contributions of three factors (geographical distance, non-Germanic, non-Indo-

European) and found that while both genealogical levels were significant, geographical distance was already accounted for to a large extent by genealogy (p. 41). Next, he considered orthography. Two measures were developed, a similarity between English and the target language, and the size of the orthographic system. As expected, he found that the more different a script is from English, the more difficult the language is to learn. The size of the orthographic inventory was more challenging to model: Korean and Arabic were outliers in the initial correlation, where no trend was found, but removing these languages resulted in a highly significant correlation with both outcome measures (FSI-levels and CM) (p. 44). Finally, Cysouw developed an overall, measure of structural differences based on a set of complex descriptive linguistics. Once again, the correlation was found to be as predicted, with a strong negative correlation between English and languages with very different structures.

After building and analyzing several prediction models, Cysouw (2013) concluded that the categories for typological and orthographic similarity resulted in a reasonably good prediction. The problem he found with the typological category is how difficult it is to develop, as not all languages are well documented, but he did share his ratings in an appendix, and they are available for researchers. A second approach was determined to be a more practical model to predict learning difficulty, relying on a combination of four binary factors (Latin script or not; Germanic or not; Indo-European or not; and same structure as English based only on 11 features).

Additional typological categories not mentioned above warrant mention here: phonological, morphological and lexical. Schepens et al. (2013) investigated the degree to which morphological difference affected L2 Dutch learning. Using data

from the World Atlas of Language Structures (Dryer & Haspelmath, 2013) , the authors developed quantitative measures of morphological similarity and complexity (based upon 29 features). They showed that morphological similarity and increasing complexity are closely related, and that the relationship with speaking proficiency is significant and negative. While their study included a large variety of languages (49 different L2), the quantitative measures cannot be used as is for the current study because Dutch, rather than English, was the reference language for their distance measures. Recreating the morphological measures for the languages in the current study is beyond the scope of this research.

One existing categorization system that takes these types of features into account is the language similarity measure available through CASL's Gateway Language Database. This measure was used in two of the studies mentioned above, Clark et al. (2016) language difficulty research and Doughty's (2019) article on cognitive language aptitude. The Gateway language similarity number is based on a comparison of features such as script, sound, word formation, word types, word order and sociolinguistic factors (Gnanadeskikan & van Rossum, 2016). Rather than taking the perspective of difficulty, the Gateway number is a measure indicating likelihood of successful cross training based upon language similarity.

For the current study, the typological measure developed by Cysouw (2013) and the Gateway Languages Database (2016) were used to provide more detailed, continuous measures, along with the four-level DLI categories, the FSI categorical system as modified into seven categories by Cysouw (2013), as well as two binary

measures, whether the languages tested use a Latin script and whether they are in the Indo-European language family.

Summary

The literature points to the need for longitudinal research to better understand proficiency growth, and there is some mixed support suggesting that general aptitude and language aptitude are predictive not only of training outcomes, but also of language growth. Given the military's investment in language and the critical role linguists play in our national security, it is important to extend the earlier research to further investigate the relationship of aptitude to growth, especially using methodologies that account for the variety of languages taught at DLIFLC.

Chapter 3: Purpose of the study

With an investment of more than $350 million dollars per year (DLIFLC Annual Program Review, 2012) in DLIFLC's language program, the DoD is motivated to ensure that only those students likely to succeed are, in fact, enrolled. Even small changes in the number of students who drop out of the program will generate savings (Bunting et al. 2011). ASVAB researchers have commented that even small improvements in predictive validity, i.e., as low as .02, have the potential to produce substantial cost savings (Held et al., 2014, p. 214) in selection of individuals for training programs.

However, with ever-increasing demands on DLI's graduates to improve their language skills beyond the levels reached upon graduation, the DoD should also be concerned with language proficiency growth post-DLI during a service member's full enlistment, not just at graduation. It may be that what predicts initial training success is different than what predicts growth over time. For training outcomes, DLAB contributed significant incremental validity beyond ASVAB in older studies (Silva & White, 1993), but does this continue to over time? Furthermore, which subtests best contribute to prediction models in longitudinal studies?

Research methodology also provides a warrant for the current study. As previously mentioned, few studies have taken a multilevel approach to the modeling of change data. However, repeated test scores are strong hierarchies, as there is potentially more variation between individuals than within individuals (Barkaoui, 2013). Assignment to a specific foreign language is yet another potential level of nesting. The advantages of multilevel modeling over more traditional analysis

(MANOVA or multiple regression) include the simultaneous modeling of both intraindividual change (how an individual changes over time) as well as interindividual change (differences in change across individuals) (Finch et al., 2019).

Repeated measures are, by their very nature, nested by individual. Ordinary Least Squares regression analysis assumes that measures are not correlated, so it is not appropriate to use this method when the data are repeated measures unless robust standard errors that allow for clustering are used (Cheslock & Rios-Aguilar, 2011).

Language professionals are inherently also nested by language in and beyond the DLI classroom, as they continue to train and work in that language. Rather than combining data into one pool, or analyzing data separately by language, multilevel modeling allows for the simultaneous analysis of test scores over time, nested by individual and again by language. One could model DLI graduates' repeated measures as scores nested within an individual, then nested by classroom and then by language to avoid adding bias to the estimates of standard error. While the classroom at DLI could serve as to nest the data in this way as a layer between the individual and the language, there are many confounding variables associated with the classroom. For example, teaching teams are not consistent and course material varies. According to DLIFLC, who has tried to measure the impact of classroom, there is too much noise in the classroom data (Dr. Seumas Rogan, Chief of Testing Design and Analysis, DLIFLC, personal communication, September 13, 2021). The relevance of classroom data beyond DLI is also questionable. It is less intuitive to consider how language would serve as a nesting variable, but DLI graduates in the field use the

language on the job and continue their language learning. This might lead to between-language differences that impact outcomes.

This dissertation investigated the role that general aptitude as measured by ASVAB and language aptitude as measured by DLAB played in predicting language proficiency growth using a multilevel approach. The following research questions were addressed:

RQ 1: To what extent is there variance in language proficiency growth over time, across individuals and languages?

RQ 2: What is the shape of language proficiency growth and how does it vary by language?

RQ 3: To what extent does language aptitude predict language proficiency growth outcomes across languages, beyond what is predicted by general aptitude?

RQ 4: To what extent does language difficulty categorization impact language proficiency growth across languages?

RQ 5: To what extent does aptitude interact with language difficulty?

Expected findings

Based upon the literature review, the following hypotheses were made relative to the research questions above:

Hypothesis 1: There is significant variance in language proficiency at time of graduation from DLIFLC and over time, both within individuals and languages.

There is ample support in the current literature for a finding of significant variance in language proficiency both at the time of graduation, as well as over time (Bloomfield et al., 2014; Masters, 2018; Wagener, 2016). This research aimed to clarify this variance by partitioning it at the levels nested within the data: by individual and by language. It was expected that the inter- and intra-individual

variance would remain, even after accounting for language studied. Studies using latent growth modeling or multilevel modeling have shown that there can be significant variance around growth, meaning that even though the overall mean growth may not be significant, there could be significant variation among individuals in their growth. It was expected that this variation would not be due to language difficulty over time, given how languages are assigned to individuals, and that this variance would remain significant.

Hypothesis 2: Language proficiency growth is non-linear.

Several studies (Bloomfield et al., 2014; Shearer, 2013; Mackey, B., 2014) have shown that the pattern of language proficiency scores for military personnel following graduation is generally flat over time, with a slight U-shaped downward curve between the first and third scores. Proficiency scores are generally lower at the time of the second test, but then often recover by the third test, though not all individuals regain their graduation level. It was therefore expected that Hypothesis 2, that language proficiency growth is not linear, would remain to be true when the multilevel structure of the data is accounted for.

Hypothesis 3: There is a significant positive relationship between cognitive ability and language proficiency over time.

Existing research on the relationship between cognitive ability and language proficiency over time is less clear, and most studies have ignored the nested nature of the data. Due to findings from studies in other fields (Carretta, 2014; Ree & Earles, 1992), it was expected that general aptitude and language aptitude would predict not only graduation outcomes, but also growth. Previous studies (Bush, 1997; Surface et

al, 2005; Silva & White, 1993; Watson et al., 2012; Wagener, 2016) have repeatedly shown that general aptitude and language aptitude incrementally predict graduation outcomes at the Defense Language Institute Foreign Language Center (DLIFLC), as well as at the United States Army John F. Kennedy Special Warfare Center and School. It was expected that this research would confirm earlier findings regarding graduation outcomes.

Hypothesis 4: There is a significant negative relationship between language difficulty and language proficiency growth over time.

Studies have demonstrated that there is a significant, negative relationship between language difficulty and language outcomes (Surface et al., 2005; Masters, 2018; Wagener, 2016). What is interesting about these findings, especially in terms of graduation from basic training, is that the schools do take steps to reduce the impact of language difficulty, such as minimum DLAB scores and adjustments in course length. However, despite these steps, language difficulty remains a significant moderator of language outcomes, and in this research, it was expected that it would continue to constrain graduation outcomes, as well as growth.

Hypothesis 5: Language difficulty will interact with general aptitude and language aptitude.

Few studies have considered interactions of predictor variables, and even fewer have used an analytic approach that would even allow for the interaction of *cross-level* predictors. The current study was designed to examine the impact of predictor variables on growth, including cross-level effects, such as the interaction of language difficulty and general or language aptitude. Wagener's (2016) analysis

suggested that there would be significant interactions. It was expected that language difficulty would interact with both general aptitude and language aptitude.

This research filled a gap in the scientific literature by addressing the longitudinal nature of proficiency, as well as the nesting created by each language. Over the decades, there has been a wealth of studies to support selection criteria in the context of graduation from language training. However, selection should be concerned not only with graduation from the basic course at DLIFLC, but also with continued growth over time. Language professionals in the military must continue to improve their language skills to be successful on the job. The nation's security depends upon it.

Chapter 4: The data

The dataset was delivered by DLI to the researcher in November 2021 in occasion-level form (i.e., long format) and divided into separate files for listening and reading test scores. Personnel were identified by DLI with random identification numbers in accordance with the approved IRB protocols from the University of Maryland as well as the DLI.

The test scores were from the individuals who graduated from a DLI basic language course between May 2010 and September 2019 and who took a DLPT between October 2010 and September 2020. The original dataset comprised over 65,00 test scores for listening and over 45,000 in reading. The availability of more listening tests in general, Arabic dialects in particular, drove the increased number of listening tests. Many individuals tested in more than one language, likely due to testing in any prior language or testing in a related language, such as Persian-Farsi and Persian-Dari. The multilingual nature of individuals in this dataset meant that scores were nested under an individual who was, in turn, nested under more than one higher-level group.

There are several ways to handle such cross-classification: delete the cases completely, attribute the data to one group only, or apply a cross-classified model (Anderson, 2012). Given the size of the dataset and the nature of the research questions, the decision was made to delete the language test scores that were not related to the language of the first basic course in the dataset. This allowed for each individual to be nested under only one language. Only those who successfully graduated were retained for this study, as those who did not graduate did not continue

to work in the field and test in a language. These two steps, to focus on the basic course and to only include graduates, reduced the original dataset provided by DLI by 24%.

These decisions had the additional benefit of focusing the analysis on the question of how the aptitude measures used for selection into basic training were related to language proficiency growth in the language of initial study. In the reading dataset, an exception was made for individuals who attended a basic course in an Arabic dialect such as Arabic-Gulf, as such individuals all tested in Modern Standard Arabic (MSA) in reading because the dialects are not assessed separately in reading. Listening was more complicated, as many individuals tested in the dialect they studied, and/or in MSA listening, and/or in a related dialect. To reduce complexity in design, the study retained only the listening test scores for the version of Arabic that was the focus of the basic course. For example, if a student was enrolled in Arabic-Gulf, but tested in both Arabic-Gulf and Arabic-Iraqi, only the Arabic-Gulf scores were included.

Very little data cleaning was necessary after the data were reduced to focus on basic course graduates. DLAB scores from four individuals were suspect and the data on the individuals with these scores were removed from both listening and reading files. To reduce imbalance in the data, languages with fewer than seven test scores were removed (German and Portuguese in the reading data and German in the listening). Test occasions greater than six were also removed. The final dataset included 34,742 test scores in listening for 9,552 individuals and 34,935 test scores in

reading for 9,564 individuals, with the bulk of the drop attributed to focusing on the language of the basic course.

There were three categories of information in the dataset: scores related to aptitude (ASVAB and DLAB), information collected at DLI via surveys, and data related to language test scores. The individuals in the listening and reading datasets almost completely overlap; therefore, only the reading data is reported below for the aptitude and survey variables. The ASVAB and DLAB data were complete; there were no missing data. Individual response data on these batteries was not provided, however, so factor analyses could not be conducted to explore this data more fully. The survey responses collected at DLI were missing data, and how this was handled is described below. The language test data was also complete although not every individual tested the same number of times.

The following sections describe the aptitude, classroom and testing data. IBM SPSS Statistics (Version 28) was used to organize the data, conduct multiple imputation and report foundational statistics (descriptives, correlations).

Aptitude-related data

The first category of information in the dataset was related to cognitive aptitude, both general aptitude (as measured by ASVAB) and language aptitude (as measured by DLAB). Scores were provided for nine ASVAB subtests and two composite scores (ASVAB-AFQT and ASVAB-VE), and the four DLAB parts with an overall DLAB score. The original dataset included DLAB scores from two sources: a self-reported score collected on the student survey and the official score. The official DLAB scores were retained and used for analysis based on the

recommendation of DLIFLC (Dr. Seumas Rogan, personal communication, December 14, 2021). Score reliability was not provided with the data for either ASVAB or DLAB, but a previously published study that included ASVAB as a predictor of language outcomes reported empirical reliability scores ranging from .70 to .92 for the subtests (Bunting et al., 2011, p. 6). The same study calculated and reported DLAB alpha reliabilities for the three forms used in their study ranging from .79 to .88 (pp. 15-26).

ASVAB Subtests

Nine ASVAB subtests were provided to the researcher and descriptive statistics from the cleaned dataset are presented below. ASVAB scores are standardized scores, with a score of 50 representing the mean for the general population of examinees, with plus or minus 10 points per standard deviation (Culver, n.d.). The mean scores in the current dataset were all approximately one standard deviation higher than for the general ASVAB test-taking population, with the exception of the AS subtest, which was on the mean (Auto and Shop).

Table 2

ASVAB subtests

ASVAB subtest	N	Mean	Median	Mode	Std Deviation
AO (Assembling Objects)	9564	62.07	63.00	68	5.523
AR (Arithmetic Reasoning)	9564	62.61	63.00	61	5.190
AS (Auto/Shop)	9564	50.64	50.00	51	7.210
EI (Electronic Information)	9564	58.63	58.00	57	7.637
GS (General Science)	9564	61.85	62.00	62	6.217
MC (Mechanical Comprehension)	9564	60.81	61.00	59	6.912
MK (Math Knowledge)	9564	63.02	63.00	64	4.935
PC (Paragraph Comprehension)	9564	61.87	62.00	62	4.683
WK (Word Knowledge)	9564	62.02	62.00	62	6.020

Two composite scores were also provided (see Table 3): the Verbal Expression score and the Armed Forces Qualifying Test score. As explained above, a Verbal Score (VE) score is formed from an optimally weighted composite of unrounded WK and PC standard scores. This Verbal Score, in turn, is double weighted in the computation of AFQT scores: AFQT = AR + MK + 2(VE) (Ostrow, 2002).

Table 3

ASVAB composite scores

Statistics

		VE (Verbal score)	AFQT/ASVAB
N	Valid	9564	9564
	Missing	0	0
Mean		62.50	90.92
Median		63.00	93.00
Mode		63	99
Std. Deviation		5.040	7.987

DLAB Subtests

Scores on the four DLAB subtests, as well as an overall DLAB score are shown below (Table 4). The minimum overall score in this sample was 71, while the mean was 117. Only 0.2% of the individuals were admitted into DLI with scores below the lowest minimum DLAB score in this time period (85), while 70% of the incoming students had DLAB scores higher than the required minimum score for even the most difficult languages (110).

Table 4

DLAB Statistics

Statistics

		Win-DLAB Part 1 NC	Win-DLAB Part 2 NC	Win-DLAB Part 3 NC	Win-DLAB Part 4 NC	Win-DLAB Std Score
N	Valid	9564	9564	9564	9564	9564
	Missing	0	0	0	0	0
Mean		3.12	13.02	46.69	19.66	117.42
Std. Deviation		1.557	2.941	5.698	4.089	11.619
Minimum		0	3	23	0	71
Maximum		7	18	64	30	159

ASVAB-DLAB Correlations

As anticipated from findings in earlier studies, there were a number of significant correlations among pairs of aptitude variables (see the full correlation table in Appendix A. The strongest correlations were within the ASVAB itself, both within domains as well as across domains: within the Science/Technical domain (AS and MC, AS and EI, EI and MC, GS and MC); within the Math domain (AR and MK), and across the Verbal and Science/Technical domains (WK and GS). DLAB Part I, the Biographical section, was the only subtest that did not show a significant correlation with all the other aptitude-related subtests. Table 5 below shows the pairs of variables with significant correlations greater than 0.5.

Table 5

ASVAB correlations greater than 0.5

Aptitude pair	Correlation
AS-MC	*.567***
AS-EI	*.594***
EI-GS	*.555***
EI-MC	*.588***
MC-GS	*.522***
AR-MK	*.557***
GS-WK	*.517***

** $p < .01$

Survey data

The second category of variables in the dataset was related to enrollment in the basic course. The dataset provided by DLI included four variables from a survey administered to students beginning their basic class. Approximately 20% of the values for these measures were missing data. While additional variables could have been included in this study, the following four variables were retained based upon previous research attesting to their influence on language proficiency.

The first variable in this category was a measure of language choice, which was used as a proxy for motivation (Table 6) to learn the language. This measure had limited utility, as individuals responded to the question at only one point in time, prior to the start of their basic course, but no better measure was available. Students at DLI considered themselves motivated to study their language, even if it was not their first choice. Only 3% of the respondents reported being unmotivated, while a quarter of the respondents reported that they were training in their first choice language and

were therefore presumably highly motivated. This variable was coded such that a higher number indicated more motivation (1=Not my choice and 5=Based on my first choice.)

Table 6

Descriptives for language learning motivation prior to training

	Frequency	Percent
Not my choice. I would prefer to do something else rather than study a foreign language	202	2.1
Not my choice. I am not motivated to study the assigned language.	78	0.8
Not my choice, but I am still motivated to study the assigned language.	2953	30.9
Based on my second or third choice.	1854	19.4
Based on my first choice	2559	26.8
System missing	1906	20.0

The second variable retained from the survey was a measure of self-assessed prior language proficiency (Table 7). Thirty percent of the respondents reported no prior proficiency, while 14% reported having proficiency in a foreign language that they self-rated as “good” or “excellent”. This variable was coded as 0=None through 4=Excellent. Other variables related to prior proficiency, such as which prior language and where that language was acquired, were omitted from the current research.

Table 7

Descriptives for prior language proficiency

	Frequency	Percent
None	2276	23.8
Poor	2045	21.4
Fair	1798	18.8
Good	917	9.6
Excellent	411	4.3
System missing	2105	22.0

The third variable was an indication as to whether English was the individual's first language (Table 8). Almost all the sample of those who responded to the survey (97.2%) reported being first language English speakers. In the analyses, this variable coded as 0=English not as a first language and 1=English as first language.

Table 8

Descriptives for first language English

	Frequency	Percent
English as first language	7424	77.6
Other first language	196	2.0
System missing	1944	20.3

The final variable retained from the survey for the study was the level of prior education (Table 9). Two-thirds of the individuals who responded to the survey had at least some college experience. This variable was coded from 1=No high school through 9=Doctorate.

Table 9

Descriptives for level of education

	Frequency	Percent
No high school	13	0.1
High school/GED	2530	26.5
One year of college	1152	12.0
Two years of college	1302	13.6
Three years of college	484	5.1
Four years of college	294	3.1
Bachelor's Degree	1702	17.8
Master's Degree	126	1.3
Doctorate	16	0.2
System missing	1945	20.3

Language testing data

The third and final category of information in the dataset was records of language testing. For each test score, the dataset included the test information (language, series, form, range, test date) and test score information (ILR Level and raw score). After data cleaning, the reading dataset contained test scores across 17

languages, with the majority of tests in Arabic (AD), Chinese (CM), Persian (PF), Spanish (QB) and Russian (RU) as shown in Table 10. The listening dataset contained 21 languages due to the availability of dialect tests such as Arabic-Gulf (see

Table 11).

Table 10

Distribution of reading test scores

Language	N	Percent
Arabic	9048	26%
Chinese	5163	15%
Dari	180	1%
French	601	2%
Hebrew	770	2%
Indonesian	124	0%
Japanese	35	0%
Korean	3603	10%
Persian	5304	15%
Punjabi	12	0%
Pushtu	2731	8%
Russian	3400	10%
Serbian-Croatian	58	0%
Spanish	3023	9%
Tagalog	110	0%
Turkish	42	0%
Urdu	728	2%

Table 11

Distribution of listening test scores

Language	N	Percent
Arabic-MSA	4287	12%
Arabic-Egyptian	400	1%
Arabic-Gulf	1438	4%
Arabic-Sudanese	65	0%

Arabic-Syrian	2660	8%
Chinese	5162	15%
French	603	2%
Hebrew	770	2%
Indonesian	135	0%
Japanese	34	0%
Korean	3612	10%
Persian-Dari	180	1%
Persian-Farsi	5307	15%
Punjabi	12	0%
Pushtu	2729	8%
Russian	3404	10%
Serbian-Croatian	58	0%
Spanish	3015	9%
Tagalog	111	0%
Turkish	42	0%
Urdu	730	2%

Languages were unevenly distributed in the listening and reading datasets as they reflect the different requirements for military language positions during this time period. Over half of the languages tested were in the most difficult language category (DLI Category IV), while 30% were in Category III and only 10% in Category I. Real-world needs were reflected in the distribution of languages studied at DLIFLC. Potential impacts of this distribution will be addressed below in the discussion of language distance measures and in the results chapter.

Test scores were provided as ILR levels and raw scores. The raw score was not used in the analysis because of the inability to equate raw scores across languages (Dr. Seumas Rogan, personal communication, March 17, 2022). ILR test scores in the original dataset ranged from ILR Level 0 to ILR Level 4, but the only scores above ILR Level 3 are from the Spanish DLPT, which had been converted to a computer-adaptive format, allowing for a range of scores from ILR 0 to ILR 4. Because scores

above ILR Level 3 were limited to Spanish, scores in this research were capped at ILR Level 3. This step, in addition the steps outlined avove to retain only DLI graduates, truncated the ILR levels even further.

Table 12

Listening language test scores

Decimal Version of ILR Score

	N	%
.00	33	0.1%
.60	181	0.5%
1.00	817	2.4%
1.60	4065	11.7%
2.00	11908	34.3%
2.60	10310	29.7%
3.00	7440	21.4%

Table 13

Reading language test scores

Decimal Version of ILR Score

	N	%
.00	43	0.1%
.60	167	0.5%
1.00	333	1.0%
1.60	2456	7.0%
2.00	10350	29.6%
2.60	12852	36.8%
3.00	8744	25.0%

In the foreign language arena, many researchers have treated proficiency scores as if they were interval-level outcomes, assuming a normal distribution on a continuous scale. Winke et al. (2022) discretized an underlying continuous self-assessed proficiency measure based on the ACTFL (2012) scale (collapsed into five levels) by using an ordinal measure (Winke et al., 2022). An ordinal measure, rather

than a continuous measure, was likely used for the self-assessment because of the small number of levels (5). The research reviewed above clearly favored the treatment of the ILR scale as a continuous measure, as suggested by the ILR Skill Level Descriptions themselves (ILR, 1985), while the research questions in this study may best be understood using ordinal models, given the interest in stages of growth (O'Connell, 2000), as well as in the unequal distribution of test scores. As Tigchelaar (2019) observed, the type of scale and statistical analysis used do matter. This motivated additional analyses to model the data with both continuous and ordinal approaches. The ILR test score variable was converted to a decimal version, with "plus" levels coded as .6 (ILR, 1985), creating a continuous variable with scores from 0.0 through 3.0.

Unlike earlier studies analyzing DLPT data (Bloomfield et al., 2014; Mackey, B., 2014), the present dataset did not include any languages that changed test generations from DLPT IV to DLPT 5, so coding for a generational version change was not needed.

A test sequence counter of the number of times an examinee took a test in a language, was created for the longitudinal analysis. A counter "Time" was created from the test sequence variable, starting with the first test as zero (0) to facilitate interpretation of the results. Rescaling time in this way meant the intercept could be interpreted as the predicted outcome at Time = 0, the first test occasion, which in this case was the ILR score upon graduation from DLIFLC. Language professionals are required to test each year while they are in the service and individuals had a varying number of test occasions in the dataset, depending on when they graduated and their

length of service. The distribution of test scores is reported below in Table 14 for listening and Table 15 for reading. There are two explanations for the differing number of tests in the dataset: first, some individuals in this dataset completed their basic training at the end of the data collection time period and therefore have had less time to continue testing; second, some individuals would have completed their tour of service during this period and therefore stopped testing. Those that have a higher number of test occasions (> 5) started their basic training at the beginning of the data collection period, reenlisted, and therefore continued to test annually. The majority of individuals in the dataset tested at least three times, and a more thorough exploration of time and its relationship with individual differences is considered in these studies is in the Results chapter below.

Table 14

Listening test occasions

Time	N	Percent (attrition)
0	9552	100
1	8256	86
2	6549	69
3	5038	53
4	3385	35
5	1974	21

Table 15

Reading test occasions

Time	N	Percent (attrition)
0	9564	100
1	8286	87
2	6581	69
3	5077	53
4	3420	36
5	2017	21

Languages were each coded with six measures of language difficulty to be used in the analyses: the FSI categories as modified in Cysouw (2013) (variable name *FSILangCat* with levels 1 to 7), a typological similarity number (Cysouw, 2013) (variable reversed and labeled *TypeRev*), the Gateway (Gnanadeskikan & van Rossum, 2016) index number (variable reversed and labeled *GateRev*), whether or not the language uses a Latin Script (variable labeled *NotLatin*, where 0=Latin), whether or not the language is in the Indo-European language family (variable labeled *NotIndo*, where 0=Indo), and the original DLI language difficulty categories (variable labeled *LangCat*). To facilitate interpretation, the Gateway and Typology measures were reversed so that all of the language measures reflected distance from English, or learning difficulty for an L1 speaker of English, given the overwhelming percentage of L1 speakers of English in the dataset. Thus, for all six language measures, a higher number indicated a more difficult language for a typical DLI student.

A note of caution is warranted here regarding the reliability of these measures, especially regarding the categorical and continuous measures, given the imprecision of the decisions behind the categories. By their very nature, language categories are

reductionist, and when many of the languages are not well documented, the categories or similarity indices were based upon expert judgement. Where categorization was not specified for a language in the dataset, a measure based on a similar language was used. For example, Arabic (MSA) was not available in the Gateway database, so it was coded in the same manner as Arabic-Egyptian, even though other Arabic languages were coded with a slightly lower number (43 versus 38). Arabic (MSA) and Egyptian are both well documented, somewhat standardized languages and therefore the expectation was that they would be slightly "easier" for L1 English speakers than the lesser-taught Arabic dialects. The complete list of languages in the dataset and their associated language distance measures is displayed below in Table 16.

Table 16

Original Language distance measures

Language	FSI LangCat	GateRev	TypeRev	NotLatin	NotIndo	DLI LangCat
Arabic Egyptian	6	43	0.50	0	0	4
Arabic Gulf	6	38	0.50	0	0	4
Arabic MSA	6	43	0.50	0	0	4
Arabic Sudanese	6	38	0.50	0	0	4
Arabic Syrian	6	38	0.50	0	0	4
Chinese	6	38	0.46	0	0	4
Dari	4	48	0.43	0	1	3
French	1	86	0.66	1	1	1
Hebrew	4	48	0.56	0	0	3
Indonesian	3	38	0.48	1	0	3
Japanese	7	33	0.39	0	0	4
Korean	6	33	0.45	0	0	4
Persian-Farsi	4	48	0.43	0	1	3
Punjabi	4	43	0.40	0	1	3
Pushtu-Afghan	4	48	0.43	0	1	4
Russian	4	67	0.65	0	1	3
Serbo-Croatian	4	67	0.61	0	1	3
Spanish	1	71	0.62	1	1	1
Tagalog	4	57	0.39	1	0	3
Turkish	4	48	0.44	1	0	3
Urdu	4	38	0.50	0	1	3

The associations across the language distance measures in the dataset were calculated using the appropriate statistics (Pearson's *r*, chi-square tests of independence, and point-biserial correlation), and significant relations were found across the measures. The two continuous measures, *GateRev* and *TypeRev*, were strongly related (*Pearson's r* = 0.795, $p < .01$), which is not surprising given their origins. Given their different relationships with the other distance measures, both variables were retained. The seven-level FSI measure was also strongly related to

GateRev (0.853, $p < .01$), but less so to *TypeRev* (0.507, $p < .01$). There was a significant association between *NotLatin* and *NotIndo* ($\chi^2(1) > = 3828.87$, $p < .001$). Point biserial correlations were conducted to test the association of the categorical and continuous variables, and all were significant (*TypeRev-NotLatin* 0.508, $p < .001$; *TypeRev-NotIndo* 0.275, $p < .001$; *GateRev-NotIndo* 0.710, $p < .001$; *GateRev-NotLatin* 0.673, $p < .001$). Given the strong, significant associations across the measures and their unknown reliability, rather than enter the measures together, the effects of each language distance measure were tested independently in the analyses below.

Chapter 5: Methodology

Previous research on language growth has shown different patterns of individual and group change, but with few exceptions, language proficiency research has not taken a longitudinal approach, nor has it employed multilevel modeling. Multilevel models are particularly useful to answer questions such as how growth is shaped and how variables might impact that growth. These models can leverage data from incomplete observations (Finch et al., 2019; Hox, 2000; Raudenbush & Bryk, 2002), so sample sizes are not reduced as dramatically as with other methods such as linear regression. (See Lett (1994) and Wagener (2016) for examples of how attrition constricts sample size). In a multilevel analysis, measurements can also be spaced irregularly. This negates the need to artificially reformat the data to accommodate evenly spaced test occasions as was seen in other studies using other analytic approaches (Bloomfield et al., 2014; Mackey, B., 2014). Group sizes are not required to be equal (Linck and Cunnings, 2012; Steele, 2007), which is another advantage. "Multi-level modeling provides an attractive alternative analysis (to ANOVA), because it allows statistical evaluation of incomplete data, without any additional complication" (Snijders & Bosker, 1999, p. 170).

While multilevel modeling can handle unbalanced measurement data, missing time-invariant predictor variables, such as the survey variables collected by DLI, cannot be. Data in a multilevel model is in "long" form, with a row holding information about the individual and *one* test occasion. Even if an individual missed test occasions, the cases for the other occasions are included in the model. This is the advantage of a multilevel model mentioned above. However, each case must be

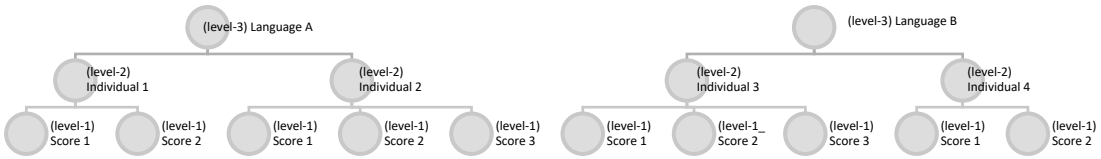
complete; it must include all of the predictors and outcomes in the model (Hoffman, 2015, p. 283). In the present dataset, there is a significant amount of missing survey data, and rather than delete these cases, other options were explored. Given that DLI considered this data to be missing at random (Dr. Seumas Rogan, personal communication, September 27, 2021) and the four variables were not the key variables in the study, the decision was made to continue with multiple imputation rather than deleting the cases. Imputation was conducted using IBM SPSS Statistics (Version 28) to create ten multiply imputed datasets. The default method was used (fully conditional specification using an iterative Markov-chain Monte Carlo) method. Methodologists currently regard multiple imputation as a state-of-the-art technique because it improves accuracy and statistical power relative to other techniques of handling missing data (Buuren, 2018).

The original proposal for this research envisioned a multilevel longitudinal framework that was characterized by a data structure in the following form: repeated measures (level-1) nested within an individual (level-2) nested within a higher-level structure (level-3). Figure 3 below illustrates how the data was to be modeled: language test scores at time *t* could serve as the repeated measure at level-1; these scores would be nested by individual *j* at level-2; and the grouping category, *k*, which would be the foreign language tested, would be at level-3. The advantage of a three-level model is that it would separate the variance associated with the language (level-3) from the variance associated with the individual (level-2). This would have made it possible to describe not only which individual difference measures were associated

with growth, but also to test the unique contribution of language distance as a predictor of growth, above and beyond those individual differences.

Figure 2

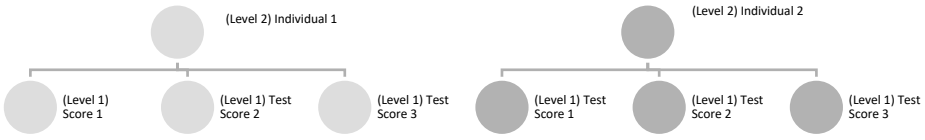
A sample three-level repeated measures model



A series of models for listening and reading were originally run as planned with a three-level model, as described above, with repeated measures at level-1, individuals at level-2 and language at level-3. While even the more complex models converged, the data failed to support estimation of fixed effects with robust standard errors, likely due to the small number of languages at level-3 (17 in reading and 21 in listening). Though the datasets have a relatively large number of languages as compared to other studies in the SLA field, they proved not to be adequate for three-level modeling. After consultation with an expert in the field, it was determined that the best approach would be to use a two-level model instead (Professor L. Stapleton, personal communication, 17 November 2022). It was recommended that rather than model language at level-3, that language distance measures be considered as a variable at level-2, since each person in the dataset was associated with only one language. The structure of the data did meet minimum requirements for using a two-level model and modeling therefore proceeded in a forward fashion with two-levels, as displayed below in Figure 4.

Figure 4

A sample two-level repeated measures model



The warrant for using the multilevel model as opposed to a repeated measures ANOVA was determined by calculating the Intraclass Correlation Coefficient (ICC). The ICC was calculated by determining the proportion of variance attributed to level-2. An ICC of zero would indicate no mean ILR level score variation across individuals, and multilevel modeling would not be needed. If the ICC is greater than zero, variation across individuals is present in the data, providing the warrant for the multilevel approach. This is one of the advantages of multilevel modeling, that it can take into account that residuals from the same person are more likely to be related than residuals from different persons (Hoffman & Stawski, 2009). Multilevel models separate the variation, in this case into two levels, within-person (level-1) and between-person (level-2).

For the level-1 models, 95% random effects confidence intervals were calculated by using the formula *Random Effect 95% CI = fixed effect ± (1.96* √random variance).* If the interval does not overlap zero, it also means that the effect is significantly different from zero (Hoffman, 2015, p. 166). Assumptions of the final models at level-1 and level-2 were checked, as misspecification at one level can bias the estimates in another (Bryk & Raudenbush, 2010; Hoffman, 2015; Peugh & Heck,

2017). Alternate covariance structures for the final models were considered and the residuals were checked for normality.

Given the number of comparisons made in the later models, corrections were made using the Benjamini-Hochberg (B-H) approach to limit the familywise Type I error rate (Thissen et al., 2002). Table 23

HLM listening FSI0[aptitude] interactions*Table 31, Table 38, andTable 45 include a column indicating the estimates in which the observed *p* value is less than the B-H critical value, giving more confidence in the relationships. According to Thissen et al. (2002) the B-H approach "…is a more powerful generalization to the context of multiplicity of the conventional test of significance" (p. 78).

For each model, a standardized effect size was calculated. This measure, a proportional reduction in variance (PRV), was calculated from a comparison of the variance in the model with fewer parameters (Var_{fewer}) relative to a model with more parameters (Var_{more}). This measure is also called Pseudo-R^2. The PRV was determined by level (i.e., level-1 or level-2) and its purpose was to describe the proportion of the outcome variance accounted for by the fixed effects of predictors (Hoffman, 2015). PRV values are always interpreted in the context of the earlier models.

As explained earlier, the outcome measure, the language proficiency test scores, were reported as ILR levels. In Study 1 and Study 2, the outcome variable was treated as a continuous measure (*ILRNum*). However, the ILR scale fails to meet the requirement for a linear scale with levels that are evenly spaced, as is seen in the

ILR's depiction of the scale as a pyramid, with each level on the scale as deeper and broader than the previous level (see website https://www.govtilr.org/Skills/IRL%20Scale%20History.htm accessed 10 October 2022). Treating the ILR scale as an ordinal measure would account for this non-linear description of the scale, as well as taking into account the ceiling and floor effects in skewed variables (Hedeker, 2015). There is justification in the literature for treating the ILR scale in this study as a continuous variable, given that it has at least seven levels (Bauer & Sterba, 2011), but there is limited research on the ILR scale. One contribution of this research to the literature, therefore, was to run the models with the ILR treated as a continuous measure and as an ordinal measure.

To accommodate the ordinal outcome variable, the analysis was conducted using hierarchical generalized linear modeling in Study 3 (Listening) and Study 4 (Reading). A comparison of results between Study 1 and Study 3 and Study 2 and Study 4 follows in the discussion chapter. For these analyses, the ILR scale was first converted from the decimal version to a range of 1 (ILR Level 0) through 7 (ILR Level 3) but given the low number of scores (0.2 percent) at ILR Level 0, the two lowest categories were combined into one level, resulting in a score range from 1 through 6.

In a hierarchical linear model (HLM) of longitudinal data, a linear model with a random intercept and random slope would be represented at level-1 by the equation:

$$ILRNUM_{ti} = \pi_{0i} + \pi_{1i}*(Time_{ti}) + e_{ti} \qquad \text{(A)}$$

where *ILRNUM* is the outcome at time *t* for individual *i*, and π are the level-1 regression coefficients, and *e* is the level-1 error term. The level-2 model is represented by the following equations for the intercept and slope:

$$\pi_{0i} = \beta_{00} + r_{0i} \quad \text{(B)}$$

$$\pi_{1i} = \beta_{10} + r_{1i} \quad \text{(C)}$$

where the β's are the level-2 regression coefficients and *r*'s are the level-2 random effects. For the sake of space, equations in this research are presented in their mixed form as below in Equation D.

$$\text{ILRNUM}_{ti} = \beta_{00} + \beta_{10}*\text{Time}_{ti} + r_{0i} + r_{1i} + e_{ti} \quad \text{(D)}$$

In a hierarchical generalized linear model (HGLM), the level-1 model consists of a sampling model, a link function and a structural model (HLM8 manual (2021), p. 108). For ordinal models, the probability that an individual falls into category *m* is cumulative, with *m-1* dummy variables. The model is based on a cumulative logit, in which "success" represents the probability of being at or below a threshold, a somewhat counterintuitive position when interpreting the estimates. In terms of language proficiency, our interest is in individuals who score beyond that threshold, so in HGLM analysis, negative logits are more desirable. In the models that follow, a negative estimate was interpreted to mean that those with higher scores on a particular variable had a positive effect on the graduation outcome or on growth.

The equations in the ordinal model become complicated quickly, but the logic is similar to that of the linear models described above. A random slope model with linear time is written out as follows.

Level-1 Model

$\text{Prob}[R_{ti} <= 1|\pi_i] = \phi^*_{1ti} = \phi_{1ti}$
$\text{Prob}[R_{ti} <= 2|\pi_i] = \phi^*_{2ti} = \phi_{1ti} + \phi_{2ti}$
$\text{Prob}[R_{ti} <= 3|\pi_i] = \phi^*_{3ti} = \phi_{1ti} + \phi_{2ti} + \phi_{3ti}$
$\text{Prob}[R_{ti} <= 4|\pi_i] = \phi^*_{4ti} = \phi_{1ti} + \phi_{2ti} + \phi_{3ti} + \phi_{4ti}$
$\text{Prob}[R_{ti} <= 5|\pi_i] = \phi^*_{5ti} = \phi_{1ti} + \phi_{2ti} + \phi_{3ti} + \phi_{4ti} + \phi_{5ti}$
$\text{Prob}[R_{ti} <= 6|\pi_i] = 1.0$

$\phi_{1ti} = \text{Prob}[ILRORD(1) = 1|\pi_i]$
$\phi_{2ti} = \text{Prob}[ILRORD(2) = 1|\pi_i]$
$\phi_{3ti} = \text{Prob}[ILRORD(3) = 1|\pi_i]$
$\phi_{4ti} = \text{Prob}[ILRORD(4) = 1|\pi_i]$
$\phi_{5ti} = \text{Prob}[ILRORD(5) = 1|\pi_i]$

$\log[\phi^*_{1ti}/(1 - \phi^*_{1ti})] = \pi_{0i} + \pi_{1i}*(TIME_{ti})$
$\log[\phi^*_{2ti}/(1 - \phi^*_{2ti})] = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + \delta_2$
$\log[\phi^*_{3ti}/(1 - \phi^*_{3ti})] = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + \delta_3$
$\log[\phi^*_{4ti}/(1 - \phi^*_{4ti})] = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + \delta_4$
$\log[\phi^*_{5ti}/(1 - \phi^*_{5ti})] = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + \delta_5$

Level-2 Model

$\pi_{0i} = \beta_{00} + r_{0i}$
$\pi_{1i} = \beta_{10} + r_{1i}$

$\delta_2 \quad \delta_3 \quad \delta_4 \quad \delta_5$

The first section of the level-1 model presents the probabilities for the seven categories. The second section describes the cumulative predicted probabilities. Since there are six ILR levels, 1 through 6, and the data is cumulatively partitioned into five "splits" as follows: $Y \leq 1,\ Y \leq 2,\ Y \leq 3,\ Y \leq 4,\ Y \leq 5$ (where Y represents each of the ILR level possibilities). Since all responses are included in $Y \leq 6,$ it is redundant (O'Connell et al., 2022, p. 220). The third section associates the five cumulative logits with their probabilities. The level-2 model is as described above in the HLM approach. HLM8 (Raudenbush & Congdon, 2021) parameters are estimated using "penalized quasi-likelihood" or PQL and there is an assumption of proportionality, which means that the effect of any predictor variable remains constant regardless of the response level. This assumption will be addressed again in the limitation chapter.

Methods for estimating non-proportional random-effects models are not available in HLM8 (Raudenbush & Congdon, 2021). Deviances are not produced with the PQL method, so model fit was determined by considering the reduction in unexplained variance. Additionally, the software also reports only unit-specific results where "the goal is to provide inferences for covariates that can change within cluster[s] (i.e., individuals)" (Bell et al., 2022, p. 226).

In all four studies, the first series of models in the research focused on level-1 and allowed for the exploration of the extent to which individuals vary, and the shape of variation around their mean scores at the time of graduation (the intercept) and their growth (the slope). After each step, only significant relationships were retained ($p < .05$). The analysis approach followed Hoffman (2015), Hox (2000), and Peugh & Heck (2017), building forward, rather than backwards. For both listening and reading datasets, the first model tested was the intercept-only model, also called a null model, where the variance within- and between-individual and language was partitioned. Time was then explored, and model fit was determined based upon a comparison of deviance statistics (where available), any reduction in unexplained variance, and theoretical concerns. A best-fitting level-1 model to capture the longitudinal nature of the data was established to answer research questions 1 and 2.

RQ 1: To what extent is there variance in language proficiency growth over time, across individuals and languages?

RQ 2: What is the shape of language proficiency growth and how does it vary by language?

Once this level-1 model was determined, predictor variables were entered into the model. Explanatory variables were entered at level-2 (best general aptitude, language aptitude, survey responses, language distance) to see if they explained any variation across individuals or languages. The aptitude variables (ASVAB and DLAB subtests) were standardized to reduce the impact of skewness. All predictor variables were entered into the models centered on the grand mean. This allowed the estimates to be interpreted in the context of mean scores. These analyses were conducted to answer research questions 3 and 4 regarding language aptitude and language difficulty.

RQ 3: To what extent does language aptitude predict language proficiency growth outcomes across languages, beyond what is predicted by general ability?

RQ 4: To what extent does language difficulty categorization impact language proficiency growth across languages?

A final stage of model building then explored whether aptitude and language distance interacted with growth. Interaction terms were first created in IBM SPSS Statistics (Version 28) since HLM8 (Raudenbush & Congdon, 2021) does not allow for the direct testing of same-level interactions. To ease interpretation in an already complex model, a dichotomous FSI measure FSI0 was created from the seven-level categorical FSI variable in the models above. FSI0 was dummy coded such that the most difficult languages, those in difficulty categories 6 and 7, were coded as 1, and all other languages were coded as 0. This new dichotomous variable was then crossed with the ASVAB and DLAB standardized subtest scores as a product term and added to the model as predictors of the intercept and slopes. The ASVAB variable names as

interactions with FSI0 were represented in the software with their subtest digraph, an underscore and FSI0; for example, *AR_FSI0*. The DLAB subtest variable names as interactions were in the software as *D1_ FSI0, D2_ FSI0, D3_ FSI0 or D4_FSI0.* All of the main effects for the ASVAB and DLAB subtests were included along with their interaction terms and the variables were centered on the grand mean. This model was used to answer the last research question.

RQ 5: To what extent does aptitude interact with language difficulty categorization?

Chapter 6: Results

To examine the influence of general aptitude, language aptitude, motivation, prior proficiency, English as a first language and level of education on proficiency outcomes at DLI graduation, growth, and subsequent growth, four modeling studies were conducted. These studies investigated listening and reading data and considered test scores separately as continuous or ordinal variables. The five research questions guided the modeling.

Study 1 Hierarchical Linear Modeling (Listening)

The first step in the analysis was to build a two-level model (Equation 1) in which the ILR outcome ($ILRNUM_{ti}$) was modeled at level-1 as a function of each individual's mean proficiency level (π_{0i}) at time *t* plus a residual that reflected the differences between each individual's observed and predicted proficiency level at a specific time (e_{ti}). At level-2, each individual's mean proficiency level was modeled as a function of the grand-mean level for all individuals (β_{00}) plus a term that reflected deviations in an individual's proficiency mean around the grand mean (r_{01}). There are different notation systems in the literature, and this research follows Raudenbush & Bryk's (2002) approach for modeling longitudinal data.

$$ILRNUM_{ti} = \beta_{00} + r_{0i} + e_{ti} \quad (1)$$

There was a maximum of 34,742 level-1 units (test scores) and 9,552 level-2 units (individuals). The average random level-1 coefficient for the intercept had a reliability estimate of over .70, indicating a moderate level of reliability (Nezlek, 2017). The mean estimate for the intercept (reported throughout the chapter with robust standard errors) for listening was 2.28 ($p < .001$), i.e., an ILR score between

ILR Levels 2 and 2+. Random effects estimates showed the ILR score variance estimates at level-1, within-person, as $\sigma^2 = 0.132$ and τ_{00} (level-2, between-person) as 0.168 with both estimates significant ($p < .001$). This means that there was significant variation around the grand mean (within-individual) as well as significant differences between each individual's observed and predicted proficiency level over time. Following Peugh (2010), the level-2 variance estimate was converted to a standard deviation to facilitate interpretation (i.e. $\sqrt{0.168} = 0.41$). Assuming normal distribution of the residuals, 95% of the individuals had mean ILR scores between 1.87 and 2.69 (i.e., 2.28 ± 1.96 [0.41]), i.e., an ILR score between 1+ and 2+. The intraclass correlation (ICC), which describes the proportion of variance that lies between individuals, was determined by dividing the variance at level-2 by the model's total variance. The level-2 ICC showed that 56% of the language proficiency variation occurred across individuals, which was interpreted as a warrant to continue with a multilevel model.

The next step in modeling was to examine variation and the rate and shape of growth. Several options were explored to describe the longitudinal nature of the data. First, time was considered as a linear function with a random intercept as in Equation 2 below. Recall that the *TIME* variable was coded such that the first test occasion in the dataset was "0" to ease interpretation of the intercept.

$$ILRNUM_{ti} = \beta_{00} + \beta_{10} * TIME_{ti} + r_{0i} + e_{ti} \qquad (2)$$

The fixed effect estimate for the intercept dropped slightly from the baseline model to 2.23 and the estimate for the slope was 0.03 ($p < .001$), which was interpreted to mean that the ILR level rose only slightly between test occasions. The models cannot

be compared with a chi-squared test of differences because there are not enough degrees of freedom, but the average deviance statistic across the ten datasets for Equation 2 was only slightly lower, and the random effect estimates were unchanged, which indicated that linear time did not explain any additional variance in the fixed effects.

Because there were up to six test occasions in the data, the quadratic (Equation 3) and cubic (Equation 4) forms of time were modeled:

$$ILRNUM_{ti} = \beta_{00} + \beta_{10}*TIME_{ti} + \beta_{20}*TIME2_{ti} + r_{0i} + e_{ti} \quad (3)$$

$$ILRNUM_{ti} = \beta_{00} + \beta_{10}*TIME_{ti} + \beta_{20}*TIME2_{ti} + \beta_{30}*TIME3_{ti} + r_{0i} + e_{ti} \quad (4)$$

The output from these models is shown below in Table 17. These models from Equations 3 and 4 did not explain any additional variance, but they did reflect the longitudinal nature of the data in different ways. In the quadratic model of change, the quadratic effect indicated a significant, but quite small acceleration in the linear rate of change on average, over time (0.01 of an ILR level). The non-significance of *Time* may be attributed to the average of rise and fall of scores over time resulting in flat growth. In the cubic polynomial model, where the effects of *Time, Time*^{2} *and Time*^{3} were modeled, the results meant that a negative cubic effect dampened the positive quadratic effect on a downward mean trajectory. Negative effects indicate that growth was constrained.

The next set of models allowed the time functions to vary (i.e., with random slopes). These models failed to converge, or had reliability estimates for the random coefficients that were below 0.10. According to the HLM manual (2021), when the reliability is very low, defined as below 0.10, it "often indicate(s) that a random

coefficient might be considered fixed in subsequent analyses" (p. 80). The random intercept model estimates with fixed forms of time (linear, quadratic and cubic) are reported below in Table 17. The deviance statistics fell slightly as the quadratic and cubic terms were added, but the addition of these terms did not affect the level-1 variance. This means that polynomial time did not explain any variance within-individuals.

Table 17

HLM listening level-1 random intercept time model parameters

	Null Model	RI Time	RI Quad	RI Cube
		Fixed Effects		
Effect				
Intercept (β_{00})	2.28*** (0.00)	2.23*** (0.01)	2.24*** (0.01)	2.25*** (0.01)
Time (β_{10})		0.03*** (0.00)	0.00 (0.00)	-0.07*** (0.01)
Time Quad (β_{20})			0.01*** (0.00)	0.05*** (0.01)
Time Cube (β_{30})				-0.01*** (0.01)
		Random Effects		
Variance components				
level-1, *e*	0.13 (0.36)	0.13 (0.36)	0.13 (0.36)	0.13 (0.36)
level-2 intercept, r_0		0.16 (0.41)	0.17 (0.41)	0.17 (0.41)
		Goodness of fit		
Deviance	44023.60359	43469.7031	43406.8211	43324.7234
Number of Parameters	2	2	2	2

Note: Standard error in parentheses for estimated coefficients.

*** $p < .001$

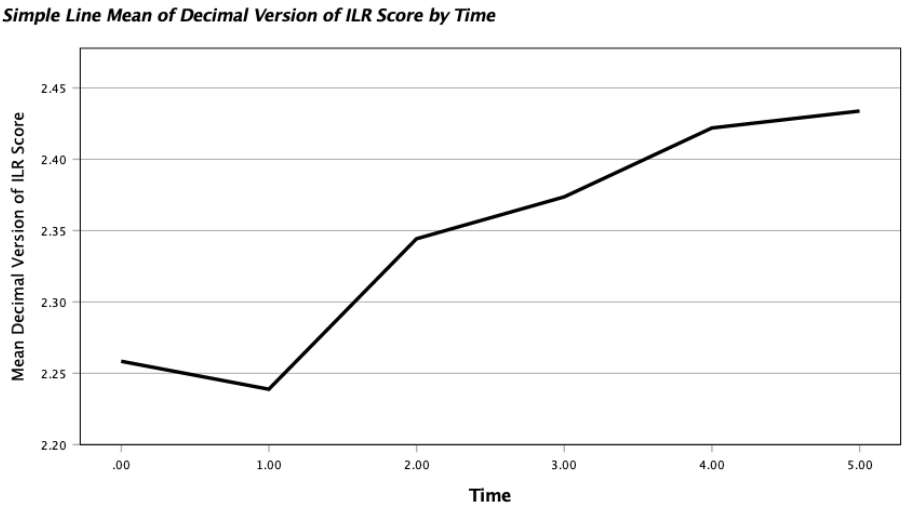
As mentioned above, multilevel models allow for unbalanced data. This is because change is represented with two levels: each individual's

"…[level-1] growth trajectory that depends upon unique set of parameters. These individual growth parameters become the outcome variables in a level-2 model, where they may depend upon some person-level characteristics… This treatment of multiple observations as nested allows the investigator to proceed without difficulty when the number and spacing of time points vary across cases" (Raudenbush & Bryk, 2002, p. 161).

Earlier studies in language assessment literature examined time as a polynomial function, testing quadratic and cubic forms of time and found that ILR levels dropped after graduation, and then recovered back to their graduation level, or increased beyond that level (Bloomfield et al, 2012; Mackey, 2014). In the present research, a similar pattern of drop and recover was seen in the mean ILR score across all languages, although both the dip and the subsequent growth was very shallow for the listening scores in this dataset. Figure 4 below displays the mean ILR score across test occasions, represented by the *Time* variable on the y-axis. The x-axis range, 2.20 to 2.45, is within the range of ILR Level 2, whose decimal form ranges from 2.0 to 2.6.

Figure 3

Mean ILR Scores over time (listening)



To better model this observed dip in growth post-DLI, rather than considering only polynomial functions, time was modeled piecewise with two slopes (Hoffman, 2015), one to represent initial growth *(Slope12)* and a second (*Slope26*) to represent subsequent growth. The new slope variables recoded the number of test occasions similarly to the *Time* variable, in that the first test occasion, graduation from DLI, was set as zero (0) to ease interpretation of the intercept. The first slope, in effect, described what happens to language growth in the first year after leaving DLI; the second slope described subsequent growth, or what happened after the first full year on the job. The coding scheme for all of the various representations of time up to this point in the study is shown below.

Table 18

Coding schemes for time

Test Occasion	Time	$Time^2$	$Time^3$	Slope12	Slope26
1	0	0	0	0	0
2	1	1	1	1	0
3	2	4	6	1	1
4	3	9	27	1	2
5	4	16	64	1	3
6	5	25	125	1	4

The resulting mixed model equation for listening with piecewise fixed slopes and a random intercept using *Slope12* and *Slope26* is shown in Equation 5. This model, the random intercept piecewise slopes model (RISlopes), answered the question "is there linear change during each time period on average?" (Hoffman, 2015, p. 232)

$$ILRNUM_{ti} = \beta_{00} + \beta_{10}*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + r_{0i} + e_{ti} . \quad (5)$$

The output for this model is shown in Table 19. The fixed effect for the intercept dropped very slightly as compared to the Null model to 2.24, meaning that the mean scores at the time of graduation from DLI were ILR Level 2, but of more interest were the estimates for the two piecewise slopes. The mean of the change of ILR level between the first two test occasions (*Slope12*) was negative (-0.04), while the mean change after the second test occasion (*Slope26)* was positive (0.05) and both estimates were significant ($p < .001$). This was interpreted to mean that there was significant linear change during each time period, on average. The initial growth was

a decrease in ILR level (-0.04 of an ILR level between the first two tests), while subsequent growth was an increase (0.05 of an ILR level per test occasion). The significant random intercept indicated that there was also variation in the average mean score at the time of graduation. While significant, however, the estimates themselves were quite small, and showed that after leaving DLI, on average, listening scores dropped very slightly and then subsequently rose very slightly. There was very little meaningful growth in ILR level, on average, in this sample.

Random slopes were tested next to allow the growth rate to vary across individuals, first allowing each slope to vary randomly (first *Slope12*, then *Slope26*) and then testing a model with both slopes varying. For the two models with a random second slope, the random level-1 coefficient reliability estimate for subsequent growth (*Slope26*) was only 0.04, suggesting that it be fixed. Therefore, the best fitting piecewise model appeared to be a model in which only initial growth (*Slope12)* was allowed to vary randomly to answer the question whether there were individual differences in linear change between the first two test occasions. The reliability estimate for the random level-1 coefficient was marginal at 0.20, but sufficient for modeling. This model's full equation is below in Equation 6, and the output reported in Table 19.

$$ILRNUM_{ti} = \beta_{00} + \beta_{10}*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (6)$$

The average ILR score across individuals was 2.26, i.e., within the range of ILR Level 2, and mean initial growth in ILR level (*Slope12*) was negative (-0.03 of an ILR level), while mean subsequent growth (*Slope26*) was positive (0.05 of an ILR

level). Significant variation in the first slope was present, indicating that individuals' initial growth varied.

Table 19

HLM listening level-1 model parameters with piecewise slopes

	RI Slopes	Slope12R
	Fixed Effects	
Intercept (β_{00})	2.26*** (0.01)	2.26*** (0.01)
Slope12 (β_{10})	-0.04*** (0.01)	-0.03*** (0.01)
Slope26 (β_{20})	0.05*** (0.00)	0.05*** (0.00)
	Random Effects	
level-1, *e*	0.13 (0.36)	0.12 (0.35)
level-2 intercept, r_0	0.17*** (0.41)	0.19*** (0.44)
level-2 *Slope12,* r_1		0.04*** (0.21)
	Goodness of fit	
Deviance	43276.91	43124.07
Number of Parameters	2	4
$\Delta\chi 2$		152.84***

Note: Standard error in parentheses for estimated coefficients

*** p < .001

Given that the piecewise model estimates appeared to capture the observed growth pattern in the dataset and the piecewise model with a varying first slope was also an improvement over the model with fixed slopes (χ^2= 152.84,, *df* =2, *p* <.001) as well as the null model (χ^2= 899.53, *df* =2, *p* <.001), the decision was made to continue with the piecewise slope model, allowing for random initial growth (*Slope12*) and fixed subsequent growth (*Slope26*) to describe the shape of time. Further evidence supporting the decision to proceed with the piecewise slopes was

found in the random effects, as the addition of a random growth slope (*Slope12*) explained 11% of the variance at level-1 as compared to the null model. Additional expressions of growth were considered and are addressed below.

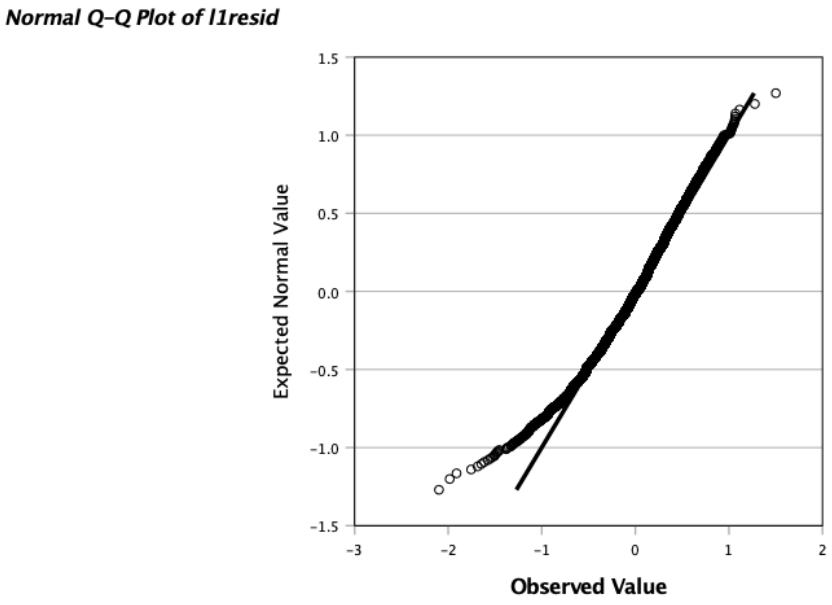
Returning now to the first two research questions, the results shown above in Table 19 confirmed the hypotheses that there was significant variability across individuals at the time of graduation (the intercept), as well in initial growth (*Slope12*). In the Slope12R model, time was described with two slopes, one to describe a slope between the first and second test occasions that was allowed to vary, and a second slope from the second to sixth test occasions that was fixed. The results indicated that individuals graduated with a score in listening, on average, in the ILR Level 2 range (2.25), which dropped, on average, 0.03 of an ILR level between the first two test occasions and rose 0.05 of an ILR level thereafter, on average. While statistically significant, likely due to the large sample size, these slope estimates were quite small and would not have a meaningful impact on the ILR level itself, given that ILR Level 2 ranges from 2.0 to 2.6. For the Slope12R listening model, a 95% random effects confidence interval for the intercept and slope indicated that 95% of the sample was expected to have individual ILR levels at graduation ranging from 1.41 to 3.11, while the 95% CI for initial growth (*Slope12*) was -0.11 to 0.05. While there was a significant rate of linear change (a decrease) between the first two tests on average, the random variation around the first slope indicated that some individuals dropped in level and others rose, as evidenced by the overlap with 0 in the random effect CI (Hoffman, 2015, p. 166). There was significant variance in ILR levels in initial growth, but not in subsequent growth. Because three-level models were not

used due to the small number of languages in the data, these models were not able to assess the proportion of variance that could be attributed to language, as opposed to individual.

Assumptions of the level-1 model were checked. There was deviation present at the tail of the residuals plot (see Figure 5), but in general the other assumptions for the level-1 model were considered to be met.

Figure 4

Normal Q-Q plot of level-1 residuals



The next set of models was designed to investigate the third research question, to what extent does language aptitude predict language proficiency growth outcomes across languages, beyond what is predicted by general aptitude? Model building continued in a stepwise fashion to consider the effects of the level-2 covariates on proficiency scores and whether the scores varied by language. The subtest scores for ASVAB were standardized and centered on the grand mean and added

simultaneously to the Slope12R model above in Equation 6 to predict the intercept and both slopes. In the dataset, the ASVAB subtests were standardized, labeled ZA_XX and entered in alphabetical order. The mixed model equation for this model is shown below in Equation 7, and the results are in Table 20. At this point it is perhaps useful to explain the notation system in the output, as the model greatly increased in its complexity. The β's represented the coefficients at the person-level; the first subscript represented a sequential count of predictors at level-1, while the second subscript represented a sequential count of predictors at level-2. As variables were added or subtracted from the model, the β's subscripts were updated, as seen by comparing Equations 7 and 8 below. Therefore, in tables combining output from several models, the subscripts were omitted from the variables.

In the model adding the ASVAB subtests (Equation 7), the intercept was represented by β_{00}, while the first slope was represented by β_{10}*$SLOPE12_{ti}$ and the second slope, β_{20}*$SLOPE26_{ti}$. Predictors of the intercept ranged from β_{01} to $\beta_{09,}$ while predictors of the first slope ranged from β_{11} to β_{19} and predictors of the second slope from $\beta_{21\ to}\beta_{29.}$ In this model, the random effects included the variance of the intercept, r_{0i}, the first slope, r_{1i}*$SLOPE12_{ti}$ and the residual, or level-1 variance, e_{ti}. Because the variables were time-invariant, they do not have random effects, as by definition they do not vary within persons.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AO_i + \beta_{02}*ZA\_AR_i + \beta_{03}*ZA\_AS_i + \beta_{04}*ZA\_EI_i + \beta_{05}*ZA\_GS_i + \beta_{06}*ZA\_MC_i + \beta_{07}*ZA\_MK_i + \beta_{08}*ZA\_PC_i + \beta_{09}*ZA\_WK_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_AO_i*SLOPE12_{ti} + \beta_{12}*ZA\_AR_i*SLOPE12_{ti} + \beta_{13}*ZA\_AS_i*SLOPE12_{ti} + \beta_{14}*ZA\_EI_i*SLOPE12_{ti} + \beta_{15}*ZA\_GS_i*SLOPE12_{ti} + \beta_{16}*ZA\_MC_i*SLOPE12_{ti} + \beta_{17}*ZA\_MK_i*SLOPE12_{ti} + \beta_{18}*ZA\_PC_i*SLOPE12_{ti} + \beta_{19}*ZA\_WK_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_AO_i*SLOPE26_{ti} + \beta_{22}*ZA\_AR_i*SLOPE26_{ti} + \beta_{23}*ZA\_AS_i*SLOPE26_{ti} + \beta_{24}*ZA\_EI_i*SLOPE26_{ti} + \beta_{25}*ZA\_GS_i*SLOPE26_{ti} + \beta_{26}*ZA\_MC_i*SLOPE26_{ti} + \beta_{27}*ZA\_MK_i*SLOPE26_{ti} + \beta_{28}*ZA\_PC_i*SLOPE26_{ti} + \beta_{29}*ZA\_WK_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (7)$$

Six of the nine subtests were significantly related to the intercept (*ASVAB-AO, -EI, -MC* were not), while only one subtest, *ASVAB-MC*, was significantly (and negatively) related to the first slope (-0.02, $p = .006$), and none to the second slope. Specifically, the mean ILR score for those with mean ASVAB scores ($\beta_{00}= 2.26$, $p < .001$) dropped between the first two test occasions ($\beta_{10}= -0.03$, $p < .001$) and subsequently rose ($\beta_{20}= 0.05$, $p < .001$) for those with average ASVAB scores. Of the significant subtests, the -*AS* subtest was a negative predictor of graduation (intercept), and the remainder were positive, meaning that those higher ASVAB scores other than -*AS* were associated with higher ILR levels, all else being equal. Variance component estimates once again confirmed significant variation in observed versus predicted ILR scores within-individual (level-1 variance $\sigma^2 = 0.12$, $p < .001$) and significant variation in ILR scores at the time of graduation from DLI (level-2 variance $r_0 = 0.18$, $p < .001$). There was also significant variation ($r_1 = 0.04$) in initial growth (i.e.,

between the first two test occasions). The level-1 variance was unchanged, as expected, because only level-2 predictors were added. The addition of the ASVAB variables explained 7% of the variance in graduation (intercept) and 2% of the variance in initial growth (first slope) as compared to the unconditional piecewise model.

The final ASVAB model (ASVABsig) with only the significant subtests is shown in Table 20 and is represented by the equation below.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MK_i + \beta_{05}*ZA\_PC_i + \beta_{06}*ZA\_WK_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_MC_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (8)$$

Given earlier studies which confirmed the additional predictive validity of adding a language aptitude measure (Silva & White, 1976; Mackey, B., 2014; Wagener, 2016), and the current study's interest in how aptitude predicts growth, the next stage was to add the four DLAB subtests, resulting in Equation 9 below. The subtests were standardized and once again centered on the grand mean. The four DLAB subtests representing DLAB Part I, Part II, Part III and Part IV are labeled *ZD_1, ZD_2, ZD_3 and ZD_4* in the software.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MK_i + \beta_{05}*ZA\_PC_i + \beta_{06}*ZA\_WK_i + \beta_{07}*ZD\_1_i + \beta_{08}*ZD\_2_i + \beta_{09}*ZD\_3_i + \beta_{010}*ZD\_4_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_MC_i*SLOPE12_{ti} + \beta_{12}*ZD\_1_i*SLOPE12_{ti} + \beta_{13}*ZD\_2_i*SLOPE12_{ti} + \beta_{14}*ZD\_3_i*SLOPE12_{ti} + \beta_{15}*ZD\_4_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZD\_1_i*SLOPE26_{ti} + \beta_{22}*ZD\_2_i*SLOPE26_{ti} + \beta_{23}*ZD\_3_i*SLOPE26_{ti} + \beta_{24}*ZD\_4_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (9)$$

The output from this model (DLAB) is presented in Table 20 below. The fixed estimates for the intercept and slopes were unchanged, and of the language aptitude subtests predicting the intercept, all but *DLABPt4* was a significant predictor of graduation (intercept), all else being equal. The relationship between DLAB and each slope was different: for initial growth (*Slope12*), *DLABPt1, DLABPt3* and -*Pt4* were significant, with the direction of the estimate for -*Pt1* positive and for -*Pt3* and -*Pt4* negative. For subsequent growth (*Slope26*), only the effect of *DLABPt3* was significant, though the estimate itself was quite small (-0.007). A model with the significant DLAB variables in addition to the ASVAB subtests was taken forward in the model building process. The reliability estimate for the random level-1 *Slope12* coefficient remained at 0.20. The addition of the DLAB models explained 2% of the variation in graduation and 1% of the variation in initial growth as compared to the model with only the ASVAB variables.

The output for an intervening model with only the significant aptitude models (DLABsig) is displayed in the fourth column of Table 20 and its equation in mixed form is below in Equation 10.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MK_i + \beta_{05}*ZA\_PC_i + \beta_{06}*ZA\_WK_i + \beta_{07}*ZD\_1_i + \beta_{08}*ZD\_2_i + \beta_{09}*ZD\_3_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_MC_i*SLOPE12_{ti} + \beta_{12}*ZD\_1_i*SLOPE12_{ti} + \beta_{13}*ZD\_3_i*SLOPE12_{ti} + \beta_{14}*ZD\_4_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZD\_3_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (10)$$

Table 20

HLM listening with aptitude subtests model parameters

	ASVAB	ASVABsig	DLAB	DLABsig
		Fixed effect		
For INTRCPT1, π0				
INTRCPT2, β_{00}	2.26*** (0.01)	2.26*** (0.01)	2.26*** (0.01)	2.26*** (0.01)
ZA_AO	-0.01 (0.01)			
ZA_AR	0.03*** (0.01)	0.02*** (0.01)	0.02** (0.01)	0.02** (0.01)
ZA_AS	-0.02** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)
ZA_EI	-0.01 (0.01)			
ZA_GS	0.03*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZA_MC	-0.01 (0.01)			
ZA_MK	0.04*** (0.01)	0.04*** (0.01)	0.03*** (0.01)	0.03*** (0.01)
ZA_PC	0.04*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)
ZA_WK	0.05*** (0.01)	0.05*** (0.01)	0.04*** (0.01)	0.04*** (0.01)
ZD_1			0.02*** (0.01)	0.02*** (0.01)
ZD_2			0.03*** (0.01)	0.02*** (0.00)
ZD_3			0.04*** (0.01)	0.04*** (0.01)
ZD_4			0.00 (0.01)	
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)
ZA_AO	-0.01 (0.01)			
ZA_AR	0.00 (0.01)			

	ASVAB	ASVABsig	DLAB	DLABsig
ZA_AS	0.00 (0.01)			
ZA_EI	0.01 (0.01)			
ZA_GS	-0.01 (0.01)			
ZA_MC	-0.02** (0.01)	-0.03*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)
ZA_MK	0.00 (0.01)			
ZA_PC	-0.01 (0.01)			
ZA_WK	0.01 (0.01)			
ZD_1			0.02*** (0.01)	0.02*** (0.01)
ZD_2			-0.01 (0.01)	
ZD_3			-0.01* (0.01)	-0.02** (0.01)
ZD_4			-0.02*** (0.01)	-0.01*** (0.00)
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)
ZA_AO	0.00 (0.00)			
ZA_AR	0.00 (0.00)			
ZA_AS	0.00 (0.00)			
ZA_EI	0.00 (0.00)			
ZA_GS	0.00 (0.00)			
ZA_MC	0.00 (0.00)			
ZA_MK	0.00 (0.00)			
ZA_PC	0.00 (0.00)			
ZA_WK	0.00 (0.00)			
ZD_1			0.00 (0.00)	

	ASVAB	ASVABsig	DLAB	DLABsig
ZD_2			0.00 (0.00)	
ZD_3			-0.01*** (0.00)	-0.01*** (0.00)
ZD_4			0.00 (0.00)	

Final estimation of variance components

	Random Effect			
level-1, *e*	0.12 (0.35)	0.12 (0.35)	0.12 (0.35)	0.12 (0.35)
INTRCPT1, r_0	0.18*** (0.42)	0.18*** (0.42)	0.17*** (0.42)	0.17*** (0.42)
SLOPE12 slope, r_1	0.04*** (0.21)	0.04*** (0.21)	0.04*** (0.21)	0.04*** (0.21)
Deviance	42854.2894	42701.1228	42635.3716	42589.3424
Parameters	4	4	4	4

Note: Standard error in parentheses for estimated coefficients. All level-2 variables were centered on the grand mean.

*** $p < .001$ ** $p < .01$ * $p < .05$

Though not the focus of the current study, other individual differences variables have been shown to moderate the prediction of aptitude and therefore were included in the model at this stage to account for their possible influence before considering language distance and interactions. As explained above, these variables were collected once at the start of the basic language course and therefore were time-invariant: the level of education prior to the basic course, the motivation to train in the language of the basic course, the level of prior language proficiency, and whether the individual's first language was English. They were entered into the model centered on the grand-mean to facilitate the interpretation of the estimates, as shown in Equation 11 below. In the software, these variables were represented by *EDUC, MOT, PRIORPRO* and *ENGY. EDUC, MOT* and *PRIORPRO* were coded such that a higher number indicated more of the measure; *ENGY* was coded such that 0 = not English as a first language and 1 = English as a First Language.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MK_i + \beta_{05}*ZA\_PC_i + \beta_{06}*ZA\_WK_i + \beta_{07}*ZD\_1_i + \beta_{08}*ZD\_2_i + \beta_{09}*ZD\_3_i + \beta_{010}*EDUC_i + \beta_{011}*MOT_i + \beta_{012}*PRIORPRO_i + \beta_{013}*ENGY_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_MC_i*SLOPE12_{ti} + \beta_{12}*ZD\_1_i*SLOPE12_{ti} + \beta_{13}*ZD\_3_i*SLOPE12_{ti} + \beta_{14}*ZD\_4_i*SLOPE12_{ti} + \beta_{15}*EDUC_i*SLOPE12_{ti} + \beta_{16}*MOT_i*SLOPE12_{ti} + \beta_{17}*PRIORPRO_i*SLOPE12_{ti} + \beta_{18}*ENGY_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZD\_3_i*SLOPE26_{ti} + \beta_{22}*EDUC_i*SLOPE26_{ti} + \beta_{23}*MOT_i*SLOPE26_{ti} + \beta_{24}*PRIORPRO_i*SLOPE26_{ti} + \beta_{25}*ENGY_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (11)$$

The output from this model is shown in Table 21. The fixed effect estimates for the intercept and two slopes were unchanged. Because the model with the survey variables was the first model in which imputed values were modeled, ten separate models were averaged and reported here; individual model statistics, including a deviance statistic for each model, were also produced. The deviance statistics varied by model, as expected given the multiply imputed data, and ranged from 42,546.42 to 42,580.53 with 4 *df* and an average of 42,565.82.

The addition of the survey variables to the DLAB model had a negligible effect on the intercept (PRV less than 1%) and none on the slope. The education and English variables had negative estimated coefficients as predictors of the intercept, all else being equal, meaning that the more education, the lower the ILR level upon graduation, and those with English as a first language were less likely to have higher ILR levels upon graduation. The motivation measure was not a significant predictor of the intercept. These three findings were counterintuitive. There was a positive effect of prior proficiency, as all else being equal, individuals with stronger prior proficiency were more likely to have a higher ILR level at graduation, a finding which would be predicted by the literature. The only survey variable to have a significant effect on growth was the education variable (*EDUC*); but its significance was marginal in the full survey model, and it dropped out of significance in subsequent modeling. The findings for the survey variables, especially in their direction, were somewhat surprising. The reliability of these measures and the multiple imputation may have impacted the results. It may also have been attributed

to the unique context of the DLIFLC learning environment, in which more traditional educational habits acquired in higher education are not as applicable.

The final survey model (Equation 12 below) with only the significant variables was taken forward for further modeling. The output from this model is shown below.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MK_i + \beta_{05}*ZA\_PC_i + \beta_{06}*ZA\_WK_i + \beta_{07}*ZD\_1_i + \beta_{08}*ZD\_2_i + \beta_{09}*ZD\_3_i + \beta_{010}*EDUC_i + \beta_{011}*PRIORPRO_i + \beta_{012}*ENGY_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_MC_i*SLOPE12_{ti} + \beta_{12}*ZD\_1_i*SLOPE12_{ti} + \beta_{13}*ZD\_3_i*SLOPE12_{ti} + \beta_{14}*ZD\_4_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZD\_3_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (12)$$

Table 21

HLM listening models with survey variables

	Survey	SurveySig
	Fixed Effect	
For INTRCPT1, π_0		
INTRCPT2, β_{00}	2.26*** (0.01)	2.26*** (0.01)
ZA_AR, β_{01}	0.02*** (0.01)	0.02*** (0.01)
ZA_AS, β_{02}	-0.01* (0.01)	-0.01* (0.01)
ZA_GS, β_{03}	0.02*** (0.01)	0.02*** (0.01)
ZA_MK, β_{04}	0.03*** (0.01)	0.03*** (0.01)
ZA_PC, β_{05}	0.03*** (0.01)	0.03*** (0.01)
ZA_WK, β_{06}	0.05*** (0.01)	0.05*** (0.01)
ZD_1, β_{07}	0.03*** (0.01)	0.03*** (0.01)
ZD_2, β_{08}	0.02*** (0.00)	0.02*** (0.00)
ZD_3, β_{09}	0.04*** (0.01)	0.04*** (0.01)
EDUC, β_{010}	-0.02*** (0.00)	-0.02*** (0.00)
MOT, β_{011}	0.00 (0.01)	
PRIORPRO, β_{012}	0.02*** (0.01)	0.02*** (0.00)
ENGY, β_{013}	-0.15*** (0.04)	-0.14*** (0.03)
For SLOPE12 slope, π_1		

	Survey	SurveySig
INTRCPT2, β_{10}	-0.03*** (0.01)	-0.03*** (0.01)
ZA_MC, β_{11}	-0.02*** (0. 00)	-0.02*** (0.00)
ZD_1, β_{12}	0.01** (0.01)	0.02*** (0.01)
ZD_3, β_{13}	-0.02*** (0.01)	-0.02*** (0.01)
ZD_4, β_{14}	-0.01*** (0.00)	-0.01*** (0.00)
EDUC, β_{15}	0.01*† (0.00)	
MOT, β_{16}	0.00 (0.01)	
PRIORPRO, β_{17}	0.01 (0.01)	
ENGY, β_{18}	0.02 (0.04)	
For SLOPE26 slope, π_2		
INTRCPT2, β_{20}	0.05*** (0.00)	0.05*** (0.00)
ZD_3, β_{21}	-0.01*** (0.00)	-0.01*** (0.00)
EDUC, β_{22}	0.00 (0.00)	
MOT, β_{23}	0.00 (0.00)	
PRIORPRO, β_{24}	0.00 (0.00)	
ENGY, β_{25}	-0.01 (0.01)	
Final estimation of variance components		
	Random Effect	
level-1, *e*	0.12 (0.35)	0.12 (0.35)
INTRCPT1, r_0	0.17*** (0.41)	0.17*** (0.41)
SLOPE12 slope, r_1	0.04*** (0.21)	0.04*** (0.21)

Note: Standard error in parentheses for estimated coefficients. All level-2 variables were centered on the grand mean.

* $p < .05$ ** $p < .01$ *** $p < .001$ † dropped out of significance in subsequent models

The next step was to add the language distance measures (centered on the grand mean) to a model with only the significant survey variables. The measures were added individually in six separate models: *FSI, GateRev, TypeRev*, *NotLatin, NotIndo, and DLI.* To facilitate interpretation, the distance measures were all coded such that the higher values represented greater distance. For example, the *NotLatin* measure was coded such that 0=Latin script and 1=non-Latin script, and the Gateway

similarity index was reversed such that languages with higher numbers were further distant from English. The mixed model equation with the DLI language difficulty category measure is shown below in Equation 13 as an example, and in the interest of space, models with the other language distance measures are not provided here. These variables were represented in the software by *LANGCAT, FSILANGC, GATEREV, TYPEREV, NOTLATIN* and *NOTINDO.*

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MK_i + \beta_{05}*ZA\_PC_i + \beta_{06}*ZA\_WK_i + \beta_{07}*ZD\_1_i + \beta_{08}*ZD\_2_i + \beta_{09}*ZD\_3_i + \beta_{010}*EDUC_i + \beta_{011}*PRIORPRO_i + \beta_{012}*ENGY_i + \beta_{013}*LANGCAT_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_MC_i*SLOPE12_{ti} + \beta_{12}*ZD\_1_i*SLOPE12_{ti} + \beta_{13}*ZD\_3_i*SLOPE12_{ti} + \beta_{14}*ZD\_4_i*SLOPE12_{ti} + \beta_{15}*LANGCAT_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZD\_3_i*SLOPE26_{ti} + \beta_{22}*LANGCAT_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti}$$ (13)

The output for the average of the ten multiple imputations can be found in Table 22 for each language distance measure*, FSI, GateRev, TypeRev*, *NotLatin, NotIndo, and DLI.* The fixed effect and variance component estimates for the intercept and slope(s) were little changed across all the models. The *FSI, TypeRev, NotIndo* and *DLI* measures were significantly related to the intercept (proficiency at graduation); all but the *NotIndo* measure were significantly related to the first slope (initial growth), and the *GateRev, TypeRev, NotLatin* and *DLI* measures all had a significant effect on the second slope (subsequent growth). The *DLABPt3* and *DLABPt4* measures relationships to the first slope were differentially impacted by the addition of the distance measures. In all six models except for the TypeRev model, these two DLAB subtests either fell completely out of significance or were of

marginal significance in the models with the language distance measures. The addition of a language difference variable explained up to 1% of the variance in intercept and up to 4% of the first slope as compared to the survey model.

Table 22

HLM listening language distance model parameters

	FSI	GateRev	TypeRev	NotLatin	NotIndo	DLI
Fixed Effect						
For INTRCPT1, π_0						
INTRCPT2, β_{00}	2.26*** (0.01)	2.26*** (0.01)	2.26*** (0.01)	2.26*** (0.01)	2.26*** (0.01)	2.26*** (0.01)
ZA_AR, β_{01}	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZA_AS, β_{02}	-0.01* (0.01)	-0.01* (0.01)	-0.01* (0.01)	-0.01* (0.01)	-0.01* (0.01)	-0.01* (0.01)
ZA_GS, β_{03}	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZA_MK, β_{04}	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)
ZA_PC, β_{05}	0.04*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.04*** (0.01)	0.03*** (0.01)
ZA_WK, β_{06}	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)
ZD_1, β_{07}	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)
ZD_2, β_{08}	0.03*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.03*** (0.00)	0.02*** (0.00)
ZD_3, β_{09}	0.05*** (0.01)	0.04*** (0.01)	0.03*** (0.01)	0.04*** (0.01)	0.05*** (0.01)	0.04*** (0.01)
EDUC, β_{010}	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02* (0.00)
PRIORPRO, β_{011}	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.03*** (0.00)	0.02*** (0.00)
ENGY, β_{012}	-0.15*** (0.03)	-0.14 *** (0.03)	-0.14*** (0.03)	-0.14*** (0.03)	-0.14*** (0.03)	-0.14*** (0.03)
[lang dist], β_{013}	-0.03*** (0.00)	0.00 (0.00)	0.00*** (0.00)	0.01 (0.02)	-0.11*** (0.01)	-0.01* (0.01)
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)
ZA_MC, β_{11}	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)

	FSI	GateRev	TypeRev	NotLatin	NotIndo	DLI
ZD_1, β_{12}	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZD_3, β_{13}	-0.01 (0.01)	-0.01* (0.01)	-0.01* (0.01)	-0.01 (0.01)	-0.02** (0.01)	-0.01 (0.01)
ZD_4, β_{14}	0.00 (0.00)	-0.01* (0.00)	-0.01** (0.00)	-0.01* (0.00)	-0.01 (0.00)	-0.01 (0.00)
[lang dist], β_{15}	-0.01*** (0.00)	0.00*** (0.00)	0.00*** (0.00)	-0.10*** (0.02)	-0.01 (0.01)	-0.03*** (0.01)
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)
ZD_3, β_{21}	-0.01** (0.00)	-0.01** (0.00)	-0.01*** (0.00)	-0.01*** (0.00)	-0.01*** (0.00)	0.00* (0.00)
[lang dist], β_{22}	0.00 (0.00)	0.00* (0.00)	0.00* (0.00)	-0.01* (0.01)	0.00 (0.00)	-0.01*** (0.00)

Final estimation of variance components

	Random Effect					
level-1, e	0.12 (0.35)	0.12 (0.35)	0.12 (0.35)	0.12 (0.35)	0.12 (0.35)	0.12 (0.35)
INTRCPT1, r_0	0.17*** (0.41)	0.17*** (0.41)	0.17*** (0.41)	0.17 *** (0.41)	0.17*** (0.41)	0.17*** (0.41)
SLOPE12 slope, r_1	0.04*** (0.20)	0.04*** (0.20)	0.04*** (0.20)	0.04*** (0.20)	0.04*** (0.21)	0.04*** (0.20)

Note: Standard error in parentheses for estimated coefficients. All level-2 variables were centered on the grand mean.

* $p < .05$ ** $p < .01$ *** $p < .005$

Interpreting effects in multilevel models is complex, especially given the different scales that each variable used in these models. In the case of *FSI, GateRev* (n/s), *NotIndo*, and *DLI* their relationship was negative for the intercept (proficiency at graduation) and first slope (initial growth), while for TypeRev and NotLatin (n/s), it was positive for the intercept and negative for the first slope. For the second slope (subsequent growth), the direction of the estimates was negative. In the case of the intercept, the model would predict that those in harder languages would have lower scores upon graduation, while for the first slope, there would be a deeper drop between the first two test occasions, and for the second slope a shallower rise thereafter, all other things being equal. In all cases, the sizes of the estimates were such that they would have little practical value to prediction.

The deviance statistic was not reported on the average output summary reported in Table 22, but individual deviance statistics for each of the ten models per distance measure were reviewed; and in combination with the coefficient and variance estimates and the PRV for the intercept and slope, the best fitting model among the models with a distance measure was determined to be the FSI model.

To investigate the final research question which examined the effect of the interaction of aptitude and language difficulty and its effect on growth, the FSI0*[aptitude] interaction terms were added to the model. The mixed model for the aptitude interactions with FSI0 is below in Equation 14 and its output is in Table 23.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AO_i + \beta_{02}*ZA\_AR_i + \beta_{03}*ZA\_AS_i + \beta_{04}*ZA\_EI_i + \beta_{05}*ZA\_GS_i + \beta_{06}*ZA\_MC_i + \beta_{07}*ZA\_MK_i + \beta_{08}*ZA\_PC_i + \beta_{09}*ZA\_WK_i + \beta_{010}*ZD\_1_i + \beta_{011}*ZD\_2_i + \beta_{012}*ZD\_3_i + \beta_{013}*ZD\_4_i + \beta_{014}*EDUC_i + \beta_{015}*PRIORPRO_i +$$

$$\beta_{016}*ENGY_i + \beta_{017}*FSI0_i + \beta_{018}*AO\_FSI0_i + \beta_{019}*AR\_FSI0_i + \beta_{020}*AS\_FSI0_i +$$

$$\beta_{021}*EI\_FSI0_i + \beta_{022}*GS\_FSI0_i + \beta_{023}*MC\_FSI0_i + \beta_{024}*MK\_FSI0_i + \beta_{025}*PC\_FSI0_i +$$

$$\beta_{026}*WK\_FSI0_i + \beta_{027}*D1\_FSI0_i + \beta_{028}*D2\_FSI0_i + \beta_{029}*D3\_FSI0_i + \beta_{030}*D4\_FSI0_i +$$

$$\beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_AO_i*SLOPE12_{ti} + \beta_{12}*ZA\_AR_i*SLOPE12_{ti} +$$

$$\beta_{13}*ZA\_AS_i*SLOPE12_{ti} + \beta_{14}*ZA\_EI_i*SLOPE12_{ti} + \beta_{15}*ZA\_GS_i*SLOPE12_{ti} +$$

$$\beta_{16}*ZA\_MC_i*SLOPE12_{ti} + \beta_{17}*ZA\_MK_i*SLOPE12_{ti} + \beta_{18}*ZA\_PC_i*SLOPE12_{ti} +$$

$$\beta_{19}*ZA\_WK_i*SLOPE12_{ti} + \beta_{110}*ZD\_1_i*SLOPE12_{ti} + \beta_{111}*ZD\_2_i*SLOPE12_{ti} +$$

$$\beta_{112}*ZD\_3_i*SLOPE12_{ti} + \beta_{113}*ZD\_4_i*SLOPE12_{ti} + \beta_{114}*FSI0_i*SLOPE12_{ti} +$$

$$\beta_{115}*AO\_FSI0_i*SLOPE12_{ti} + \beta_{116}*AR\_FSI0_i*SLOPE12_{ti} + \beta_{117}*AS\_FSI0_i*SLOPE12_{ti} +$$

$$\beta_{118}*EI\_FSI0_i*SLOPE12_{ti} + \beta_{119}*GS\_FSI0_i*SLOPE12_{ti} + \beta_{120}*MC\_FSI0_i*SLOPE12_{ti} +$$

$$\beta_{121}*MK\_FSI0_i*SLOPE12_{ti} + \beta_{122}*PC\_FSI0_i*SLOPE12_{ti} +$$

$$\beta_{123}*WK\_FSI0_i*SLOPE12_{ti} + \beta_{124}*D1\_FSI0_i*SLOPE12_{ti} +$$

$$\beta_{125}*D2\_FSI0_i*SLOPE12_{ti} + \beta_{126}*D3\_FSI0_i*SLOPE12_{ti} + \beta_{127}*D4\_FSI0_i*SLOPE12_{ti} +$$

$$\beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_AO_i*SLOPE26_{ti} + \beta_{22}*ZA\_AR_i*SLOPE26_{ti} +$$

$$\beta_{23}*ZA\_AS_i*SLOPE26_{ti} + \beta_{24}*ZA\_EI_i*SLOPE26_{ti} + \beta_{25}*ZA\_GS_i*SLOPE26_{ti} +$$

$$\beta_{26}*ZA\_MC_i*SLOPE26_{ti} + \beta_{27}*ZA\_MK_i*SLOPE26_{ti} + \beta_{28}*ZA\_PC_i*SLOPE26_{ti} +$$

$$\beta_{29}*ZA\_WK_i*SLOPE26_{ti} + \beta_{210}*ZD\_1_i*SLOPE26_{ti} + \beta_{211}*ZD\_2_i*SLOPE26_{ti} +$$

$$\beta_{212}*ZD\_3_i*SLOPE26_{ti} + \beta_{213}*ZD\_4_i*SLOPE26_{ti} + \beta_{214}*FSI0_i*SLOPE26_{ti} +$$

$$\beta_{215}*AO\_FSI0_i*SLOPE26_{ti} + \beta_{216}*AR\_FSI0_i*SLOPE26_{ti} + \beta_{217}*AS\_FSI0_i*SLOPE26_{ti} +$$

$$\beta_{218}*EI\_FSI0_i*SLOPE26_{ti} + \beta_{219}*GS\_FSI0_i*SLOPE26_{ti} + \beta_{220}*MC\_FSI0_i*SLOPE26_{ti} +$$

$$\beta_{221}*MK\_FSI0_i*SLOPE26_{ti} + \beta_{222}*PC\_FSI0_i*SLOPE26_{ti} +$$

$$\beta_{223}*WK\_FSI0_i*SLOPE26_{ti} + \beta_{224}*D1\_FSI0_i*SLOPE26_{ti} +$$

β_{225}*D2_FSI0$_i$*SLOPE26$_{ti}$ + β_{226}*D3_FSI0$_i$*SLOPE26$_{ti}$ + β_{227}*D4_FSI0$_i$*SLOPE26$_{ti}$ +

r_{0i} + r_{1i}*SLOPE12$_{ti}$ + e_{ti}

Table 23

HLM listening FSI0[aptitude] interactions*

	Coefficient	p < B-H
	Fixed Effect	
For INTRCPT1, π_0		
INTRCPT2, β_{00}	2.26*** (0.01)	†
ZA_AO, β_{01}	-0.01 (0.01)	
ZA_AR, β_{02}	0.03* (0.01)	
ZA_AS, β_{03}	0.01 (0.01)	
ZA_EI, β_{04}	-0.01 (0.01)	
ZA_GS, β_{05}	0.02 (0.01)	
ZA_MC, β_{06}	0.00 (0.01)	
ZA_MK, β_{07}	0.02* (0.01)	
ZA_PC, β_{08}	0.04*** (0.01)	†
ZA_WK, β_{09}	0.05*** (0.01)	†
ZD_1, β_{010}	0.04*** (0.01)	†
ZD_2, β_{011}	0.02*** (0.01)	†
ZD_3, β_{012}	0.07*** (0.01)	†
ZD_4, β_{013}	0.03*** (0.01)	†
EDUC, β_{014}	-0.02*** (0.00)	†
PRIORPRO, β_{015}	0.03*** (0.00)	†
ENGY, β_{016}	-0.14*** (0.03)	†
FSI0, β_{017}	-0.16*** (0.01)	†
AO_FSI0, β_{018}	-0.01 (0.01)	
AR_FSI0, β_{019}	0.00 (0.01)	
AS_FSI0, β_{020}	-0.04* (0.01)	
EI_FSI0, β_{021}	0.01 (0.02)	
GS_FSI0, β_{022}	0.01 (0.02)	
MC_FSI0, β_{023}	-0.02 (0.02)	
MK_FSI0, β_{024}	0.01 (0.01)	
PC_FSI0, β_{025}	0.00 (0.01)	
WK_FSI0, β_{026}	-0.01 (0.01)	
D1_FSI0, β_{027}	0.00 (0.01)	

	Coefficient	p < B-H
D2_FSI0, β_{028}	0.03* (0.01)	
D3_FSI0, β_{029}	-0.03** (0.01)	
D4_FSI0, β_{030}	-0.02 (0.01)	
For SLOPE12 slope, π_1		
INTRCPT2, β_{10}	-0.03*** (0.01)	†
ZA_AO, β_{11}	-0.01 (0.01)	
ZA_AR, β_{12}	0.01 (0.01)	
ZA_AS, β_{13}	0.00 (0.01)	
ZA_EI, β_{14}	0.01 (0.01)	
ZA_GS, β_{15}	-0.01 (0.01)	
ZA_MC, β_{16}	-0.03** (0.01)	†
ZA_MK, β_{17}	0.00 (0.01)	
ZA_PC, β_{18}	-0.01 (0.01)	
ZA_WK, β_{19}	0.02* (0.01)	
ZD_1, β_{110}	0.01 (0.01)	
ZD_2, β_{111}	-0.01 (0.01)	
ZD_3, β_{112}	-0.03*** (0.01)	†
ZD_4, β_{113}	-0.03*** (0.01)	†
FSI0, β_{114}	0.00 (0.01)	
AO_FSI0, β_{115}	0.00 (0.01)	
AR_FSI0, β_{116}	-0.02 (0.01)	
AS_FSI0, β_{117}	0.00 (0.01)	
EI_FSI0, β_{118}	-0.01 (0.02)	
GS_FSI0, β_{119}	0.00 (0.01)	
MC_FSI0, β_{120}	0.03 (0.02)	
MK_FSI0, β_{121}	0.00 (0.01)	
PC_FSI0, β_{122}	0.01 (0.01)	
WK_FSI0, β_{123}	-0.03* (0.01)	
D1_FSI0, β_{124}	0.02* (0.01)	
D2_FSI0, β_{125}	0.00 (0.01)	
D3_FSI0, β_{126}	0.03** (0.01)	
D4_FSI0, β_{127}	0.03* (0.01)	
For SLOPE26 slope, π_2		
INTRCPT2, β_{20}	0.05*** (0.00)	†
ZA_AO, β_{21}	0.00 (0.00)	
ZA_AR, β_{22}	0.00 (0.00)	
ZA_AS, β_{23}	0.00 (0.00)	

	Coefficient	p < B-H
ZA_EI, β_{24}	0.00 (0.00)	
ZA_GS, β_{25}	0.00 (0.00)	
ZA_MC, β_{26}	0.01 (0.00)	
ZA_MK, β_{27}	0.00 (0.00)	
ZA_PC, β_{28}	0.00 (0.00)	
ZA_WK, β_{29}	0.00 (0.00)	
ZD_1, β_{210}	0.00 (0.00)	
ZD_2, β_{211}	0.00 (0.00)	
ZD_3, β_{212}	0.00 (0.00)	
ZD_4, β_{213}	0.00 (0.00)	
FSI0, β_{214}	0.00 (0.00)	
AO_FSI0, β_{215}	0.00 (0.00)	
AR_FSI0, β_{216}	0.00 (0.01)	
AS_FSI0, β_{217}	-0.01 (0.01)	
EI_FSI0, β_{218}	0.00 (0.01)	
GS_FSI0, β_{219}	0.00 (0.01)	
MC_FSI0, β_{220}	0.00 (0.01)	
MK_FSI0, β_{221}	0.00 (0.00)	
PC_FSI0, β_{222}	0.00 (0.00)	
WK_FSI0, β_{223}	0.00 (0.01)	
D1_FSI0, β_{224}	0.00 (0.00)	
D2_FSI0, β_{225}	0.00 (0.00)	
D3_FSI0, β_{226}	-0.01 (0.00)	
D4_FSI0, β_{227}	0.00 (0.00)	

Final estimation of variance components

Random Effect	
level-1, e	0.12 (0.35)
INTRCPT1, r_0	0.17*** (0.41)
SLOPE12 slope, r_1	0.04*** (0.20)

Note: Standard error in parentheses for estimated coefficients. All level-2 variables were centered on the grand mean.

** p < .05 ** p < .01 *** p < .005 † p < B-H critical value*

The interaction model was corrected for multiple comparisons and a number of estimates, including several DLAB/FSI interaction terms, failed to meet the criteria although their *p*-value estimates were significant ($p < .01$). As a result of the correction, no interaction terms had a significant effect on graduation (intercept) or growth (either slope). *ASVAB-MC, DLAB Part 3* and *DLAB Part 4* were significant predictors of the first slope, all else being equal. All three estimates were negative, which indicated that individuals with higher scores on these subtests experienced a larger drop in score between the first two tests, though the size of that drop remained small (0.03 of an ILR Level). The addition of the interaction terms also resulted in several of the ASVAB subtests as predictors of the intercept dropping out of significance (including ones that had been significant up until this model): *ASVAB-AO, -AS, -EI, -GS, -MC*. The *FSI0* measure entered as a main effect was also significant as a predictor of the intercept in this model (coeff. -0.16, $p < .001$).

The level-2 variance estimates were reduced for this final model and compared to the FSI model, the interactions explained another 2% of the variance in intercept and 1% in slope. Assumptions in this final model were checked. The level-2 residuals and the covariance matrix for the random slopes-piecewise time were examined for the presence of heteroscedasticity and autocorrelation. An alternate covariance structure was modeled but was rejected. There was deviation present at the tail of the plot as seen above at level-1, but in general, the assumptions for the final model were considered to be met.

Although multilevel modeling allows for unbalanced outcome data, to understand the population more thoroughly, individual differences for those with only

a few test occasions (up to three tests, n=5358) were compared to those with more test occasions (four or more, n=4194) using independent t-tests. A number of the ASVAB subtests were significant: *ASVAB-AO, -AS, -EI, -PC* and -*WK* and of the four DLAB subtests, Parts 1, 2 and 3 were significant. These findings indicated that there were differences between those who tested up to three times, and those who tested more than three times. The direction of the difference varied by subtest, however, with those with fewer tests having higher mean scores on *ASVAB-AO, t*(9550) =-2.52*, p* < .01*, DLAB Part 2, t*(9550) = -2.20*, p* < .01*,* and DLAB Part *3, t*(9550) = -2.41*, p* < .01, and lower scores on *ASVAB-AS, t*(9550) = 2.6*, p* < .01, *-EI, t*(9550) = 2.75*, p* < .01*, -PC, t*(9550) = 3.78*, p* < .01*, -WK*, t(9550) = 3.24, p < .01, and *DLAB Part 1 t*(9550) = 3.77, *p* < .001. There could be many reasons for the differing number of test occasions per individual. Some individuals attended training at the beginning of the time period captured by this dataset and completed their tour of duty, while others stayed in the service and continued testing. Other individuals tested at the end of the time period and therefore did not have the opportunity to be tested more than two or three times. However, the differences in these two populations were statistically significant.

Before moving on to the reading data, therefore, additional formulations of the slopes were compared to better assess the impact of attrition in the outcome data. The first slope, representing initial growth and coded as *Slope12*, was the same in all three trials. Three versions of subsequent growth were trialed, one which included only scores up through the fourth test (*Slope24*), one which included scores up through the fifth test (*Slope25*), and one which included all of the scores in the dataset (*Slope26*).

Table 20 below displays the coding system used. Missing test occasions from test occasions 5 and 6 were dropped from the *Slope24* and *Slope25* analyses.

Table 24

Alternate Recoding Schemes

Test Occasion	Slope12	Slope24	Slope25	Slope26
1	0	0	0	0
2	1	0	0	0
3	1	1	1	1
4	1	2	2	2
5	1	-	3	3
6	1	-	-	4
Number of tests	34,742	29,383	32,768	34742

The same full set of models conducted above for Study 1 (Listening HLM) with *Slope12/Slope26* were conducted using the piecewise slopes *Slope12/Slope24* and then again with *Slope12/Slope25*. The final model in all three trials of the slopes (which included all of the ASVAB, DLAB, and language distance interactions, for a total of 88 comparisons) was corrected for multiple comparisons and a summary chart of the significant variables in all three slope variations is shown below with the full chart in Appendix B. Had the estimates not been corrected, the table would include identical significant variables.

The results for the fixed effect estimates for the second slope differed only by 0.008 across the three models, and therefore, given the robustness of HLM with unevenly spaced outcomes (Finch, 2013; Hox, 2000; Raudenbush & Bryk, 2002) as

well as these findings, modeling continued with the reading data, and test occasions were modeled as two piecewise slopes, *Slope12* and *Slope26* to take advantage of the full dataset.

Table 25

Comparison chart of significant variables in variations of slopes

	Slope24	Slope25	Slope26
Intercept β_{00}	†	†	†
ZA_PC, β_{08}	†	†	†
ZA_WK, β_{09}	†	†	†
ZD_1, β_{010}	†	†	†
ZD_2, β_{011}		†	†
ZD_3, β_{012}		†	†
ZD_4, β_{013}	†	†	†
EDUC, β_{014}	†	†	†
PRIORPRO, β_{015}	†	†	†
ENGY, β_{016}	†	†	†
FSI0, β_{017}	†	†	†
D2_FSI0, β_{028}	†		
D3_FSI0, β_{029}	†		
For SLOPE12 slope, π_1			
INTRCPT2, β_{10}	†	†	†
ZA_MC, β_{16}	†	†	†
ZD_3, β_{111}	†	†	†
ZD_4, β_{112}	†	†	†
For SLOPE24 slope, π_2			
INTRCPT2, β_{20}	†	†	†

Study 2 Hierarchical Linear Modeling (Reading)

The reading data was modeled in a similar approach, first settling on the level-1 model and then adding in predictors at level-2. The first model, Equation 1 as above, served to set a baseline and outline the multilevel framework and provide the warrant for a multilevel model. There was a maximum of 34,935 level-1 units (test scores) and 9,564 level-2 units (individuals) in the reading data. The random level-1 coefficient for the intercept had a moderate reliability estimate of 0.76. The mean ILR Score (β_{00}) (reported throughout the research with robust standard errors unless otherwise noted) for reading was 2.40 ($p < .001$), an ILR score between ILR Levels 2 and 2+, a score that was slightly higher than the listening score. Random effects estimates showed the ILR score variance estimates at level-1, within-person, as $\sigma^2 = 0.121$ and τ_{00} (level-2, between-person) as 0.134 with both estimates significant ($p < .001$). Assuming normal distribution of the residuals, 95% of the individuals had mean ILR scores between 1.67 and 3.12 (i.e., 2.4 ± 1.96 [0.37]), which would mean scores between ILR Level 1+ and Level 3. The intraclass correlation (ICC), which describes the proportion of variance that lies between individuals, was calculated by dividing the variance at level-2 by the model's total variance and found to be 52%, which was taken as a warrant to continue with a multilevel model. The next step in modeling was to build out the level-1 model to investigate the rate and shape of growth.

For the reading data, time was first modeled as a linear function with a random intercept (repeating Equation 2 above). The fixed effect estimate for the intercept dropped slightly to 2.35, and the estimate for the slope was 0.03 ($p < .001$),

meaning that the ILR level rose, on average, only slightly across the six test occasions in the dataset. With only one variable describing time, the drop-and-recover pattern in the data is not visible. Because there were up to six test occasions in the data, time could be modeled with polynomial functions, so models with quadratic and cubic forms of time (Equations 3 and 4 above) were first tested with random intercepts. The output from these models is shown below in Table 26. These models did not explain additional level-1 variance. The cubic model showed a negative estimate for linear time, with an acceleration in the quadratic function followed by a slight decline in slope. As seen in the listening model above, this meant that the cubic effect dampened the acceleration (quadratic effect) of the linear downward slope. Models that allowed the time functions *Time, Time*^{2} and *Time*^{3} to vary had reliability estimates below 0.10 and therefore random slope models were not pursued further.

Table 26

HLM reading models of time

	Null Model	RI Time	RI Quad	RI Cube
		Fixed Effects		
Intercept	2.40*** (0.00)	2.35*** (0.00)	2.37*** (0.00)	2.38*** (0.01)
Time		0.03*** (0.00)	-0.01 (0.00)	-0.08*** (0.01)
Time Quad			0.01*** (0.00)	0.06*** (0.00)
Time Cube				-0.01*** (0.00)
		Random Effects		
Level-1, *e*	0.12	0.12	0.12	0.12
Level-2 intercept, *r*	0.13	0.13	0.13	0.13
		Goodness of fit		
Deviance	40039.1785	39396.8288	39290.4588	39184.259
Number of parameters	2	2	2	2

Note: Standard error in parentheses for estimated coefficients.

*** p < .001

As with the listening data, time was also modeled for reading with two piecewise slopes as in Equation 6 above. The output for this model is shown in the first column in Table 27. The fixed effect for the intercept dropped very slightly as compared to the Null model to 2.38, and the mean change of ILR level between the first two test occasions (*Slope12*) was negative (-0.04), while the mean change after the second test occasion (*Slope26)* was positive (0.05) and both estimates were significant ($p < .001$). The average growth trajectories were very shallow, with a very slight decrease in initial growth by a small increase in subsequent growth per test occasion.

Random slopes were tested next to allow the growth rate to vary across individuals, first allowing each slope to vary randomly (first *Slope12*, then *Slope26*)

and then testing a model with both slopes varying. The model with *Slope26* varying required an increase in iteration settings and the model converged at iteration 597, but the reliability estimate for the random level-1 coefficient *Slope26* was below 0.10 and therefore this model was rejected. A model with both slopes varying failed to converge. Therefore, the best fitting piecewise model appeared to be a model in which only *Slope12* was allowed to vary randomly. This model's output is reported in Table 27 along with the null model for comparison, and the random intercept, piecewise time model.

The intercept, the average ILR score in reading across individuals at the time of graduation, was 2.38; and the mean change of ILR level between the first two test occasions (*Slope12*) was negative (0.04 of an ILR level), while the mean change after the second test occasion (*Slope26*) was positive (0.05 of an ILR level). The piecewise model with a random first slope explained an additional 14% of the level-1 variance as compared to the null model, and the model was a better fit than the random intercept model with two slopes ($\chi^2 = 96.75$, $p < .01$, *df* 2). This model responded to the first two research questions, showing that reading scores varied by individual, and after dropping slightly following the first test occasion, rose slightly over time. The reading scores were, on average, higher than the listening scores, but the pattern of growth was quite similar. The 95% random effects confidence intervals for the intercept and random linear time indicated that 95% of the sample was expected to have individual intercepts ranging from 2.08 to 2.68 and the slope representing initial growth (*Slope12*) ranging from -0.13 to 0.05. While there was a significant rate of linear increase between the first two tests on average,

the random variation around this initial growth indicated that some individuals dropped in level and others rose. Assumptions of the final level-1 model (Slope12R) were checked as with the listening model, and in general the assumptions for the level-1 model were considered to have been met.

Table 27

HLM reading with piecewise slopes

	Null Model	Random Intercept, two slopes	Slope12 Random
		Fixed Effects	
Intercept	2.40*** (0.00)	2.38*** (0.00)	2.38*** (0.00)
Slope12		-0.04*** (0.00)	-0.04*** (0.00)
Slope26		0.05*** (0.00)	0.05*** (0.00)
		Random Effects	
Level-1, *e*	0.12	0.12	0.11
Level-2 intercept, r_0	0.13	0.13	0.15
Slope12, r_1			0.05
		Goodness of fit	
Deviance	40039.1785	39185.6549	39088.9063
Number of parameters	2	2	4

Note: Standard error in parentheses for estimated coefficients, standard deviation in parentheses for estimated variances.

*** $p < .001$

The next set of models was designed to model the extent to which aptitude was related to growth. Model building continued in a stepwise fashion to consider the effects of the level-2 covariates on proficiency scores and whether the scores varied by language. The subtest scores for ASVAB were standardized and centered on the grand mean and added simultaneously to the Slope12R model above to predict the intercept and both slopes. The mixed model equation for this model is above in Equation 5 and the results are in Table 28.

Seven of the nine ASVAB subtests were significantly related to the intercept (-*AO* and -*EI,* were not). None of the subtests were significantly related to the first slope and only -*AR*, -*PC* and -*WK* were significantly related to the second slope. The addition of the subtest variables changed the interpretation of the parameter estimates: the intercept was now understood as the average ILR score for those with mean ASVAB subtest scores, and the slopes were now the average change in those persons' ILR level between test occasions 1 and 2, and test occasions 2 and 6. Specifically, for those with average ASVAB scores, there was a significant mean ILR score (β_{00}= 2.38, $p < .001$) that dropped between test occasions 1 and 2 (β_{10}= -0.04, $p < .001$) and rose between occasions 2 and 6 (β_{20}= 0.05, $p < .001$), on average. Of the significant subtests, the -*AS* and -*MC* subtests were negative predictors of the intercept, while -*AR, -PC* and -*WK* were negative predictors of *Slope26*. This was interpreted to mean that those with higher scores on these subtests had lower ILR scores at graduation (-*AS* and -*MC*) or less of an increase in subsequent growth (-*AR, -PC* and -*WK)*. Variance component estimates once again confirmed significant variance in observed versus predicted ILR scores within-individual (level-1 variance σ^2 = .11, $p < .001$)

and significant variance in ILR scores at the time of graduation from DLI (level-2 variance $r_0 = .13$, $p < .001$). The estimate for the first slope, *Slope12*, was also significant and indicated that there was variance ($r_1 = .05$, $p < .001$) in proficiency scores between the first two test occasions. The level-1 variance was unchanged, as expected, because only level-2 predictors were added. The level-2 variance estimates decreased, and the addition of the ASVAB subtests explained 14% of the variation in level-2 intercept and 1% of the slope as compared to the Slope12R model above. Before adding the DLAB subtests to the model, a model with only the significant ASVAB subtests, shown below in Equation 15, was run and results are displayed in Table 28.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MC_i + \beta_{05}*ZA\_MK_i + \beta_{06}*ZA\_PC_i + \beta_{07}*ZA\_WK_i + \beta_{10}*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_AR_i*SLOPE26_{ti} + \beta_{22}*ZA\_PC_i*SLOPE26_{ti} + \beta_{23}*ZA\_WK_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (15)$$

As with the listening models, the next stage was to add the four DLAB subtests, resulting in Equation 16 below. The subtests were centered on the grand mean.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MC_i + \beta_{05}*ZA\_MK_i + \beta_{06}*ZA\_PC_i + \beta_{07}*ZA\_WK_i + \beta_{08}*ZD\_1_i + \beta_{09}*ZD\_2_i + \beta_{010}*ZD\_3_i + \beta_{011}*ZD\_4_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZD\_1_i*SLOPE12_{ti} + \beta_{12}*ZD\_2_i*SLOPE12_{ti} + \beta_{13}*ZD\_3_i*SLOPE12_{ti} + \beta_{14}*ZD\_4_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_AR_i*SLOPE26_{ti} + \beta_{22}*ZA\_PC_i*SLOPE26_{ti} + \beta_{23}*ZA\_WK_i*SLOPE26_{ti} +$$

β_{24}*ZD_1$_i$*SLOPE26$_{ti}$ + β_{25}*ZD_2$_i$*SLOPE26$_{ti}$ + β_{26}*ZD_3$_i$*SLOPE26$_{ti}$ +

β_{27}*ZD_4$_i$*SLOPE26$_{ti}$ + r_{0i} + r_{1i}*SLOPE12$_{ti}$ + e_{ti} (16)

The output from this model is in Table 28 below. The fixed estimates for the intercept and slopes were unchanged, and all four DLAB subtests were significant predictors of the intercept. The coefficients for *ASVAB-PC, ASVAB-WK and DLABPt3* were the highest estimates among the significant predictors of the intercept. Several of the ASVAB estimates shifted slightly after adding the DLAB subtests into the model, and *Slope26*ASVAB-AR* and *Slope26*ASVAB-PC* fell out of significance, though in practical terms these estimates were so small to begin with (approaching zero) the change in significance was not necessarily important.

The relationship between the DLAB subtests and each slope was different: for *Slope12*, *DLABPt1* and *-Pt3* were significant, with the direction of the estimate for *-Pt1* positive and for *-Pt3* negative. For *Slope26*, only the effect of *DLABPt3* was marginally significant, though the estimate itself was approaching zero (-0.004, p = .05). The addition of the DLAB variables explained 3% of the level 2 intercept variance as compared to the model with only ASVAB scores, but DLAB did not explain any variance in slope.

The output for an intervening model with only the significant aptitude models (DLABsig) is displayed in the fourth column of Table 28 and its equation in mixed form is shown below in Equation 17. This model's intercept would now be understood as the average ILR score in listening for individuals with mean scores in the following aptitude subtests: *ASVAB-AR, -AS, -GS, -MC, -MK, -PC and -WK*, as well as the four DLAB subtests. The coefficients $\beta_{01(VARIABLE)}$ through $\beta_{14(VARIABLE)}$

were the mean differences at graduation. Results showed that individuals with higher scores in *ASVAB-AR, -GS, -MK, -PC and -WK*, as well as DLAB scores in Parts 1, 2, 3, and 4 had significant increases in ILR level at graduation, though these estimates were quite small, ranging from 0.02 to 0.05. There were significant decreases between the first and second test occasions (-0.01, $p = .001$) and significant increases between the second and sixth test occasion (.05, $p < .001$).

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MC_i + \beta_{05}*ZA\_MK_i + \beta_{06}*ZA\_PC_i + \beta_{07}*ZA\_WK_i + \beta_{08}*ZD\_1_i + \beta_{09}*ZD\_2_i + \beta_{010}*ZD\_3_i + \beta_{011}*ZD\_4_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZD\_1_i*SLOPE12_{ti} + \beta_{12}*ZD\_3_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_WK_i*SLOPE26_{ti} + \beta_{22}*ZD\_3_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (17)$$

Table 28

HLM reading with aptitude subtests model parameters

	ASVAB	ASVABsig	DLAB	DLABsig
		Fixed Effect		
For INTRCPT1 π_0				
INTRCPT2 β_{00}	2.38*** (0.00)	2.38*** (0.00)	2.38*** (0.00)	2.38*** (0.00)
ZA_AO	0.00 (0.01)			
ZA_AR	0.04*** (0.01)	0.04*** (0.01)	0.03*** (0.01)	0.03 (0.01)
ZA_AS	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)
ZA_EI	0.00 (0.01)			
ZA_GS	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZA_MC	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)
ZA_MK	0.04*** (0.01)	0.05*** (0.00)	0.04*** (0.00)	0.04*** (0.00)
ZA_PC	0.05*** (0.01)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)
ZA_WK	0.07*** (0.01)	0.07*** (0.01)	0.05*** (0.01)	0.05*** (0.01)
ZD_1			0.02*** (0.01)	0.02*** (0.01)
ZD_2			0.02*** (0.01)	0.01** (0.00)
ZD_3,			0.05*** (0.01)	0.05*** (0.01)
ZD_4			0.03*** (0.01)	0.02*** (0.00)
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	-0.04*** (0.01)	-0.04*** (0.01)	-0.04*** (0.01)	-0.04*** (0.01)
ZA_AO	-0.01 (0.01)			
ZA_AR	0.00 (0.01)			

	ASVAB	ASVABsig	DLAB	DLABsig
ZA_AS	-0.01 (0.01)			
ZA_EI	0.00 (0.01)			
ZA_GS	0.00 (0.01)			
ZA_MC	-0.01 (0.01)			
ZA_MK	0.01 (0.01)			
ZA_PC	0.00 (0.01)			
ZA_WK	0.00 (0.01)			
ZD_1			0.02*** (0.01)	0.02*** (0.00)
ZD_2			-0.01 (0.01)	
ZD_3,			-0.01* (0.01)	-0.02*** (0.01)
ZD_4			-0.01 (0.01)	
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)
ZA_AO	0.00 (0.00)			
ZA_AR	0.00* (0.00)	0.00* (0.00)	0.00 (0.00)	
ZA_AS	0.00 (0.00)			
ZA_EI	0.00 (0.00)			
ZA_GS	0.00 (0.00)			
ZA_MC	0.00 (0.00)			
ZA_MK	0.00 (0.00)			
ZA_PC	0.00* (0.00)	0.00*** (0.00)	0.00 (0.00)	
ZA_WK	-0.01*** (0.00)	-0.01*** (0.00)	0.00* (0.00)	-0.01*** (0.00)
ZD_1			0.00 (0.00)	

	ASVAB	ASVABsig	DLAB	DLABsig
ZD_2			0.00 (0.00)	
ZD_3,			0.00* (0.00)	0.00*** (0.00)
ZD_4			0.00 (0.00)	
	Random Effect			
level-1, *e*	0.11 (0.33)	0.11 (0.33)	0.11 (0.33)	0.11 (0.33)
INTRCPT1, r_0	0.13*** (0.36)	0.13*** (0.36)	0.13*** (0.36)	0.13*** (0.36)
SLOPE12 slope, r_1	0.05*** (0.22)	0.05*** (0.22)	0.05*** (0.22)	0.05*** (0.22)

Note: Standard error in parentheses for estimated coefficients. Parameter designators from ASVAB model. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

The next step in model building was to add the four individual difference variables collected via the survey, *EDUC, MOT, PRIORPRO,* and *ENGY,* for the reasons explained above in Study 1. The mixed model equation for reading is below.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MC_i + \beta_{05}*ZA\_MK_i + \beta_{06}*ZA\_PC_i + \beta_{07}*ZA\_WK_i + \beta_{08}*ZD\_1_i + \beta_{09}*ZD\_2_i + \beta_{010}*ZD\_3_i + \beta_{011}*ZD\_4_i + \beta_{012}*EDUC_i + \beta_{013}*MOT_i + \beta_{014}*PRIPROF_i + \beta_{015}*ENGY_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZD\_1_i*SLOPE12_{ti} + \beta_{12}*ZD\_3_i*SLOPE12_{ti} + \beta_{13}*EDUC_i*SLOPE12_{ti} + \beta_{14}*MOT_i*SLOPE12_{ti} + \beta_{15}*PRIPROF_i*SLOPE12_{ti} + \beta_{16}*ENGY_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_WK_i*SLOPE26_{ti} + \beta_{22}*ZD\_3_i*SLOPE26_{ti} + \beta_{23}*EDUC_i*SLOPE26_{ti} + \beta_{24}*MOT_i*SLOPE26_{ti} + \beta_{25}*PRIPROF_i*SLOPE26_{ti} + \beta_{26}*ENGY_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti}$$

(18)

The output from this model is in Table 29. The estimated fixed effect coefficients and variance components for the aptitude variables were basically unchanged from the last model. For the reading data, the fixed effects for the motivation and prior proficiency variables were not significant as predictors of graduation (the intercept) or either slope (initial growth or subsequent growth), while education and English were. Those with higher levels of education and who had English as a first language were less likely to have higher ILR levels upon graduation, though this estimate was quite small and its significance marginal (-.005, $p = .05$). This was counterintuitive and contrary to findings in other research, but given that *EDUC* fell out of significance in subsequent models, it may be an artifact of the

dataset. Education was also significantly related to growth, such that those with higher education experienced less of a decrease in initial growth, subsequent growth was not as steep. The final model including only the significant survey variables is shown below in Equation 19. The addition of the survey variables did not change the variance for either the intercept or for the slope.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MC_i + \beta_{05}*ZA\_MK_i + \beta_{06}*ZA\_PC_i + \beta_{07}*ZA\_WK_i + \beta_{08}*ZD\_1_i + \beta_{09}*ZD\_2_i + \beta_{010}*ZD\_3_i + \beta_{011}*ZD\_4_i + \beta_{012}*ENGY_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZD\_1_i*SLOPE12_{ti} + \beta_{12}*ZD\_3_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_WK_i*SLOPE26_{ti} + \beta_{22}*ZD\_3_i*SLOPE26_{ti} + \beta_{23}*EDUC_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (19)$$

Table 29

HLM reading survey variables model parameters

	SURVEY	SURVEYSig (Final)
	Fixed Effect	
For INTRCPT1, π_0		
INTRCPT2, β_{00}	2.38*** (0.00)	2.38*** (0.00)
ZA_AR	0.03*** (0.01)	0.03*** (0.01)
ZA_AS	-0.02*** (0.01)	-0.02*** (0.01)
ZA_GS	0.02*** (0.01)	0.02*** (0.01)
ZA_MC	-0.03*** (0.01)	-0.03*** (0.01)
ZA_MK	0.04*** (0.00)	0.04*** (0.00)
ZA_PC	0.05*** (0.00)	0.05*** (0.00)
ZA_WK	0.06*** (0.01)	0.05*** (0.01)
ZD_1	0.02*** (0.01)	0.02*** (0.01)
ZD_2	0.01** (0.00)	0.01** (0.00)
ZD_3	0.05*** (0.01)	0.05*** (0.01)
ZD_4	0.02*** (0.00)	0.02*** (0.00)
EDUC	-0.01* (0.00) †	

MOT	0.00 (0.01)	
PRIPROF	0.00 (0.00)	
ENGY	-0.10*** (0.03)	-0.10*** (0.03)
For SLOPE12 slope, π_1		
INTRCPT2, β_{10}	-0.04*** (0.01)	-0.04*** (0.01)
ZD_1	0.01* (0.01)	0.02*** (0.00)
ZD_3	-0.02*** (0.01)	-0.02*** (0.01)
EDUC	0.01* (0.00) †	0.00 (0.00)
MOT	0.00 (0.01)	
PRIPROF	0.00 (0.00)	
ENGY	-0.01 (0.03)	
For SLOPE26 slope, π_2		
INTRCPT2	0.05*** (0.00)	0.05*** (0.00)
ZA_WK	-0.01** (0.00)	0.00** (0.00)
ZD_3	-0.01*** (0.00)	-0.01*** (0.00)
EDUC	0.00*** (0.00)	0.00*** (0.00)
MOT	0.00 (0.00)	
PRIPROF	0.00 (0.00)	
ENGY	0.02 (0.01)	
Final estimation of variance components		
	Random Effect	
level-1, *e*	0.11 (0.32)	0.11 (0.32)
INTRCPT1, r_0	0.13*** (0.36)	0.13*** (0.36)
SLOPE12 slope, r_1	0.05*** (0.22)	0.05*** (0.22)

Note: Standard error in parentheses for estimated coefficients. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05 † dropped out of significance in next model

The next model added the language distance measures (centered on the grand mean). These measures were treated individually. The mixed model equation for the FSILangCat measure is shown below in Equation 20. The output each of the six language distance measures: *FSI, GateRev, TypeRevRev, NotLatin, NotIndo,* and

LangCat (DLI) can be found below. For brevity, the mixed model equations for the other language distance measures are not repeated here.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AR_i + \beta_{02}*ZA\_AS_i + \beta_{03}*ZA\_GS_i + \beta_{04}*ZA\_MC_i + \beta_{05}*ZA\_MK_i + \beta_{06}*ZA\_PC_i + \beta_{07}*ZA\_WK_i + \beta_{08}*ZD\_1_i + \beta_{09}*ZD\_2_i + \beta_{010}*ZD\_3_i + \beta_{011}*ZD\_4_i + \beta_{012}*ENGY_i + \beta_{013}*FSI_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZD\_1_i*SLOPE12_{ti} + \beta_{12}*ZD\_3_i*SLOPE12_{ti} + \beta_{13}*FSI_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_WK_i*SLOPE26_{ti} + \beta_{22}*ZD\_3_i*SLOPE26_{ti} + \beta_{23}*EDUC_i*SLOPE26_{ti} + \beta_{24}*FSI_i*SLOPE26_{ti} + r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti} \quad (20)$$

The fixed effect and variance component estimates for the intercept and slope(s) were little changed across all of the models with the exception of DLAB Part 3. For reasons that are not clear, in all but one of the distance measures (*TypeRev*), the estimated coefficients for *DLABPt3* dropped out of significance in relationship to *Slope12*, while it also fell out of significance for *Slope26* in three of the measures (all except for *TypeRev*, *NotLatin* and *NotIndo*).

The relationships between each language distance measure and the intercept and slopes varied, possibly indicating that each measure tapped into a different aspect of language difficulty, though the estimates themselves were quite small. Three measures had a significant effect on the intercept (*FSI, TypeRev, NotIndo*); all six measures had significant estimates for *Slope12,* while five had significant estimates for Slope26 (*FSI, GateRev, TypeRev, NotLatin, DLI*). The direction of the language distance estimates varied for the intercept, but was consistently negative for the slopes, meaning that language distance constrained growth.

Interpreting effects and comparing models in multilevel analyses is complex, especially given the use of continuous and categorical variables. Language distance measures were interpreted as an estimate of the relationship of language distance with the ILR level for individuals with average scores on the predictor variables (aptitude and survey) at the time graduation (intercept), as well in growth (both slopes). In the case of two of the three significant predictors of the intercept, *FSI* and *NotIndo*, this relationship was negative. This meant that for those in harder languages (higher FSI categories or non-IndoEuropean languages, respectively), all else being equal, ILR levels at graduation were slightly lower. In the case of the slopes, significant relationships were negative, meaning a steeper drop between the first two test occasions for those in harder languages or a shallower rise thereafter. In almost all cases, however, practical significance was slight, as most estimates ranged from 0.00 to 0.05. The exception was the estimate for *Slope12*NotLatin*, which was -0.13.

The variance component estimates for the intercept were mostly unchanged as compared to the Survey model, which indicated that the language distance measures did not explain any additional variation in the intercept; but there was a reduction in the variance in the slope, as the PRV for *Slope12* ranged from 0.00 to 0.06, depending on the distance measure modeled.

Table 30

HLM reading language distance models

	FSI	GateRev	TypeRev	NotLatin	NotIndo	DLI
			Fixed Effect			
INTRCPT1π_0						
INTRCPT2, β_{00}	2.38*** (0.00)	2.38*** (0.00)	2.38*** (0.00)	2.38*** (0.00)	2.38*** (0.00)	2.38*** (0.00)
ZA_AR, β_{01}	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)
ZA_AS, β_{02}	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)	-0.02*** (0.01)
ZA_GS, β_{03}	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZA_MC, β_{04}	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)	-0.03*** (0.01)
ZA_MK, β_{05}	0.04*** (0.00)	0.04*** (0.00)	0.04*** (0.00)	0.04*** (0.00)	0.04*** (0.00)	0.04*** (0.00)
ZA_PC, β_{06}	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)
ZA_WK, β_{07}	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.05*** (0.01)
ZD_1, β_{08}	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)
ZD_2, β_{09}	0.02*** (0.00)	0.01*** (0.00)	0.01** (0.00)	0.01*** (0.00)	0.01*** (0.00)	0.01*** (0.00)
ZD_3, β_{010}	0.05*** (0.01)	0.05*** (0.01)	0.04*** (0.01)	0.05*** (0.01)	0.05*** (0.01)	0.04*** (0.01)
ZD_4, β_{011}	0.03*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.03*** (0.00)	0.02*** (0.00)
ENGY, β_{012}	-0.10 *** (0.03)	-0.10*** (0.03)	-0.10*** (0.03)	-0.10*** (0.03)	-0.10*** (0.03)	-0.10*** (0.03)
[lang dist]	-0.01*** (0.00)	0.00 (0.00)	0.00*** (0.00)	0.01 (0.02)	-0.03** (0.01)	0.01 (0.01)
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	-0.04 *** (0.01)	-0.04*** (0.01)	-0.04*** (0.01)	-0.04*** (0.01)	-0.04*** (0.01)	-0.04*** (0.01)
ZD_1, β_{11}	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)

	FSI	GateRev	TypeRev	NotLatin	NotIndo	DLI
ZD_3, β_{12}	0.00 (0.01)	0.00 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	0.00 (0.01)
[lang dist]	-0.02*** (0.00)	0.00*** (0.00)	0.00*** (0.00)	-0.13*** (0.02)	-0.04*** (0.01)	-0.05*** (0.01)
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)	0.05*** (0.00)
ZA_WK, β_{21}	0.00* (0.00)	0.00** (0.00)	0.00* (0.00)	0.00* (0.00)	0.00* (0.00)	0.00* (0.00)
ZD_3, β_{22}	0.00 (0.00)	0.00 (0.00)	0.00* (0.00)	0.00* (0.00)	0.00* (0.00)	0.00 (0.00)
EDUC, β_{23}	0.00*** (0.00)	0.00*** (0.00)	0.00*** (0.00)	0.00*** (0.00)	0.00*** (0.00)	0.00*** (0.00)
[lang dist]	0.00** (0.00)	0.00*** (0.00)	0.00** (0.00)	-0.02** (0.01)	0.00 (0.00)	-0.01*** (0.00)
Final estimation of variance components						
			Random Effect			
level-1, e	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)
INTRCPT1, r_0	0.13*** (0.36)	0.13*** (0.36)	0.13*** (0.36)	0.13*** (0.36)	0.13*** (0.36)	0.13*** (0.36)
SLOPE12 slope, r_1	0.05*** (0.22)	0.05*** (0.22)	0.05 *** (0.22)	0.05*** (0.22)	0.05*** (0.22)	0.05*** (0.22)

Note: Standard error in parentheses for estimated coefficients. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

The next stage in modeling addressed the last research question concerning the presence of an interaction effect on proficiency. The FSI0*[aptitude] terms were once again added to the Surveysig model as seen above in Study 1, along with the main effects for the DLAB and ASVAB subtests. The new model is below in Equation 21.

$$ILRNUM_{ti} = \beta_{00} + \beta_{01}*ZA\_AO_i + \beta_{02}*ZA\_AR_i + \beta_{03}*ZA\_AS_i + \beta_{04}*ZA\_EI_i + \beta_{05}*ZA\_GS_i + \beta_{06}*ZA\_MC_i + \beta_{07}*ZA\_MK_i + \beta_{08}*ZA\_PC_i + \beta_{09}*ZA\_WK_i + \beta_{010}*ZD\_1_i + \beta_{011}*ZD\_2_i + \beta_{012}*ZD\_3_i + \beta_{013}*ZD\_4_i + \beta_{014}*EDUC_i + \beta_{015}*ENGY_i + \beta_{016}*FSI0_i + \beta_{017}*AO\_FSI0_i + \beta_{018}*AR\_FSI0_i + \beta_{019}*AS\_FSI0_i + \beta_{020}*EI\_FSI0_i + \beta_{021}*GS\_FSI0_i + \beta_{022}*MC\_FSI0_i + \beta_{023}*MK\_FSI0_i + \beta_{024}*PC\_FSI0_i + \beta_{025}*WK\_FSI0_i + \beta_{026}*D1\_FSI0_i + \beta_{027}*D2\_FSI0_i + \beta_{028}*D3\_FSI0_i + \beta_{029}*D4\_FSI0_i + \beta_{10}*SLOPE12_{ti} + \beta_{11}*ZA\_AO_i*SLOPE12_{ti} + \beta_{12}*ZA\_AR_i*SLOPE12_{ti} + \beta_{13}*ZA\_AS_i*SLOPE12_{ti} + \beta_{14}*ZA\_EI_i*SLOPE12_{ti} + \beta_{15}*ZA\_GS_i*SLOPE12_{ti} + \beta_{16}*ZA\_MC_i*SLOPE12_{ti} + \beta_{17}*ZA\_MK_i*SLOPE12_{ti} + \beta_{18}*ZA\_PC_i*SLOPE12_{ti} + \beta_{19}*ZA\_WK_i*SLOPE12_{ti} + \beta_{110}*ZD\_1_i*SLOPE12_{ti} + \beta_{111}*ZD\_2_i*SLOPE12_{ti} + \beta_{112}*ZD\_3_i*SLOPE12_{ti} + \beta_{113}*ZD\_4_i*SLOPE12_{ti} + \beta_{114}*EDUC_i*SLOPE12_{ti} + \beta_{115}*FSI0_i*SLOPE12_{ti} + \beta_{116}*AO\_FSI0_i*SLOPE12_{ti} + \beta_{117}*AR\_FSI0_i*SLOPE12_{ti} + \beta_{118}*AS\_FSI0_i*SLOPE12_{ti} + \beta_{119}*EI\_FSI0_i*SLOPE12_{ti} + \beta_{120}*GS\_FSI0_i*SLOPE12_{ti} + \beta_{121}*MC\_FSI0_i*SLOPE12_{ti} + \beta_{122}*MK\_FSI0_i*SLOPE12_{ti} + \beta_{123}*PC\_FSI0_i*SLOPE12_{ti} + \beta_{124}*WK\_FSI0_i*SLOPE12_{ti} + \beta_{125}*D1\_FSI0_i*SLOPE12_{ti} + \beta_{126}*D2\_FSI0_i*SLOPE12_{ti} + \beta_{127}*D3\_FSI0_i*SLOPE12_{ti} + \beta_{128}*D4\_FSI0_i*SLOPE12_{ti} + \beta_{20}*SLOPE26_{ti} + \beta_{21}*ZA\_AO_i*SLOPE26_{ti} + \beta_{22}*ZA\_AR_i*SLOPE26_{ti} + \beta_{23}*ZA\_AS_i*SLOPE26_{ti} + \beta_{24}*ZA\_EI_i*SLOPE26_{ti} + \beta_{25}*ZA\_GS_i*SLOPE26_{ti} +$$

$\beta_{26}*ZA\_MC_i*SLOPE26_{ti} + \beta_{27}*ZA\_MK_i*SLOPE26_{ti} + \beta_{28}*ZA\_PC_i*SLOPE26_{ti} +$

$\beta_{29}*ZA\_WK_i*SLOPE26_{ti} + \beta_{210}*ZD\_1_i*SLOPE26_{ti} + \beta_{211}*ZD\_2_i*SLOPE26_{ti} +$

$\beta_{212}*ZD\_3_i*SLOPE26_{ti} + \beta_{213}*ZD\_4_i*SLOPE26_{ti} + \beta_{214}*EDUC_i*SLOPE26_{ti} +$

$\beta_{215}*FSI0_i*SLOPE26_{ti} + \beta_{216}*AO\_FSI0_i*SLOPE26_{ti} + \beta_{217}*AR\_FSI0_i*SLOPE26_{ti} +$

$\beta_{218}*GS\_FSI0_i*SLOPE26_{ti} + \beta_{219}*MC\_FSI0_i*SLOPE26_{ti} +$

$\beta_{220}*MK\_FSI0_i*SLOPE26_{ti} + \beta_{221}*PC\_FSI0_i*SLOPE26_{ti} +$

$\beta_{222}*WK\_FSI0_i*SLOPE26_{ti} + \beta_{223}*D1\_FSI0_i*SLOPE26_{ti} +$

$\beta_{224}*D2\_FSI0_i*SLOPE26_{ti} + \beta_{225}*D3\_FSI0_i*SLOPE26_{ti} + \beta_{226}*D4\_FSI0_i*SLOPE26_{ti} +$

$r_{0i} + r_{1i}*SLOPE12_{ti} + e_{ti}$ (21)

Reviewing the main effects for the intercept, the *ASVAB-AO, -AS, -EI* and *-GS* estimated coefficients for the intercept were all non-significant, while *-MC*, *-MK* had *p*-values less than the B-H critical value and therefore were also considered to be non-significant. The four DLAB subtests' estimated coefficients for the intercept were significant, though *DLABPt2* had a *p*-value less than the critical value. None of the interaction terms were significant predictors of the intercept after the correction. For the two slopes, the only remaining significant variable with an effect on either slope after correcting for the false discovery rate was the *EDUC* variable, which had a negative estimated coefficient for the second slope, and the estimate itself was approaching zero (0.004, $p = 0.002$). None of the aptitude variables, whether as main effects or as interaction terms, had an effect on either slope. The results are displayed below in Table 31. The model with interactions explained an additional 1% of the variance in intercept and 4% of the first slope as compared to the FSI model.

Table 31

HLM reading FSI0 interaction model

	Coefficient	p < BH
	Fixed Effect	
For INTRCPT1, π_0		
INTRCPT2, β_{00}	2.38*** (0.00)	†
ZA_AO, β_{01}	0.00 (0.01)	
ZA_AR, β_{02}	0.03*** (0.01)	†
ZA_AS, β_{03}	0.00 (0.01)	
ZA_EI, β_{04}	0.01 (0.01)	
ZA_GS, β_{05}	0.01 (0.01)	
ZA_MC, β_{06}	-0.02* (0.01)	
ZA_MK, β_{07}	0.02* (0.01)	
ZA_PC, β_{08}	0.05*** (0.01)	†
ZA_WK, β_{09}	0.07*** (0.01)	†
ZD_1, β_{010}	0.03*** (0.01)	†
ZD_2, β_{011}	0.02*** (0.01)	
ZD_3, β_{012}	0.06*** (0.01)	†
ZD_4, β_{013}	0.04*** (0.01)	†
EDUC, β_{014}	-0.01* (0.00)	
ENGY, β_{015}	-0.10*** (0.03)	†
FSI0, β_{016}	-0.08*** (0.01)	†
AO_FSI0, β_{017}	0.00 (0.01)	
AR_FSI0, β_{018}	0.00 (0.01)	
AS_FSI0, β_{019}	-0.02 (0.01)	
EI_FSI0, β_{020}	0.00 (0.01)	
GS_FSI0, β_{021}	0.02 (0.01)	
MC_FSI0, β_{022}	-0.02 (0.01)	
MK_FSI0, β_{023}	0.02* (0.01)	
PC_FSI0, β_{024}	-0.01 (0.01)	
WK_FSI0, β_{025}	-0.03* (0.01)	
D1_FSI0, β_{026}	0.00 (0.01)	
D2_FSI0, β_{027}	0.00 (0.01)	
D3_FSI0, β_{028}	0.00 (0.01)	
D4_FSI0, β_{029}	-0.02* (0.01)	
For SLOPE12 slope, π_1		

	Coefficient	p < BH
INTRCPT2, β_{10}	-0.04*** (0.01)	†
ZA_AO, β_{11}	-0.01 (0.01)	
ZA_AR, β_{12}	0.00 (0.01)	
ZA_AS, β_{13}	-0.01 (0.01)	
ZA_EI, β_{14}	0.01 (0.01)	
ZA_GS, β_{15}	0.00 (0.01)	
ZA_MC, β_{16}	-0.01 (0.01)	
ZA_MK, β_{17}	0.01 (0.01)	
ZA_PC, β_{18}	0.00 (0.01)	
ZA_WK, β_{19}	-0.01 (0.01)	
ZD_1, β_{110}	0.01 (0.01)	
ZD_2, β_{111}	0.00 (0.01)	
ZD_3, β_{112}	-0.01 (0.01)	
ZD_4, β_{113}	-0.01 (0.01)	
EDUC, β_{114}	0.01** (0.00)	
FSI0, β_{115}	-0.03*** (0.01)	†
AO_FSI0, β_{116}	0.00 (0.01)	
AR_FSI0, β_{117}	-0.01 (0.01)	
AS_FSI0, β_{118}	-0.01 (0.01)	
EI_FSI0, β_{119}	-0.01 (0.01)	
GS_FSI0, β_{120}	-0.01 (0.01)	
MC_FSI0, β_{121}	0.00 (0.01)	
MK_FSI0, β_{122}	0.00 (0.01)	
PC_FSI0, β_{123}	0.01 (0.01)	
WK_FSI0, β_{124}	0.00 (0.01)	
D1_FSI0, β_{125}	0.01 (0.01)	
D2_FSI0, β_{126}	0.00 (0.01)	
D3_FSI0, β_{127}	0.01 (0.01)	
D4_FSI0, β_{128}	0.02 (0.01)	
For SLOPE26 slope, π_2		
INTRCPT2, β_{20}	0.05*** (0.00)	†
ZA_AO, β_{21}	0.00 (0.00)	
ZA_AR, β_{22}	0.00 (0.00)	
ZA_AS, β_{23}	0.00 (0.00)	
ZA_EI, β_{24}	0.00 (0.00)	
ZA_GS, β_{25}	0.00 (0.00)	
ZA_MC, β_{26}	0.00 (0.00)	

	Coefficient	p < BH
ZA_MK, β_{27}	0.00 (0.00)	
ZA_PC, β_{28}	0.00 (0.00)	
ZA_WK, β_{29}	-0.01** (0.00)	
ZD_1, β_{210}	0.00 (0.00)	
ZD_2, β_{211}	0.00 (0.00)	
ZD_3, β_{212}	0.00 (0.00)	
ZD_4, β_{213}	0.00 (0.00)	
EDUC, β_{214}	0.00*** (0.00)	†
FSI0, β_{215}	0.00 (0.00)	
AO_FSI0, β_{216}	0.00 (0.00)	
AR_FSI0, β_{217}	0.00 (0.00)	
GS_FSI0, β_{218}	0.00 (0.00)	
MC_FSI0, β_{219}	0.00 (0.00)	
MK_FSI0, β_{220}	0.00 (0.00)	
PC_FSI0, β_{221}	0.00 (0.00)	
WK_FSI0, β_{222}	0.01 (0.00)	
D1_FSI0, β_{223}	0.00 (0.00)	
D2_FSI0, β_{224}	0.01 (0.00)	
D3_FSI0, β_{225}	0.00 (0.00)	
D4_FSI0, β_{226}	0.00 (0.00)	

Final estimation of variance components	
	Random Effect
level-1, e	0.11 (0.32)
INTRCPT1, r_0	0.13*** (0.36)
SLOPE12 slope, r_1	0.05*** (0.22)

Note: Standard error in parentheses for estimated coefficients. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05 † p < BH critical

Assumptions in this final model were checked. The Q-Q plot still retained the deviance at the tail as seen above in the level-1 model. Alternate covariance structures

were attempted but models failed to converge. Further modeling to explore other language distance measure interactions was not conducted.

Study 3 Hierarchical Generalized Linear Modeling (Listening)

HLM8 (Raudenbush & Congdon, 2021) was used to model the listening data with the outcome treated as an ordinal measure. Due to the low number of language clusters in the data, a two-level generalized linear model was used rather than a three-level model. The first model tested in listening was an empty random intercept model, and the results for this model are displayed in Table 33. The model produced the estimates for the effect of explanatory variables on the odds of an individual being at or below each ILR level. The odds ratio (OR) provided an estimated probability of being at or below a proficiency level. In the null model for the listening data, the results showed that across all the individuals in the dataset, the expected log odds of being at Level 0/0+ was negative (–6.36, $p < .001$), meaning that it was more likely that individuals were at least ILR Level 1 or higher. The probability of an individual being at or below ILR Level 0/0+ was 0.0017 or .02%. There was significant variance ($p < .001$) in the intercept (level at graduation) at level-2 (4.13, $p < .001$). Given the importance of reaching ILR Level 2, the model predictions for being at or below ILR Level 2 were calculated and the predicted probability of being at or below ILR Level 2 was 612.89, or 61.29%. The analysis also showed significant variation between individuals in their intercepts, $\tau_{00} = 4.132$ ($SE = 2.03$). $\chi^2 = 42595.522$, $p < .001$. The level-2 ICC was calculated for this model following Snijders & Bosker (1999), as cited in O'Connell (2010) where "in a logistic model, the level-1 residuals are assumed to follow the standard logistic distribution, which has a mean of 0 and a variance of $\pi 2 / 3 = 3.29$" (p. 5). Therefore, the total variance in the null model was 7.42 [3.29 (level-1) + 4.13 (level-2)]. The level-2 ICC for the null model was then

calculated as .56 which indicated that 56% of the variance was attributed to the individual. This value matched the ICC seen in Study 1 above and provided the warrant to continue with a two-level model of growth. Before moving on to examine how predictors might account for the heterogeneity between individuals, the longitudinal nature of the data was modeled with two slopes, *Slope12* (initial growth) and *Slope26* (subsequent growth).

Table 32

HGLM listening null model

	Coefficient (*se*)	Odds Ratio	C.I.
Fixed Effects			
Intercept	-6.36*** (0.09)	0.00	(0.001,0.002)
For THOLD2, δ_2	1.78*** (0.07)	5.91	(5.118,6.827)
For THOLD3, δ_3	3.91*** (0.08)	49.96	(42.602,58.586)
For THOLD4, δ_4	6.42*** (0.08)	612.89	(519.630,722.888)
For THOLD5, δ_5	8.56*** (0.09)	5,229.51	(4418.467,6189.365)
Random effects			
Intercept Variance, r_0	4.13		

Note: Standard error in parentheses for estimated coefficients.

*** p < .001

The output for the random intercept model with piecewise slopes (RISlopes) is reported below. Both slopes were significant (χ^2 25,184, $p < .001$), with a positive coefficient for the first slope and negative for the second. As explained above, the

interpretation of the coefficients in the HGLM model is opposite that of the earlier models, thus these results indicated that the likelihood of falling test scores was higher between the first two test occasions (positive estimate), and rising test scores thereafter (negative estimate), the same pattern seen above. The results could also be used to predict the probability of being at or below a given ILR level at various time points. For example, the predicted logit for an individual at or below ILR Level 2 at the third test occasion (time=2) was -0.209 [$\beta_{00} + \beta_{10}*2 + \beta_{20}*2 + \delta 4$], with the predicted probability as 81%. Overall, the fixed and random effects estimates changed only slightly from those in the Null Model, as can be seen in Table 33 below.

Table 33

HGLM listening piecewise slopes, random intercept and random Slope12 models

	RISlopes		Slope12R	
	Coefficient	Odds Ratio	Coefficient	Odds Ratio
Fixed Effects				
For INTRCPT1 slope, π_0				
INTRCPT2, β_{00}	-6.35***	0.00	-6.43***	0.00
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	0.19***	1.21	0.19***	1.20
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	-0.27***	0.76	-0.27***	0.76
For THOLD2,				
δ_2	1.79***	5.96	1.80***	6.08
For THOLD3,				
δ_3	3.94***	51.42	3.99***	54.14
For THOLD4,				
δ_4	6.49***	657.54	6.58***	717.57
For THOLD5,				
δ_5	8.67***	5833.70	8.79***	6578.16
Final estimation of variance components				
Random Effects				

INTRCPT1, r_0	4.23***	4.61***
SLOPE12 slope, r_1		0.66***

*** p < .001

The next series of models tested the effects of allowing the slopes to vary randomly, first with *Slope12* random, then with *Slope26* random, then with both slopes random. Iteration settings had to be raised from the default of 100 to 500 to reach convergence. The random level-1 estimated coefficient reliability estimate was only marginal, but it remained above 0.10, and the random slope variance component estimate was significant ($p < .001$). Models with a varying second slope failed to converge, so it was fixed in later modeling. The Slope12R model is above in Table 33. The estimated coefficients changed slightly between the random intercept and random first slope models. Although convergence took more iterations, the random slope variance estimate was significant, and based on the earlier modeling and the patterns in the observed data, modeling continued with piecewise slopes, with the first slope allowed to vary and the second slope fixed.

The next set of models was designed to explore aptitude. The general aptitude subtests (ASVAB), and then language aptitude battery (DLAB), were added to the model as in the studies above. The ASVAB subtests were first added to the model as predictors of graduation (the intercept) and growth (two slopes), and the model output is shown below. The standardized scores from the subtests were centered by grand mean when added to the model. As seen in Study 1 above, the *ASVAB-AO, EI* and -*MC* failed to reach significance as predictors of the intercept. Of the six other subtests, the estimates for -*AR, -GS, -MK -PC* and -*WK* were all negative, and the estimate for *ASVAB-AS* was positive. The subtests with the largest effects were -*PC*

and -*WK*. Only *ASVAB-MC* was significantly related to *Slope12,* again parallel with the findings in Study 1. None of the ASVAB subtests were significantly related to the second slope. The ASVAB model described the effect of the nine subtests on the intercept and slopes. *ASVAB-AR, -GS, -MK, -PC* and -*WK* tended to decrease the logit, making it more likely that individuals with higher scores on these subtests would be in higher proficiency categories relative to those with lower scores at the time of graduation from DLI. Those with higher *ASVAB-AS* scores would be more likely to have lower ILR levels relative to those with lower scores. This may be an artifact of selection based on the *ASVAB-AFQT* score into the basic course at DLIFLC. In the ASVAB model, the slope coefficients were interpreted to be the estimated slopes for those with mean ASVAB scores. Only one subtest, *ASVAB-MC*, had a statistically significant relationship with growth, in this case the first slope, and the results indicated that individuals with higher scores on the -*MC* subtest would be less likely to be in higher proficiency categories between the first two test occasions. The addition of the ASVAB variables to the Slope12R model reduced the variance in the level-2 intercept by 8% and the variation in the first slope by 7%.

An intervening model with only the significant ASVAB variables was run, and the results are shown in Table 34 below. Language aptitude predictor variables were then added to the model as predictors of the intercept and both slopes. The estimated coefficients for the intercept and slopes did not change with the addition of the language aptitude variables, but many of general aptitude estimates changed slightly. *DLABPt1, DLABPt2* and *DLABPt3* were all significantly related to the intercept, while *DLABPt1*, *DLABPt3* and *DLABPt4* were significantly related to the

first slope. Only *DLABPt3* was significantly related to the second slope. These results mirrored the findings in the HLM listening models in Study 1. The results are to be interpreted as follows: positive, significant effects make the likelihood of being in a higher proficiency category lower for those with below average aptitude scores, and negative effects make the likelihood of being in a higher ILR category higher for those with below average scores. The logic is similar for the slopes. The first slope coefficient was interpreted as the estimated slope for an individual with average aptitude scores. Controlling for the other subtests, a negative significant effect of a subtest meant that the likelihood was greater that an individual with higher scores on that subtest improved over time.

The addition of the DLAB variables reduced the level-2 variances in the model by 2% as compared to the model with only the ASVAB variables. The four survey variables were next entered into the model as in the studies above.

Table 34

HGLM listening ASVAB and ASVABsig models

	ASVAB			ASVABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
			Fixed Effect			
For INTRCPT1 slope, π_0)						
INTRCPT2, β_{00}	-6.43***	0.00	(0.001,0.002)	-6.43***	0.00	(0.001,0.002)
ZA_AO, β_{01}	0.04	1.04	(0.974,1.104)			
ZA_AR, β_{02}	-0.15***	0.86	(0.797,0.926)	-0.12***	0.89	(0.838,0.944)
ZA_AS, β_{03}	0.10**	1.11	(1.027,1.196)	0.10***	1.11	(1.051,1.170)
ZA_EI, β_{04}	0.03	1.03	(0.951,1.119)			
ZA_GS, β_{05}	-0.13***	0.88	(0.811,0.948)	-0.10***	0.90	(0.849,0.958)
ZA_MC, β_{06}	0.03	1.03	(0.945,1.114)			
ZA_MK, β_{07}	-0.19***	0.83	(0.774,0.891)	-0.20***	0.82	(0.776,0.871)
ZA_PC, β_{08}	-0.21***	0.81	(0.758,0.864)	-0.18***	0.84	(0.795,0.885)
ZA_WK, β_{09}	-0.25***	0.78	(0.728,0.841)	-0.25***	0.78	(0.732,0.824)
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	0.19***	1.21	(1.145,1.277)	0.19***	1.21	(1.144,1.277)
ZA_AO, β_{11}	0.06	1.06	(0.998,1.127)			
ZA_AR, β_{12}	0.01	1.01	(0.937,1.083)			
ZA_AS, β_{13}	-0.01	0.99	(0.919,1.066)			
ZA_EI, β_{14}	-0.03	0.97	(0.895,1.048)			
ZA_GS, β_{15}	0.02	1.02	(0.948,1.104)			

	ASVAB			ASVABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
ZA_MC, β_{16}	0.11**	1.12	(1.030,1.213)	0.13***	1.14	(1.087,1.192)
ZA_MK, β_{17}	-0.01	0.99	(0.923,1.059)			
ZA_PC, β_{18}	0.06	1.06	(0.994,1.130)			
ZA_WK, β_{19}	-0.03	0.98	(0.911,1.044)			
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.27***	0.76	(0.746,0.777)	-0.27***	0.76	(0.745,0.776)
ZA_AO, β_{21}	0.00	1.00	(0.980,1.026)			
ZA_AR, β_{22}	0.02	1.02	(0.992,1.045)			
ZA_AS, β_{23}	-0.01	0.99	(0.967,1.020)			
ZA_EI, β_{24}	-0.01	0.99	(0.964,1.020)			
ZA_GS, β_{25}	0.02	1.02	(0.988,1.044)			
ZA_MC, β_{26}	-0.02	0.98	(0.955,1.013)			
ZA_MK, β_{27}	-0.01	0.99	(0.963,1.013)			
ZA_PC, β_{28}	-0.01	0.99	(0.972,1.018)			
ZA_WK, β_{29}	0.01	1.01	(0.988,1.039)			
For THOLD2,						
δ_2	1.80***	6.07	(5.25,7.029)	1.80***	6.07	(5.25,7.03)
For THOLD3,						
δ_3	3.99***	54.14	(46.07,63.64)	3.99***	54.11	(46.04,63.60)
For THOLD4,						
δ_4	6.58***	719.80	(608.73,851.13)	6.58***	718.95	(607.98,850.18)

	ASVAB			ASVABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
For THOLD5,						
δ_5	8.80***	6626.00	(5582.72,7864.26)	8.80***	6616.15	(5574.11,7853.00)

Final estimation of variance components

Random Effect	Variance	d.f.	χ^2	Variance	d.f.	χ^2
INTRCPT1, r_0	4.26*** (2.06)	8234	15160.17212	4.26*** (2.06)	8237	15153.24869
SLOPE12 slope, r_1	0.62*** (0.79)	8234	8656.99908	0.62*** (0.79)	8242	8666.536

Note: Standard deviation in parentheses for estimated variances. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

Table 35

HGLM listening DLAB and DLAB sig models

	DLAB			DLABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
			Fixed Effect			
For INTRCPT1 slope, π_0						
INTRCPT2, β_{00}	-6.43***	0.00	(0.001,0.002)	-6.43	0.00	(0.001,0.002)

	DLAB			DLABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
ZA_AO, β_{01}						
ZA_AR, β_{02}	-0.08**	0.92	(0.866,0.976)	-0.09	0.92	(0.863,0.972)
ZA_AS, β_{03}	0.09***	1.10	(1.041,1.159)	0.09	1.10	(1.041,1.159)
ZA_EI, β_{04}						
ZA_GS, β_{05}	-0.11***	0.90	(0.846,0.955)	-0.11	0.90	(0.844,0.953)
ZA_MC, β_{06}						
ZA_MK, β_{07}	-0.16***	0.86	(0.808,0.907)	-0.16	0.85	(0.806,0.905)
ZA_PC, β_{08}	-0.16***	0.85	(0.808,0.900)	-0.16	0.85	(0.807,0.898)
ZA_WK, β_{09}	-0.20***	0.82	(0.775,0.872)	-0.20	0.82	(0.773,0.870)
ZD_1, β_{07}	-0.13***	0.88	(0.831,0.932)	-0.13***	0.88	(0.830,0.931)
ZD_2, β_{08}	-0.15***	0.86	(0.814,0.915)	-0.12***	0.89	(0.848,0.934)
ZD_3, β_{09}	-0.21***	0.81	(0.762,0.861)	-0.22***	0.80	(0.754,0.851)
ZD_4, β_{010}	-0.03	0.97	(0.912,1.033)			
For SLOPE12 slope, π_1				0.19***	1.21	(1.143,1.276)
INTRCPT2, β_{10}	0.19***	1.21	(1.144,1.277)			
ZA_AO, β_{11}						
ZA_AR, β_{12}						
ZA_AS, β_{13}						
ZA_EI, β_{14}						
ZA_GS, β_{15}						
ZA_MC, β_{16}	0.09***	1.09	(1.043,1.148)	0.10***	1.10	(1.049,1.154)

	DLAB			DLABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
ZA_MK, β_{17}						
ZA_PC, β_{18}						
ZA_WK, β_{19}						
ZD_1, β_{12}	-0.09***	0.92	(0.866,0.968)	-0.08***	0.92	(0.875,0.970)
ZD_2, β_{13}	0.04	1.05	(0.989,1.107)			
ZD_3, β_{14}	0.06*	1.07	(1.005,1.129)	0.08*	1.08	(1.021,1.142)
ZD_4, β_{15}	0.09***	1.10	(1.036,1.165)	0.07***	1.08	(1.028,1.126)
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.27***	0.76	(0.747,0.778)	-0.27***	0.76	(0.747,0.778)
ZA_AO, β_{21}						
ZA_AR, β_{22}						
ZA_AS, β_{23}						
ZA_EI, β_{24}						
ZA_GS, β_{25}						
ZA_MC, β_{26}						
ZA_MK, β_{27}						
ZA_PC, β_{28}						
ZA_WK, β_{29}						
ZD_1, β_{21}	0.00	1.00	(0.985,1.025)			
ZD_2, β_{22}	0.00	1.00	(0.981,1.023)			
ZD_3, β_{23}	0.03*	1.03	(1.006,1.050)	0.03*	1.03	(1.007,1.049)

	DLAB			DLABsig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
ZD_4, β_{24}	0.00	1.00	(0.976,1.016)			
For THOLD2, δ_2	1.80***	6.06	(5.241,7.017)	1.80***	6.06	(5.239,7.015)
For THOLD3, δ_3	3.99***	53.99	(45.941,63.448)	3.99***	53.96	(45.912,63.416)
For THOLD4, δ_4	6.58***	718.25	(607.460,849.240)	6.58***	717.54	(606.834,848.442)
For THOLD5, δ_5	8.80***	6628.72	(5585.224,7867.173)	8.80***	6620.83	(5578.306,7858.200)

Final estimation of variance components

	Random Effect					
	Variance	d.f.	χ^2	Variance	d.f.	χ^2
INTRCPT1, r_0	4.18*** (2.04)	8233	15037.47825	4.18*** (2.04)	8234	15031.80561
SLOPE12 slope, r_1	0.60*** (0.78)	8238	8635.46489	0.60*** (0.78)	8239	8635.78561

Note: Standard deviation in parentheses for estimated variances. All level-2 variables centered on grand mean.

*** p < .005 * p < .05

The four survey variables were next added (group centered) into the last aptitude model (DLABsig) as predictors of graduation (intercept) and growth (both slopes). The results are shown in Table 36. Of the four survey variables, *EDUC, PRIORPRO and ENGY* were significantly related to the Intercept ($p < .001$). The estimate for the fixed effect of *PRIORPRO* as a predictor of graduation (*Intercept*) was negative, and *EDUC* and *ENGY* were both positive. For initial growth (*Slope12*), the estimated coefficient for *EDUC* was -0.03 and of marginal significance at $p = .052$. None of the survey variables were significant in their relationship with subsequent growth (*Slope26*). The fixed effect coefficient estimates for the aptitude variables shifted with the addition of the survey variables, either increasing slightly or decreasing, but very slightly given the scale of these estimates. The addition of the survey variables reduced the level-2 intercept variance by 1% and the level-2 slope12 variance was unchanged. The *Slope12*EDUC* variable dropped out of significance in an intervening model, so in later models there were no remaining survey variables in as predictors of growth.

Intercept: *ASVAB-AR, -AS, -GS, -MK, -PC, -WK, DLABPt1, DLABPt 2, DLABPt3, EDUC, PriorPro, ENGY*

Slope12: *ASVAB-MC*, *DLABPt1, DLABPt 3, DLABPt4*

Slope26: *DLABPt 3*

Table 36

HGLM listening with survey variables

	Survey			SurveySig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
			Fixed effect			
For INTRCPT1 slope, π_0						
INTRCPT2, β_{00}	-6.43***	0.00	(0.001,0.002)	-6.43***	0.00	(0.001,0.002)
ZA_AR, β_{01}	-0.11***	0.90	(0.848,0.955)	-0.11***	0.90	(0.847,0.954)
ZA_AS, β_{02}	0.06*	1.07	(1.011,1.126)	0.07*	1.07	(1.011,1.126)
ZA_GS, β_{03}	-0.10***	0.91	(0.855,0.963)	-0.10***	0.91	(0.854,0.963)
ZA_MK, β_{04}	-0.15***	0.86	(0.815,0.914)	-0.15***	0.86	(0.815,0.914)
ZA_PC, β_{05}	-0.19***	0.83	(0.787,0.877)	-0.19***	0.83	(0.788,0.877)
ZA_WK, β_{06}	-0.26***	0.77	(0.724,0.817)	-0.26***	0.77	(0.723,0.817)
ZD_1, β_{07}	-0.18***	0.83	(0.783,0.889)	-0.18***	0.84	(0.786,0.892)
ZD_2, β_{08}	-0.11***	0.90	(0.854,0.941)	-0.11***	0.90	(0.855,0.942)
ZD_3, β_{09}	-0.21***	0.81	(0.763,0.862)	-0.21***	0.81	(0.765,0.863)
EDUC, β_{010}	0.12***	1.13	(1.089,1.167)	0.12***	1.13	(1.089,1.167)
MOT, β_{011}	0.01	1.01	(0.948,1.070)			
PRIORPRO, β_{012}	-0.11***	0.90	(0.847,0.946)	-0.12***	0.88	(0.845,0.923)
ENGY, β_{013}	0.78***	2.17	(1.484,3.187)	0.74***	2.10	(1.499,2.940)
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	0.19***	1.21	(1.143,1.276)	0.19***	1.21	(1.142,1.275)
ZA_MC, β_{11}	0.10***	1.10	(1.050,1.155)	0.10***	1.10	(1.052,1.157)

	Survey			SurveySig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
ZD_1, β_{12}	-0.06*	0.94	(0.890,0.998)	-0.06*	0.94	(0.888,0.992)
ZD_3, β_{13}	0.08**	1.09	(1.026,1.150)	0.08**	1.08	(1.022,1.143)
ZD_4, β_{14}	0.08***	1.08	(1.031,1.129)	0.08***	1.08	(1.031,1.130)
EDUC, β_{15}	-0.03*	0.97	(0.942,1.000)	-0.03	0.98	(0.948,1.003)
MOT, β_{16}	0.00	1.00	(0.937,1.067)			
PRIORPRO, β_{17}	-0.04	0.96	(0.911,1.015)			
ENGY, β_{18}	-0.10	0.90	(0.621,1.306)			
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.27***	0.76	(0.746,0.777)	-0.27***	0.76	(0.747,0.778)
ZD_3, β_{21}	0.02*	1.02	(1.002,1.044)	0.03**	1.03	(1.007,1.048)
EDUC, β_{22}	0.00	1.00	(0.993,1.016)			
MOT, β_{23}	0.02	1.02	(0.994,1.040)			
PRIORPRO, β_{24}	0.01	1.01	(0.996,1.033)			
ENGY, β_{25}	0.04	1.04	(0.915,1.186)			
For THOLD2, δ_2	1.80***	6.05	(5.232,7.007)	1.80***	6.05	(5.232,7.007)
For THOLD3, δ_3	3.99***	53.87	(45.826,63.315)	3.99***	53.86	(45.824,63.317)
For THOLD4, δ_4	6.58***	717.21	(606.426,848.222)	6.57***	716.84	(606.107,847.806)
For THOLD5, δ_5	8.80***	6623.61	(5579.170,7863.576)	8.80***	6618.35	(5574.672,7857.411)

	Survey			SurveySig		
	Coefficient	Odds	C.I.	Coefficient	Odds	C.I.
Final estimation of variance components						
			Random Effect			
	Variance (s.d.)	d.f.	χ^2	Variance (s.d.)	d.f.	χ^2
INTRCPT1, r_0	4.10*** (2.03)	8230	14889.28536	4.10*** (2.03)	8231	14890.35649
SLOPE12						
slope, r_1	0.60*** (0.78)	8235	8628.7225	0.61*** (0.78)	8238	8634.89429

Note: Standard deviation in parentheses for estimated variances. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

The next stage of models considered the language distance measures. As in the above studies, the distance measures, *FSI, GateRev, TypeRev, NotIndo, NotLatin,* and *DLI* were modeled separately. The models all required an increase in the default number of estimation settings to converge, and the models' level-1 random coefficient reliability estimate remained marginal in a range of 0.098 to 0.105. Reduced output from these five models is displayed in Table 37. The main effect for the intercept at the first threshold was similar to all other models, with an estimated coefficient of either -6.44 or -6.43, depending on the measure. The two slopes were consistent with earlier models, showing the greater likelihood of scores dropping between the first two test occasions and rising thereafter. For the intercept, in addition to the *NotLatin* measure, the *GateRev* measure was also not significant; for the first slope, the results were identical (all but the *NotIndo* measure were significantly related to the first slope), and for the second slope, there were two measures found to be significant predictors of the second slope (*NotLatin* and *DLI*).

The direction of the estimates also provided information about the effect of language distance. For the significant measures for the intercept, *FSI, NotIndo* and *DLI* were all positive, which meant that all things being equal, individuals in the harder languages had a greater probability of being at or below a threshold. *TypeRev* was in the opposite direction, i.e., negative. All of the significant predictors of the either slope, including *TypeRev*, were positive, confirming the hypothesis that language distance would constrain growth. The addition of the distance measures effected the estimated coefficients and their significance in several cases, most notably in the *DLAB-Slope12* terms: in the FSI, NotLatin and DLI models, the

*DLABPt3*Slope12* coefficient dropped out of significance, while in *GateRev* it was marginal ($p < .05$); in the FSI and DLI models, *DLABPt4* Slope12* also fell out of significance, while in NotIndo model it was marginal. In the two models with continuous distance measures, *GateRev* or *TypeRev*, the two DLAB subtest estimates remained significant in their effect on the first slope. The *DLABPt3*Slope26* estimated coefficient, which had been significant ($p = .03$) in the earlier models, dropped out of significance for the FSI, GateRev, NotLatin and DLI models. The FSI variable was selected to take forward to the next stage in model in order to mirror the earlier studies.

Table 37

HGLM listening language distance models

	FSI		GateRev		TypeRev	
	Coeff	Odds	Coeff	Odds	Coeff	Odds
Fixed effect						
For INTRCPT1 slope, π_0						
INTRCPT2, β_{00}	-6.44***	0.00	-6.43***	0.00	-6.43***	0.00
ZA_AR, β_{01}	-0.11***	0.89	-0.11***	0.90	-0.10***	0.90
ZA_AS, β_{02}	0.07*	1.07	0.07**	1.07	0.06*	1.07
ZA_GS, β_{03}	-0.10***	0.91	-0.10***	0.91	-0.10***	0.91
ZA_MK, β_{04}	-0.16***	0.86	-0.15***	0.86	-0.14***	0.87
ZA_PC, β_{05}	-0.19***	0.82	-0.19***	0.83	-0.18***	0.83
ZA_WK, β_{06}	-0.27***	0.77	-0.26***	0.77	-0.26***	0.77
ZD_1, β_{07}	-0.19***	0.83	-0.17***	0.84	-0.16***	0.85
ZD_2, β_{08}	-0.14***	0.87	-0.12***	0.89	-0.10***	0.90
ZD_3, β_{09}	-0.28***	0.76	-0.22***	0.80	-0.18***	0.83
EDUC, β_{010}	0.10***	1.11	0.10***	1.11	0.11***	1.11
PRIORPRO, β_{011}	-0.13***	0.88	-0.12***	0.88	-0.13***	0.88
ENGY, β_{012}	0.78***	2.18	0.75***	2.12	0.73***	2.08
[langdist], β_{013}	0.16***	1.18	0.00	1.00	-0.02***	0.98
For SLOPE12 slope, π_1						

	FSI		GateRev		TypeRev	
INTRCPT2, β_{10}	0.19***	1.21	0.19***	1.21	0.19***	1.21
ZA_MC, β_{11}	0.10***	1.11	0.10***	1.10	0.09***	1.10
ZD_1, β_{12}	-0.09***	0.91	-0.09***	0.91	-0.09***	0.91
ZD_3, β_{13}	0.05	1.05	0.06*	1.06	0.06*	1.06
ZD_4, β_{14}	0.03	1.03	0.05*	1.05	0.07***	1.07
FSILANGC, β_{15}	0.06***	1.07	0.01**	1.01	0.01***	1.01
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.27***	0.76	-0.27***	0.76	-0.27***	0.76
ZD_3, β_{21}	0.02	1.02	0.02	1.02	0.02*	1.02
FSILANGC, β_{22}	0.01*	1.01	0.00	1.00	0.00	1.00
For THOLD2,						
δ_2	1.81***	6.08	1.80***	6.07	1.80***	6.04
For THOLD3,						
δ_3	3.99***	54.22	3.99***	54.02	3.98***	53.60
For THOLD4,						
δ_4	6.58***	723.44	6.58***	719.58	6.57***	714.21
For THOLD5,						
δ_5	8.81***	6691.08	8.80***	6649.67	8.80***	6604.21
Final estimation of variance components						
Random Effect	Variance		Variance		Variance	
INTRCPT1, r_0	4.06***		4.11***		4.07***	
SLOPE12 slope, r_1	0.59***		0.60***		0.60***	

	NotLatin		NotIndo		DLI	
	Coeff	Odds	Coeff	Odds	Coeff	Odds
Fixed effect						
For INTRCPT1 slope, π_0						
INTRCPT2, β_{00}	-6.43***	0.00	-6.43***	0.00	-6.44***	0.00
ZA_AR, β_{01}	-0.11***	0.90	-0.12***	0.89	-0.11***	0.90
ZA_AS, β_{02}	0.07*	1.07	0.06*	1.07	0.07*	1.07
ZA_GS, β_{03}	-0.10***	0.91	-0.10***	0.90	-0.10***	0.91
ZA_MK, β_{04}	-0.15***	0.86	-0.15***	0.86	-0.15***	0.86
ZA_PC, β_{05}	-0.18***	0.83	-0.19***	0.82	-0.19***	0.83
ZA_WK, β_{06}	-0.26***	0.77	-0.27***	0.77	-0.26***	0.77
ZD_1, β_{07}	-0.17***	0.85	-0.18***	0.83	-0.17***	0.84
ZD_2, β_{08}	-0.12***	0.89	-0.14***	0.87	-0.13***	0.88
ZD_3, β_{09}	-0.21***	0.81	-0.27***	0.77	-0.22***	0.80
EDUC, β_{010}	0.10***	1.11	0.10***	1.11	0.10***	1.11
PRIORPRO, β_{011}	-0.12***	0.89	-0.14***	0.87	-0.12***	0.88
ENGY, β_{012}	0.76***	2.15	0.74***	2.09	0.76***	2.14
[langdist], β_{013}	0.01	1.01	0.53***	1.71	0.08*	1.08
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	0.19***	1.21	0.19***	1.21	0.19***	1.21
ZA_MC, β_{11}	0.10***	1.10	0.10***	1.10	0.10***	1.10
ZD_1, β_{12}	-0.09***	0.91	-0.08***	0.92	-0.10***	0.91
ZD_3, β_{13}	0.05	1.05	0.07**	1.08	0.04	1.04

	NotLatin		NotIndo		DLI	
ZD_4, β_{14}	0.05*	1.05	0.05*	1.05	0.03	1.04
[langdist], β_{15}	0.46**	1.59	0.06	1.06	0.15***	1.16
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.27***	0.76	-0.27***	0.76	-0.27***	0.76
ZD_3, β_{21}	0.02	1.02	0.02*	1.02	0.01	1.01
[langdist], β_{22}	0.09**	1.09	0.02	1.02	0.04***	1.04
For THOLD2,						
δ_2	1.80***	6.06	1.80***	6.07	1.80***	6.07
For THOLD3,						
δ_3	3.99***	53.95	3.99***	54.03	3.99***	54.01
For THOLD4,						
δ_4	6.58***	719.44	6.58***	720.48	6.58***	720.57
For THOLD5,						
δ_5	8.80***	6653.74	8.80***	6653.03	8.81***	6668.67
Final estimation of variance components						
Random Effect	Variance		Variance		Variance	
INTRCPT1, r_0	4.10***		4.04***		4.11***	
SLOPE12 slope, r_1	0.57***		0.60***		0.57***	

Note:. All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

The final stage of modeling investigated whether there were any significant interactions for aptitude subtests and language distance in the ordinal model of listening. In addition to the interaction terms, the main effects of the variables were also modeled even if their *p*-values were larger than 0.05 in earlier models. The FSI measure was selected for this model, and as with the first two studies, FSI was coded as a dichotomous variable. A finding of significant interaction would mean that the relationship between the scores on an aptitude subtest and either the intercept or slopes was different in harder languages (*FSI0=1*) than in the easier languages (*FSI0=0*). Given the number of variables now in the model, corrections were made for multiple comparisons. The B-H approach was applied to this final model to reduce the false discovery rate, and the last two columns in Table 38 below display the p-value as reported in the model and whether the p-value is below the B-H critical value. Adding the interaction terms complicated the interpretation of the estimates. With the interaction terms, the effect of the aptitude predictor variables on the outcome is different for different values of language distance, all other things being equal.

Table 38

HGLM Listening with FSI0 interactions model

Fixed Effect	Coefficient	Odds	*p*-value	p < BH
For INTRCPT1 slope, π_0)				
INTRCPT2, β_{00}	-6.44***	0.00	<0.001	†
ZA_AO, β_{01}	0.05	1.05	0.31	
ZA_AR, β_{02}	-0.13*	0.88	0.015	
ZA_AS, β_{03}	-0.07	0.94	0.23	
ZA_EI, β_{04}	0.03	1.03	0.653	
ZA_GS, β_{05}	-0.08	0.92	0.126	

Fixed Effect	Coefficient	Odds	*p*-value	p < BH
ZA_MC, β_{06}	0.00	1.00	0.968	
ZA_MK, β_{07}	-0.11	0.90	0.035	
ZA_PC, β_{08}	-0.22***	0.80	<0.001	†
ZA_WK, β_{09}	-0.25***	0.78	<0.001	†
ZD_1, β_{010}	-0.19***	0.83	<0.001	†
ZD_2, β_{011}	-0.11***	0.90	0.002	†
ZD_3, β_{012}	-0.38***	0.68	<0.001	†
ZD_4, β_{013}	-0.17***	0.84	<0.001	†
EDUC, β_{014}	0.10***	1.11	<0.001	†
PRIORPRO, β_{015}	-0.14***	0.87	<0.001	†
ENGY, β_{016}	0.74***	2.09	<0.001	
FSI0, β_{017}	0.81***	2.25	<0.001	†
AO_FSI0, β_{018}	0.06	1.06	0.344	
AR_FSI0, β_{019}	-0.01	0.99	0.921	
AS_FSI0, β_{020}	0.20**	1.22	0.009	
EI_FSI0, β_{021}	-0.03	0.98	0.759	
GS_FSI0, β_{022}	-0.06	0.94	0.453	
MC_FSI0, β_{023}	0.08	1.08	0.361	
MK_FSI0, β_{024}	-0.04	0.96	0.589	
PC_FSI0, β_{025}	0.01	1.01	0.859	
WK_FSI0, β_{026}	0.02	1.02	0.807	
D1_FSI0, β_{027}	0.00	1.00	0.972	
D2_FSI0, β_{028}	-0.15**	0.86	0.007	
D3_FSI0, β_{029}	0.17**	1.18	0.011	
D4_FSI0, β_{030}	0.11	1.11	0.105	
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	0.19***	1.21	<0.001	†
ZA_AO, β_{11}	0.04	1.04	0.365	
ZA_AR, β_{12}	-0.06	0.94	0.242	
ZA_AS, β_{13}	0.01	1.01	0.787	
ZA_EI, β_{14}	-0.06	0.94	0.275	
ZA_GS, β_{15}	0.02	1.02	0.679	
ZA_MC, β_{16}	0.16**	1.18	0.006	
ZA_MK, β_{17}	-0.02	0.98	0.708	
ZA_PC, β_{18}	0.06	1.06	0.211	
ZA_WK, β_{19}	-0.11*	0.89	0.021	
ZD_1, β_{110}	-0.03	0.97	0.441	

Fixed Effect	Coefficient	Odds	*p*-value	p < BH
ZD_3, β_{111}	0.13***	1.14	0.002	†
ZD_4, β_{112}	0.15***	1.17	<0.001	†
FSI0, β_{113}	-0.01	0.99	0.911	
AO_FSI0, β_{114}	-0.01	0.99	0.817	
AR_FSI0, β_{115}	0.11	1.12	0.129	
AS_FSI0, β_{116}	-0.01	0.99	0.921	
EI_FSI0, β_{117}	0.05	1.05	0.549	
GS_FSI0, β_{118}	-0.02	0.98	0.788	
MC_FSI0, β_{119}	-0.14	0.87	0.105	
MK_FSI0, β_{120}	-0.02	0.98	0.814	
PC_FSI0, β_{121}	-0.02	0.98	0.747	
WK_FSI0, β_{122}	0.16*	1.18	0.022	
D1_FSI0, β_{123}	-0.10	0.91	0.085	
D2_FSI0, β_{124}	0.04	1.04	0.391	
D3_FSI0, β_{125}	-0.13*	0.88	0.038	
D4_FSI0, β_{126}	-0.12*	0.88	0.054	
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	-0.27***	0.76	<0.001	†
ZA_AO, β_{21}	0.01	1.01	0.662	
ZA_AR, β_{22}	0.01	1.01	0.606	
ZA_AS, β_{23}	-0.03	0.98	0.163	
ZA_EI, β_{24}	0.00	1.00	0.947	
ZA_GS, β_{25}	0.02	1.02	0.411	
ZA_MC, β_{26}	-0.02	0.98	0.405	
ZA_MK, β_{27}	-0.01	0.99	0.477	
ZA_PC, β_{28}	0.01	1.01	0.712	
ZA_WK, β_{29}	0.01	1.01	0.411	
ZD_1, β_{210}	0.01	1.01	0.655	
ZD_2, β_{211}	0.01	1.01	0.625	
ZD_3, β_{212}	0.01	1.01	0.427	
ZD_4, β_{213}	-0.02	0.98	0.231	
FSI0, β_{214}	0.02	1.02	0.388	
AO_FSI0, β_{215}	-0.01	0.99	0.554	
AR_FSI0, β_{216}	0.01	1.01	0.602	
AS_FSI0, β_{217}	0.04	1.04	0.116	
EI_FSI0, β_{218}	-0.01	0.99	0.651	
GS_FSI0, β_{219}	0.00	1.00	0.977	

Fixed Effect	Coefficient	Odds	*p*-value	p < BH
MC_FSI0, β_{220}	0.00	1.00	0.978	
MK_FSI0, β_{221}	-0.01	0.99	0.718	
PC_FSI0, β_{222}	-0.02	0.98	0.323	
WK_FSI0, β_{223}	-0.01	0.99	0.578	
D1_FSI0, β_{224}	-0.01	0.99	0.569	
D2_FSI0, β_{225}	0.00	1.00	0.932	
D3_FSI0, β_{226}	0.02	1.02	0.375	
D4_FSI0, β_{227}	0.03	1.03	0.238	
For THOLD2, δ_2	1.81***	6.08	<0.001	†
For THOLD3, δ_3	3.99***	54.18	<0.001	†
For THOLD4, δ_4	6.59***	724.54	<0.001	†
For THOLD5, δ_5	8.81***	6717.22	<0.001	†
Random Effect				
	Variance	d.f.	$\chi 2$	p-value
INTRCPT1, r0	3.95 (1.99)	8213	14700.4623	<0.001
SLOPE12 slope, r1	0.58 (0.76)	8217	8596.71833	0.002

Note: All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05 † p < BH critical

A number of estimates dropped out of significance in this model. With the FSI*aptitude interaction terms added to the model, the only remaining significant effects in the model for the intercept were *ASVAB-PC, ASVAB-WK*, and all four DLAB subtests. These estimates were all negative, which was meant that all else being equal, those with higher scores on each subtest were more likely to be at higher proficiency levels at the time of graduation, as expected. The three survey variables in the model for the intercepts were also still significant. The *DLABPt3* and *DLABPt4* as predictors of the first slope, *Slope12,* and the direction of the estimates indicated that those with higher aptitude scores were more likely to have a lower scores between

the first two tests, which had the effect of steepening Slope12. None of the interaction terms themselves were significant in the model for the intercepts, the model for *Slope12* or the model for *Slope26*. The PRV for this final model's level-2 intercept was 0.14 and for the level-2 slope it was 0.12, indicating that this model did explain additional variance in mean proficiency level at the time of graduation and in mean initial growth. Significant unexplained variance remained in the model, but as this model responded to the final research question, no further exploration was conducted.

Study 4 Hierarchical Generalized Linear Modeling (Reading)

In the final study, HLM8 (Raudenbush & Congdon, 2021) was used once again to model the reading data with the outcome treated as an ordinal measure with six categories (Levels 0/0+, Level 1, Level 1+, Level 2, Level 2+ and Level 3). The first model was an empty random intercept model and the results for this model are displayed in Table 39.

Table 39

HGLM reading null model

	Coefficient	Odds Ratio	Confidence Intervals
	Fixed Effect		
For INTRCPT1 slope, π_0)			
INTRCPT2, β_{00}	-6.32***	0.00	(0.002, 0.002)
For THOLD2, δ_2	1.04***	2.83	(2.493, 3.216)
For THOLD3, δ_3	3.12***	22.68	(19.384, 26.533)
For THOLD4, δ_4	5.68***	292.22	(247.962, 344.368)
For THOLD5, δ_5	8.10***	3285.96	(2778.490, 3886.109)
	Random Effect		
INTRCPT1, r_0	3.69 (8.10)		

*** p < .001

The model produced the estimates for an individual being at or below each ILR level. In the null model for the reading data, the results showed that across all the individuals in the dataset, the expected log odds of being at Level 0/0+ was negative (–6.32, $p < .001$), meaning that it was more likely that individuals were at least ILR

Level 1 or higher. The analysis also showed significant variation between individuals in their intercepts (proficiency at graduation), $\tau_{00} = 3.693$ ($SE = 1.92$), $\chi^2 = 38828.94$, $p < .001$. The level-2 ICC was calculated for this model using the definition presented above as 0.53, which indicated that 53% of the variance was attributed to the individual, which is just slightly above the ICC calculated for reading in Study 2. Before moving on to examine how predictors might account for the heterogeneity between individuals, the longitudinal nature of the data was modeled with two slopes, *Slope12* (initial growth) and *Slope26* (subsequent growth).

The output for the random intercept model with two piecewise slopes (RISlopes) is reported below. Both slopes were significant, with a positive coefficient for the first slope and negative for the second. As explained in Study 3, the direction of the coefficients in the ordinal HGLM model is opposite that of the earlier HLM models, so these results indicated that the likelihood of falling test scores was higher between the first two test occasions, and rising test scores were seen thereafter. Overall, the fixed effect estimates for the intercept and thresholds changed only slightly from those in the Null Model, as can be seen in Table 40 below.

Table 40

HGLM reading random intercept, piecewise slopes model

Fixed Effect	Coefficient	Odds	Confidence
For INTRCPT1 slope, π_0			
INTRCPT2, β_{00}	-6.32***	0.00	(0.002,0.002)
For SLOPE12 slope, π_1			
INTRCPT2, β_{10}	0.22***	1.25	(1.180,1.314)
For SLOPE26 slope, π_2			
INTRCPT2, β_{20}	-0.31***	0.74	(0.721,0.751)
For THOLD2,			
δ_2	1.05***	2.85	(2.515,3.239)

Fixed Effect	Coefficient	Odds	Confidence
For THOLD3, δ_3	3.15***	23.38	(19.999,27.326)
For THOLD4, δ_4	5.76***	318.31	(270.273,374.878)
For THOLD5, δ_5	8.25***	3819.11	(3230.437,4515.044)
Final estimation of variance components			
	Random Effect		
INTRCPT1, r_0	3.90*** (1.97)		

*** p < .001

The next series of models tested the effects of allowing the slopes to vary randomly, first with *Slope12* random, then with *Slope26* random, then with both slopes random. Models with either slope random would not converge, even after raising the default iteration settings. Therefore, model building in reading continued with a random intercept and two fixed slopes.

The next set of models were designed to respond to RQ 3. The general aptitude subtests from ASVAB, and then language aptitude battery, DLAB, were added to the model as in the studies above. The ASVAB subtests were first all added to the model as predictors of the intercept and slopes, and the model output is in the first two columns of Table 41. The standardized scores from the subtests were centered by grand mean when added to the model. The estimates for the main effect of the intercept, *Slope12,* or *Slope26* did not fundamentally change. The results showed that of the nine ASVAB subtests, only the *ASVAB-AO* and *EI* subtests failed to reach significance as predictors of the intercept. The subtests with the largest effects on the intercept were -*PC* and -*WK*. Only *ASVAB-AO* was significantly related

to *Slope12* and only *ASVAB-WK* was a significant predictor of *Slope26.* The addition of the ASVAB subtests to the model reduced the variance by 16%.

Before moving on to add the DLAB subtests, an interim model was tested in which only the significant ASVAB subtests were included. This model's output is displayed in the third and fourth columns of the table below. Slight changes to the ASVAB subtest intercept estimates were observed, but the intercept, slopes and threshold estimates, as well as the variance estimates and significance, were unchanged.

Table 41

HGLM reading with ASVAB subtests

	ASVAB		ASVABsig	
	Coefficient	O.R	Coefficient	O.R.
For INTRCPT1 slope, π_0)		Fixed effect		
INTRCPT2, β_{00}	-6.35***	0.00	-6.34***	0.00
ZA_AO	-0.02	0.98		
ZA_AR	-0.22***	0.81	-0.21***	0.81
ZA_AS	0.13***	1.14	0.13***	1.14
ZA_EI	-0.02	0.98		
ZA_GS	-0.11***	0.89	-0.11***	0.89
ZA_MC	0.12***	1.13	0.14***	1.15
ZA_MK	-0.22***	0.80	-0.25***	0.78
ZA_PC	-0.29***	0.75	-0.28***	0.76
ZA_WK	-0.36***	0.69	-0.38***	0.68
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	0.22***	1.25	0.22***	1.25
ZA_AO	0.06*	1.06	0.06***	1.06
ZA_AR	0.00	1.00		
ZA_AS	0.03	1.03		
ZA_EI	0.00	1.00		
ZA_GS	0.03	1.03		
ZA_MC	0.07	1.08		

ZA_MK	-0.05	0.95		
ZA_PC	0.01	1.01		
ZA_WK	-0.03	0.97		
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	-0.31***	0.74	-0.31***	0.74
ZA_AO	-0.01	0.99		
ZA_AR	0.02	1.02		
ZA_AS	-0.01	0.99		
ZA_EI	-0.01	0.99		
ZA_GS	-0.01	0.99		
ZA_MC	-0.02	0.98		
ZA_MK	0.01	1.01		
ZA_PC	0.01	1.01		
ZA_WK	0.04***	1.04	0.03***	1.03
For THOLD2,				
δ_2	1.05***	2.85	1.05***	2.85
For THOLD3,				
δ_3	3.16***	23.47	3.15***	23.42
For THOLD4,				
δ_4	5.78***	323.69	5.78***	322.22
For THOLD5,				
δ_5	8.27***	3915.94	8.27***	3894.38

Final estimation of variance components

	Random Effect Var (s.d)
INTRCPT1, r_0	3.36*** (1.83)

Note: All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

Next, the DLAB subtest measures were added as predictors of intercept and slopes to the random intercept, piecewise slopes model with only the significant ASVAB subtests. The standardized DLAB subtests were grand mean centered when added to the model, and the output is presented in Table 42. All four of the DLAB subtests were significant predictors of the intercept, and only *DLABPt1* was significantly related to the first slope. None of the DLAB subtests had a significant

effect on the second slope. The logits for the subtests -*AR, -GS, -MK -PC and -WK* remained negative in their relationship with the intercept while *ASVAB-AS* and -*MC* were positive. The four DLAB subtests were all negative in relation to the intercept and *DLABPt3* had the largest effect. Language aptitude explained 3% of the variance in intercept between individuals. Before moving on to add the survey variables, a model was built to include only the significant aptitude subtests these results (DLABsig) are also shown below.

Table 42

HGLM reading with DLAB subtests

	DLAB		DLABsig	
	Coefficient	O.R.	Coefficient	O.R.
	Fixed Effect			
For INTRCPT1 slope, π_0				
INTRCPT2, β_{00}	-6.34***	0.00	-6.34***	0.00
ZA_AR	-0.15***	0.86	-0.16***	0.86
ZA_AS	0.10***	1.10	0.10***	1.10
ZA_GS	-0.11***	0.90	-0.10***	0.90
ZA_MC	0.15***	1.16	0.15***	1.16
ZA_MK	-0.19***	0.83	-0.19***	0.83
ZA_PC,	-0.25***	0.78	-0.25***	0.78
ZA_WK	-0.29***	0.75	-0.30***	0.74
ZD_1	-0.12***	0.89	-0.13***	0.88
ZD_2	-0.09***	0.91	-0.07***	0.93
ZD_3	-0.25***	0.78	-0.21***	0.81
ZD_4	-0.14***	0.87	-0.12***	0.88
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	0.22***	1.25	0.22***	1.25
ZA_AO	0.06***	1.06	0.08***	1.08
ZD_1	-0.09***	0.91	-0.07***	0.93
ZD_2	0.03	1.03		
ZD_3	0.05	1.06		
ZD_4	0.01	1.01		

	DLAB		DLABsig	
	Coefficient	O.R.	Coefficient	O.R.
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	-0.31***	0.74	-0.31***	0.73
ZA_WK	0.02***	1.02	0.03***	1.03
ZD_1	0.01	1.01		
ZD_2	0.00	1.00		
ZD_3	0.02	1.02		
ZD_4	0.01	1.01		
For THOLD2,				
δ_2	1.05***	2.85	1.05***	2.85
For THOLD3,				
δ_3	3.15***	23.31	3.15***	23.29
For THOLD4,				
δ_4	5.77***	321.14	5.77***	320.17
For THOLD5,				
δ_5	8.27***	3895.50	8.26***	3882.07
Final estimation of variance components				
	Random Effect (s.d.)			
INTRCPT1, r_0	3.26*** (1.80)		3.26*** (1.80)	

Note: All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

The four survey variables were next added (group centered) into the last aptitude model (*DLABsig*) as predictors of intercept and both slopes. Of the four survey variables, education, motivation, prior proficiency and English as a first language, only *ENGY* (0=Not English) was significantly related to the Intercept (β_{015} -0.55, p = .003). *EDUC* was significantly related to *Slope26*, with a positive estimate (β_{22} 0.02, p < .001), which indicated that contrary to expectations, individuals with higher levels of education and all other things being equal, individuals were more likely to be at a lower proficiency level in reading. While unexpected, these findings mirrored those in the earlier HLM study. The fixed effect coefficient estimates for the

aptitude variables shifted very slightly with the addition of the survey variables, either increasing or decreasing. The survey variables did not explain any additional level-2 variance in reading. A model with only the significant level-2 predictors resulted in the *Slope12/EDUCM* relationship dropping out of significance, and the final SurveySig model is shown below in columns 3 and 4.

Table 43

HGLM reading with survey variables

	Survey		SurveySig	
	Coefficient	OR	Coefficient	OR
	Fixed effect			
For INTRCPT1 slope, π_0				
INTRCPT2, β_{00}	-6.34***	0.00	-6.34***	0.00
ZA_AR	-0.16***	0.85	-0.16***	0.86
ZA_AS	0.09***	1.09	0.09***	1.10
ZA_GS	-0.10***	0.90	-0.11***	0.90
ZA_MC	0.15***	1.16	0.15***	1.16
ZA_MK	-0.19***	0.83	-0.19***	0.83
ZA_PC	-0.25***	0.78	-0.25***	0.78
ZA_WK	-0.31***	0.73	-0.31***	0.74
ZD_1	-0.14***	0.87	-0.13***	0.88
ZD_2	-0.07***	0.93	-0.07***	0.93
ZD_3	-0.21***	0.81	-0.21***	0.81
ZD_4	-0.12***	0.88	-0.12***	0.88
EDUC	0.03	1.03		
MOT	0.01	1.01		
PRIPROF	-0.02	0.98		
ENGY	0.55***	1.73	0.55***	1.74
For SLOPE12 slope, π_1				
INTRCPT2, β_{10}	0.22***	1.25	0.22***	1.25
ZA_AO	0.08***	1.08	0.08***	1.08
ZD_1	-0.06*	0.94	-0.09***	0.92
EDUC	-0.04*	0.96		
MOT	0.00	1.00		

	Survey		SurveySig	
	Coefficient	OR	Coefficient	OR
PRIPROF	-0.01	0.99		
ENGY	0.08	1.08		
For SLOPE26 slope, π_2				
INTRCPT2, β_{20}	-0.31***	0.73	-0.31***	0.73
ZA_WK	0.02*	1.02	0.02*	1.02
EDUC	0.02***	1.02	0.02***	1.02
MOT	0.01	1.01		
PRIPROF	-0.01	0.99		
ENGY	-0.10	0.90		
For THOLD2,				
δ_2	1.05***	2.85	1.05***	2.85
For THOLD3,				
δ_3	3.15***	23.32	3.15***	23.31
For THOLD4,				
δ_4	5.77***	321.10	5.77***	320.72
For THOLD5,				
δ_5	8.27***	3901.27	8.27***	3893.49
Final estimation of variance components				
	Random Effect (s.d)			
INTRCPT1, r_0	3.26*** (1.80)		3.26*** (1.80)	

Note: All level-2 variables centered on grand mean.

*** p < .005 ** p < .01 * p < .05

The next stage of models examined the influence of the language distance measures on reading language proficiency, while considering significant ASVAB, DLAB and other individual differences. As in the above studies, the distance measures, *FSI, GateRev, TypeRev, NotLatin, NotIndo*, and *DLI,* were modeled separately. The estimated coefficients for graduation (the intercept) and growth (two slopes) changed only slightly as compared to the SurveySig model above. The new

models accounting for language distance explained up to 1% of the intercept variance between individuals.

Reduced output from these five models is displayed in Table 44. *FSI*, *TypeRev*, and *NotIndo* all had significant effects on the intercept, though the direction of the estimates varied: *FSI* and *NotIndo* were both negative, while *TypeRev* was positive, as seen in the listening models. This meant that for *FSI* and *NotIndo*, those in harder languages were expected to be in lower ILR categories. The differing scales of the estimates complicated any comparison of effects, but as noted earlier in the other studies, none of the estimates of language distance were large relative to the coefficient for the intercept. All six of the distance measures were significant, positive predictors of the first slope, again meaning that those in the more difficult languages were more likely to show a decrease in initial growth. Five of the six distance measures (*FSI, GateRev, TypeRev, NotLatin* and *DLI*) were significantly related to subsequent growth, and the estimates for *Slope26* were positive, which as explained above, was interpreted to mean that all else equal, individuals in this population in the harder languages were more likely to have their growth constrained between the second and sixth tests as well. The effect of *ASVAB-WK* as a main effect of the second slope dropped out of significance in each of the six models when distance was added to the model. The FSI variable was selected to take forward to the next model, given that it was chosen in the earlier models for the exploration of interaction effects.

Table 44

HGLM reading with language distance measures

	FSI		GateRev		TypeRev	
	Coefficient	OR	Coefficient	OR	Coefficient	OR
			Fixed effect			
For INTRCPT1 slope, π_0)						
INTRCPT2, β_{00}	-6.36***	0.00	-6.36***	0.00	-6.36***	0.00
ZA_AR, β_{01}	-0.16***	0.86	-0.16***	0.85	-0.15***	0.86
ZA_AS, β_{02}	0.09***	1.09	0.09***	1.10	0.09***	1.10
ZA_GS, β_{03}	-0.11***	0.90	-0.11***	0.90	-0.11***	0.90
ZA_MC, β_{04}	0.16***	1.17	0.15***	1.16	0.15***	1.16
ZA_MK, β_{05}	-0.19***	0.83	-0.19***	0.83	-0.19***	0.83
ZA_PC, β_{06}	-0.25***	0.77	-0.25***	0.78	-0.25***	0.78
ZA_WK, β_{07}	-0.30***	0.74	-0.30***	0.74	-0.31***	0.74
ZD_1, β_{08}	-0.13***	0.88	-0.12***	0.89	-0.11***	0.89
ZD_2, β_{09}	-0.11***	0.90	-0.08***	0.92	-0.07***	0.93
ZD_3, β_{010}	-0.28***	0.75	-0.24***	0.79	-0.21***	0.81
ZD_4, β_{011}	-0.18***	0.84	-0.14***	0.87	-0.12***	0.88
ENGY, β_{012}	0.59***	1.80	0.56***	1.75	0.56***	1.74
[langdist], β_{013}	0.08***	1.09	0.00	1.00	-0.02***	0.98
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	0.23***	1.25	0.23***	1.25	0.23***	1.25
ZA_AO, β_{11}	0.06***	1.07	0.06***	1.06	0.07***	1.07

	FSI		GateRev		TypeRev	
	Coefficient	OR	Coefficient	OR	Coefficient	OR
ZD_1, β_{12}	-0.12***	0.89	-0.11***	0.89	-0.10***	0.90
[langdist] β_{13}	0.12***	1.13	0.02***	1.02	0.03***	1.03
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.31***	0.73	-0.31***	0.73	-0.31***	0.73
ZA_WK, β_{21}	0.01	1.01	0.02	1.02	0.02	1.02
EDUC, β_{22}	0.02***	1.02	0.02***	1.02	0.02***	1.02
[langdist], β_{23}	0.02***	1.02	0.00***	1.00	0.01***	1.01
For THOLD2,						
δ_2	1.05***	2.86	1.05***	2.85	1.05***	2.85
For THOLD3,						
δ_3	3.16***	23.55	3.16***	23.46	3.15***	23.43
For THOLD4,						
δ_4	5.79***	326.53	5.79***	325.77	5.79***	325.69
For THOLD5,						
δ_5	8.29***	3998.60	8.29***	3987.31	8.29***	3991.25

Final estimation of variance components			
	Random Effect		
INTRCPT1, r_0	3.23*** (1.80)	3.28*** (1.81)	3.29*** (1.81)

	NotLatin		NotIndo		DLI	
	Coefficient	OR	Coefficient	OR	Coefficient	OR
			Fixed Effect			
For INTRCPT1 slope, π_0)						
INTRCPT2, β_{00}	-6.36***	0.00	-6.35***	0.00	-6.36***	0.00
ZA_AR, β_{01}	-0.15***	0.86	-0.16***	0.85	-0.15***	0.86
ZA_AS, β_{02}	0.09***	1.10	0.09***	1.09	0.09***	1.10
ZA_GS, β_{03}	-0.11***	0.90	-0.11***	0.90	-0.11***	0.90
ZA_MC, β_{04}	0.15***	1.16	0.15***	1.16	0.15***	1.16
ZA_MK, β_{05}	-0.19***	0.83	-0.19***	0.83	-0.19***	0.83
ZA_PC, β_{06}	-0.25***	0.78	-0.25***	0.78	-0.25***	0.78
ZA_WK, β_{07}	-0.30***	0.74	-0.30***	0.74	-0.30***	0.74
ZD_1, β_{08}	-0.12***	0.89	-0.13***	0.88	-0.12***	0.89
ZD_2, β_{09}	-0.09***	0.92	-0.09***	0.92	-0.09***	0.91
ZD_3, β_{010}	-0.24***	0.79	-0.24***	0.78	-0.25***	0.78
ZD_4, β_{011}	-0.14***	0.87	-0.15***	0.86	-0.15***	0.86
ENGY, β_{012}	0.58***	1.79	0.56***	1.74	0.57***	1.77
[langdist], β_{013}	-0.01	0.99	0.13*	1.14	-0.04	0.96
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	0.22***	1.25	0.23***	1.26	0.23***	1.25
ZA_AO, β_{11}	0.07***	1.07	0.07***	1.07	0.06***	1.06
ZD_1, β_{12}	-0.11***	0.90	-0.10***	0.90	-0.12***	0.89
[langdist], β_{13}	0.68***	1.97	0.24***	1.27	0.24***	1.27

	NotLatin		NotIndo		DLI	
	Coefficient	OR	Coefficient	OR	Coefficient	OR
For SLOPE26 slope, π_2						
INTRCPT2, β_{20}	-0.31***	0.73	-0.31***	0.73	-0.31***	0.73
ZA_WK, β_{21}	0.01	1.01	0.02	1.02	0.01	1.01
EDUC, β_{22}	0.02***	1.02	0.02***	1.02	0.02***	1.02
NOTLATIN, β_{23}	0.12***	1.13	0.03	1.03	0.05***	1.05
For THOLD2,						
δ_2	1.05***	2.85	1.05***	2.86	1.05***	2.85
For THOLD3,						
δ_3	3.15***	23.40	3.16***	23.50	3.16***	23.49
For THOLD4,						
δ_4	5.78***	324.65	5.78***	324.58	5.79***	326.59
For THOLD5,						
δ_5	8.29***	3978.70	8.28***	3950.36	8.30***	4005.63
Final estimation of variance components						
Random effects						
INTRCPT1, r_0	3.26*** (1.81)		3.26*** (1.81)		3.27*** (1.81)	

Note: All level-2 variables centered on grand mean. *** p < .005 ** p < .01 * p < .05

The final stage of modeling investigated whether there were any interactions between language distance and aptitude. All the main effects for the ASVAB and DLAB variables were re-entered into the model along with their interaction terms. The B-H approach was applied to identify those estimates where the p-value was lower than the B-H critical value to reduce the false discovery rate. Even before the application of the correction, there were no significant interaction terms for either slope in reading. Two interaction terms for the Intercept, *ASVAB-MK*FSI0* and *DLABPt4*FSI0* had significant p-values ($p < .05$) that did not meet the criteria after the correction. The addition of the interaction terms also resulted in a number of the ASVAB subtests to fall out of significance as predictors of the intercept: *ASVAB-AS, ASVAB-GS, ASVAB-MC, and ASVAB-MK*.

With the full model with interactions, the level-2 intercept variance was reduced by 1% as compared to the FSI model, and by 2% as compared to the Survey model, which suggested that the interactions did explain some of the variance in intercept at the individual level.

Table 45

HGLM reading with aptitude-FSI0 distance interactions

	Coefficient Fixed Effect	*O.R.*	Confidence		P < BH
For INTRCPT1 slope, π_0)					
INTRCPT2, β_{00}	-6.37***		0.00	(0.001,0.002)	†
ZA_AO, β_{01}	0.03		1.03	(0.945,1.114)	
ZA_AR, β_{02}	-0.17***		0.84	(0.766,0.929)	†
ZA_AS, β_{03}	0.01		1.01	(0.916,1.121)	
ZA_EI, β_{04}	-0.05		0.95	(0.855,1.061)	
ZA_GS, β_{05}	-0.05		0.95	(0.863,1.052)	
ZA_MC, β_{06}	0.10		1.11	(0.993,1.234)	
ZA_MK, β_{07}	-0.09		0.91	(0.835,1.002)	
ZA_PC, β_{08}	-0.30***		0.74	(0.682,0.808)	†
ZA_WK, β_{09}	-0.34***		0.71	(0.646,0.780)	†
ZD_1, β_{010}	-0.16***		0.85	(0.791,0.923)	†
ZD_2, β_{011}	-0.12***		0.89	(0.824,0.957)	†
ZD_3, β_{012}	-0.33***		0.72	(0.659,0.778)	†
ZD_4, β_{013}	-0.24***		0.79	(0.724,0.854)	†
ENGY, β_{014}	0.56***		1.76	(1.300,2.381)	†
FSI0, β_{015}	0.48***		1.62	(1.445,1.813)	†
AO_FSI0, β_{016}	0.00		1.00	(0.892,1.122)	

	Coefficient Fixed Effect	*O.R.*	Confidence		P < BH
AR_FSI0, β_{017}	0.01		1.01	(0.878,1.158)	
AS_FSI0, β_{018}	0.13		1.14	(0.989,1.307)	
EI_FSI0, β_{019}	0.04		1.04	(0.898,1.208)	
GS_FSI0, β_{020}	-0.10		0.90	(0.782,1.041)	
MC_FSI0, β_{021}	0.06		1.06	(0.907,1.233)	
MK_FSI0, β_{022}	-0.13*		0.88	(0.773,1.000)	
PC_FSI0, β_{023}	0.07		1.07	(0.947,1.208)	
WK_FSI0, β_{024}	0.11		1.11	(0.975,1.273)	
D1_FSI0, β_{025}	0.02		1.02	(0.916,1.135)	
D2_FSI0, β_{026}	-0.01		0.99	(0.890,1.104)	
D3_FSI0, β_{027}	0.02		1.02	(0.908,1.155)	
D4_FSI0, β_{028}	0.13*		1.14	(1.013,1.284)	
For SLOPE12 slope, π_1					
INTRCPT2, β_{10}	0.23***		1.26	(1.190,1.326)	†
ZA_AO, β_{11}	0.03		1.03	(0.954,1.122)	
ZA_AR, β_{12}	-0.01		0.99	(0.900,1.090)	
ZA_AS, β_{13}	0.02		1.02	(0.923,1.126)	
ZA_EI, β_{14}	-0.03		0.97	(0.873,1.086)	
ZA_GS, β_{15}	0.01		1.01	(0.911,1.114)	
ZA_MC, β_{16}	0.05		1.05	(0.944,1.171)	
ZA_MK, β_{17}	-0.05		0.95	(0.865,1.040)	

	Coefficient Fixed Effect	*O.R.*	Confidence		P < BH
ZA_PC, β_{18}	0.03		1.03	(0.945,1.113)	
ZA_WK, β_{19}	-0.04		0.96	(0.879,1.058)	
ZD_1, β_{110}	-0.07		0.93	(0.865,1.010)	
ZD_2, β_{111}	0.02		1.02	(0.947,1.104)	
ZD_3, β_{112}	0.07		1.07	(0.983,1.159)	
ZD_4, β_{113}	0.04		1.05	(0.963,1.135)	
FSI0, β_{114}	0.17***		1.18	(1.050,1.331)	
AO_FSI0, β_{115}	0.04		1.04	(0.919,1.169)	
AR_FSI0, β_{116}	0.01		1.01	(0.876,1.168)	
AS_FSI0, β_{117}	0.04		1.04	(0.899,1.200)	
EI_FSI0, β_{118}	0.05		1.05	(0.900,1.231)	
GS_FSI0, β_{119}	0.03		1.03	(0.886,1.191)	
MC_FSI0, β_{120}	0.03		1.03	(0.875,1.202)	
MK_FSI0, β_{121}	-0.01		0.99	(0.863,1.132)	
PC_FSI0, β_{122}	-0.04		0.96	(0.845,1.089)	
WK_FSI0, β_{123}	0.01		1.01	(0.882,1.163)	
D1_FSI0, β_{124}	-0.05		0.95	(0.853,1.064)	
D2_FSI0, β_{125}	0.00		1.00	(0.888,1.117)	
D3_FSI0, β_{126}	-0.05		0.95	(0.837,1.077)	
D4_FSI0, β_{127}	-0.11		0.90	(0.793,1.016)	
For SLOPE26 slope, π_2					

	Coefficient Fixed Effect	*O.R.*	Confidence		P < BH
INTRCPT2, β_{20}	-0.31***		0.74	(0.720,0.751)	†
ZA_AO, β_{21}	-0.01		0.99	(0.958,1.019)	
ZA_AR, β_{22}	0.01		1.01	(0.971,1.043)	
ZA_AS, β_{23}	-0.02		0.98	(0.948,1.020)	
ZA_EI, β_{24}	0.01		1.01	(0.965,1.049)	
ZA_GS, β_{25}	-0.02		0.98	(0.941,1.018)	
ZA_MC, β_{26}	-0.02		0.98	(0.939,1.022)	
ZA_MK, β_{27}	0.01		1.01	(0.974,1.045)	
ZA_PC, β_{28}	0.01		1.01	(0.977,1.040)	
ZA_WK, β_{29}	0.04*		1.04	(1.004,1.078)	
ZD_1, β_{210}	0.00		1.00	(0.969,1.028)	
ZD_2, β_{211}	0.02		1.02	(0.990,1.048)	
ZD_3, β_{212}	0.02		1.02	(0.985,1.052)	
ZD_4, β_{213}	0.00		1.00	(0.974,1.034)	
EDUC, β_{214}	0.02***		1.02	(1.005,1.027)	†
FSI0, β_{215}	0.01		1.01	(0.967,1.059)	
AO_FSI0, β_{216}	0.01		1.01	(0.964,1.055)	
AR_FSI0, β_{217}	0.02		1.02	(0.967,1.075)	
AS_FSI0, β_{218}	0.01		1.01	(0.959,1.068)	
EI_FSI0, β_{219}	-0.03		0.97	(0.918,1.032)	
GS_FSI0, β_{220}	0.02		1.02	(0.965,1.079)	

	Coefficient	*O.R.*	Confidence	P < BH
	Fixed Effect			
MC_FSI0, β_{221}	0.00	1.00	(0.945,1.068)	
MK_FSI0, β_{222}	-0.01	0.99	(0.943,1.045)	
PC_FSI0, β_{223}	0.01	1.01	(0.960,1.057)	
WK_FSI0, β_{224}	-0.03	0.97	(0.917,1.019)	
D1_FSI0, β_{225}	-0.01	0.99	(0.954,1.036)	
D2_FSI0, β_{226}	-0.03	0.97	(0.928,1.009)	
D3_FSI0, β_{227}	-0.01	0.99	(0.945,1.039)	
D4_FSI0, β_{228}	0.01	1.01	(0.964,1.055)	
For THOLD2,				
δ_2	1.05***	2.86	(2.522,3.245)	†
For THOLD3,				
δ_3	3.16***	23.65	(20.242,27.636)	†
For THOLD4,				
δ_4	5.79***	327.89	(278.410,386.156)	†
For THOLD5,				
δ_5	8.30***	4004.60	(3385.464,4736.962)	†
Final estimation of variance components		Random effects		
INTRCPT1, r_0	3.20*** (1.79)			

Note: All level-2 variables centered on grand mean.
*** p < .005 ** p < .01 * p < .05 † *p*-value < B-H critical value

Summary

There was very little growth in language proficiency in either listening or reading. Language proficiency followed a drop-and-recover pattern, but in general, scores were between ILR Level 2 and ILR Level 2+. ILR levels dropped slightly between the first two test scores, and then rose slightly in subsequent testing, and a piecewise approach with two slopes was used to model the data. The first slope, representing initial growth, described the slope between the first two test occasions, and the second slope, representing subsequent growth, described the slope from the second test occasion through the sixth. Three of the aptitude subtests, *ASVAB-MC*, *DLAB Part 3* and *DLAB Part 4*, were found to have a significant, negative relationship with initial growth in listening, regardless of how the ILR scale was treated (continuous or ordinal). No other significant relationships between aptitude (general or language aptitude) and growth were found.

The studies did confirm earlier findings in which general and language aptitude predicted language proficiency at the time of graduation. Higher scores on the two verbal subtests *ASVAB-PC* and *ASVAB-WK*, as well as higher scores on all four parts of the DLAB, predicted higher ILR proficiency in listening and reading at the time of graduation. In addition to these six subtests, the *ASVAB-AR* subtest was also found to significantly predict ILR levels in reading at the time of graduation.

Language distance constrained proficiency levels at the time of graduation, as well as constraining growth. There was no effect of an interaction of aptitude and distance on proficiency. **Table 46** summarizes the main findings of this research, and

Table 47 explains in more detail which language distance measures were significant predictors of graduation and growth.

Table 46

Summary of findings

	HLM Listening (Study 1)	HGLM Listening (Study 3)
Graduation (Intercept)	ASVAB-PC, -WK DLAB Part 1, Part 2, Part 3, Part 4 Language distance	ASVAB-PC, -WK DLAB Part 1, Part 2, Part 3, Part 4 Language distance
Growth (Slope12)	ASVAB-MC Language distance	ASVAB-MC Language distance
Subsequent growth (Slope26)	Language distance	Language distance
	HLM Reading (Study 2)	**HGLM Reading (Study 4)**
Graduation (Intercept)	ASVAB-AR, -PC, -WK DLAB Part 1, Part 3, Part 4 Language distance	ASVAB-PC, -WK DLAB Part 1, Part 2, Part 3, Part 4 Language distance
Growth (Slope12)	Language distance	Language distance
Subsequent growth (Slope26)	Language distance	Language distance

Table 47

Summary of significant language distance effects

	HLM Listening	HGLM Listening	HLM Reading	HGLM Reading
Graduation	FSI, TypeRev, NotIndo, DLI	FSI, TypeRev, NotIndo, DLI	FSI, TypeRev, Indo	FSI, TypeRev, Indo
Growth	FSI, GateRev, TypeRev, NotLatin, DLI	FSI, GateRev, TypeRev, NotLatin, DLI	FSI, GateRev, TypeRev, NotLatin, NotIndo, DLI	FSI, GateRev, TypeRev, NotLatin, NotIndo, DLI
Subsequent growth	GateRev, TypeRev, NotLatin, DLI	FSI, NotLatin, DLI	FSI, GateRev, TypeRev, NotLatin, DLI	FSI, GateRev, TypeRev, NotLatin, DLI

Chapter 7: Discussion

Research questions

This research looked beyond findings in earlier studies that cognitive abilities (general and language aptitude) predicted training outcomes to consider how these aptitude measures were related to growth after graduation. It also explored how language difficulty moderated these relationships. General linear mixed models and generalized linear mixed models were used to examine predictors of language proficiency growth in listening and reading. This chapter reviews the findings across all four studies in the context of the five research questions and compares the results by modality and methodology.

The first two research questions explored the longitudinal nature of the datasets. The first question examined to what extent there was variation in language proficiency growth after training, across individuals and languages. Because this is generally the case for adult language learning, it was hypothesized that there would be significant variation in language proficiency growth. "Growth" in these studies indicated both increases and decreases in ILR level. This hypothesis was partially supported, as significant variation was found in initial growth (*Slope12*) in three of the four studies. Later modeling in this research explored which aptitude variables explained that variation. The lack of support for variation in subsequent growth (*Slope26*) may have been due to the lack of variability in scores, or it may have been that using two piecewise slopes to describe the small amount of change resulted in convergence problems in the models.

The second part of the first research question asked to what extent there was variation by language. How individuals' scores varied by language was not tested directly (i.e., in a three-level model) due to the small number of languages in the study. In a three-level model, variation would have been partitioned by level, and the specific variation attributed to the nesting by language captured separately. In the two-level model used in this research, the variation was collapsed into level-2, and it was not possible to partition the variation by individual and by language.

It is worth noting at the outset of this chapter that there was actually very little growth over time in either listening or reading. While the two slopes, one marking initial growth (*Slope12*) and the other marking subsequent growth (*Slope26)*, were statistically significant, it was possibly due to the large sample size rather than to any meaningful growth.

Although not the focus of this research question, there was also significant variation found around the proficiency level at the time of graduation (i.e., the Intercept). This means that individuals' graduation scores varied around the mean, with some scores higher and some scores lower. The predictor variables included in this research explained some of that variation, but significant unexplained variance in the mean score at the time of graduation remained in the final model. This indicates that future research should address other possible predictors.

The second research question considered the shape of language proficiency growth and how it varied by language. A series of unconditional longitudinal models was run to consider the effects of time. Polynomial models of change with linear, linear and quadratic, and linear, quadratic and cubic polynomials were compared.

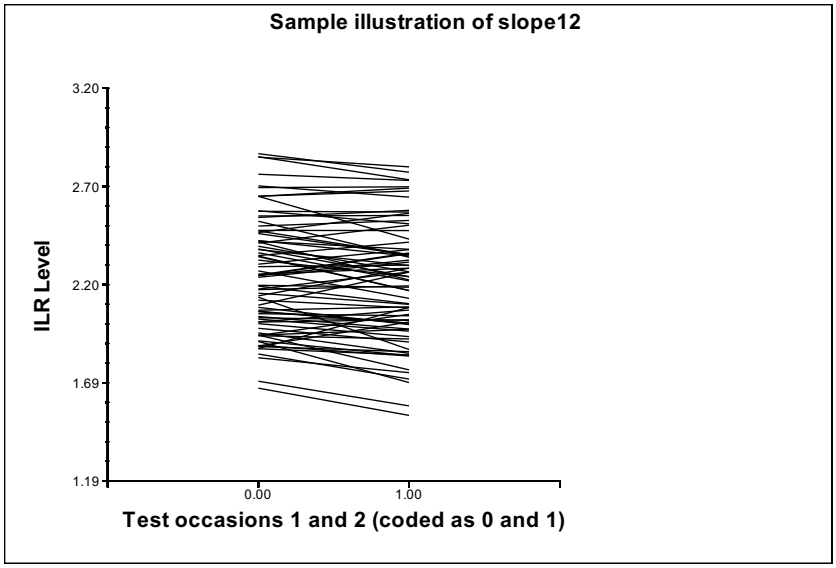
Given the drop-and-recover patterns reported in earlier studies, piecewise models with two slopes were also compared. The best fitting model was determined to be a piecewise slopes model, with two slopes, one representing initial growth (*Slope12*) and one representing subsequent growth (*Slope26*). Initial growth captured changes in proficiency between the first two tests, and subsequent growth represented changes from the second to the sixth testing occasion.

In the models for both listening and reading which treated the ILR scale as a continuous variable (Study 1 and Study 2), the estimated coefficient for the fixed effects for initial growth showed a very slight drop in proficiency. Listening scores dropped by 0.03 of an ILR level ($p < .001$) between the first two tests, while reading scores fell by 0.04 of an ILR level ($p < .001$). Subsequently, listening scores rose 0.05 of an ILR level per test occasion, while reading scores increased by 0.06 of an ILR level, both estimates significant at $p < .001$.

To better illustrate the meaning of the growth estimates for the two slopes, two figures are presented below. While drawn from real data, these figures are for illustrative purposes only as they represent only a small portion of the population. In Figure 5, each individual in this sample has their own slope between the first two test occasions (coded as Time 0 and Time 1). Some slopes are flat, others rise or fall, and the slope was allowed to vary randomly in the model to capture the variation around the mean. There is an overall downward trend seen, which would be represented in the model by a negative estimate for the mean slope, as was the case in these studies for initial growth (*Slope12*).

Figure 5

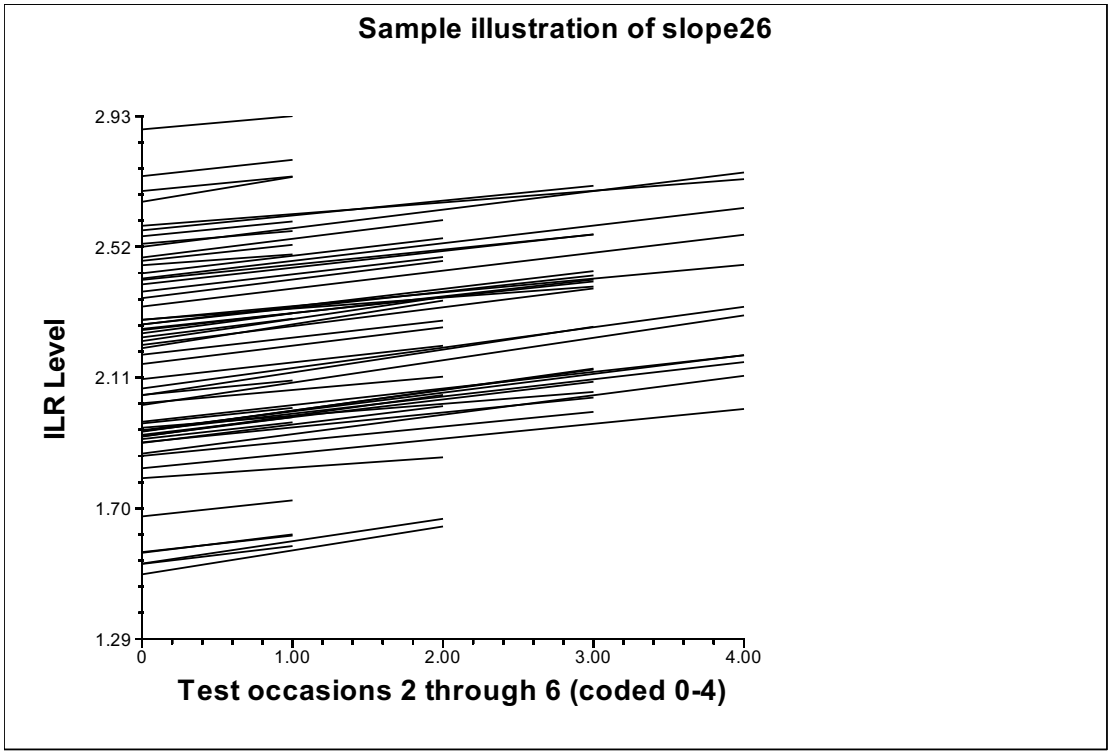
An illustration of test scores at the first and second occasions (by individual)



The second graph (Figure 6) illustrates subsequent growth (test scores between the second and sixth test occasion, or *Slope26*). There are up to four test occasions in this sample, as illustrated by the length of the slope along the x-axis. The slopes are almost parallel, indicating little to no variation in the slope, and the mean slope has a slight upward trend. These two features are represented in the estimates for this model, with the fixed effect of the slope for subsequent growth slightly more than zero and positive, while the random effect was non-significant. The second slope was fixed in future models due to this lack of variance around the mean.

Figure 6

An illustration of test scores at the second to six test occasions



Similar patterns were seen in growth for both modalities in the two studies (Study 3 and Study 4) which modeled the ILR scale as an ordinal variable, with individuals found, on average, to be more likely to have lower scores between the first two tests and higher scores thereafter. In the listening study (Study 3), the first slope was allowed to vary, and the second slope was fixed; in the reading study both slopes were fixed.

The findings that initial growth was characterized by a decrease, and subsequent growth by an increase in proficiency can perhaps be explained by the context in which the first two tests are taken. The first test is the final hurdle for graduation, and without a successful test score (ILR Level 2), the DLI student fails

and is recycled into an "easier" language, is given more time in training in their current language of study or is released to another position in the military. There is considerable pressure on students to "pass" the DLPT—their course of language study is designed to help them be successful, practice tests are taken, and emphasis is placed on graduation. It is then perhaps not surprising that when they take their second test a year later, out of this intensive language training environment, that their scores are lower. Thus, the first test score may reflect the emphasis placed on the DLPT in language training, and the second test, approximately a year later, might be considered as the more realistic estimate of the language level. After the second test, a variety of influences such as on-going language training, job assignments, foreign language incentive pay, and/or the desire to be promoted, differentially impact the trajectories.

The explanation for the drop-and-recover pattern could also be due to the selection process that was mentioned above. As explained in the first chapter, the majority of DLI graduates attend job-specific training at Advanced Intelligence Training for up to three months before they report to their duty station. While the job-specific training and job assignments involve the foreign language, individuals are no longer focused solely on general language proficiency, as it is measured by the DLPT. While initial growth and subsequent growth were statistically significant in this research, the trajectories were, in practical terms, almost flat (estimates for the slopes in the first two studies ranged from ± 0.03 to 0.05 of an ILR level). The subsequent growth indicated by the slope estimates of 0.03 or 0.05 of an ILR level was quite small, even over a few test occasions.

There may, in fact, be growth in language that is not captured by the DLPT, as the language used on the job likely differs from the more general form of the language that is tested on the DLPT. New performance-based measures may be needed to measure language skills on-the-job, either instead of, or in addition to the DLPT. As one of the early members of the ILR testing committee who trained DLPT test developers explained,

> "Proficiency is essentially what a language user can do with (for present purposes) a second language he or she has acquired or learned, without reference to a particular task or mission. Performance, on the other hand, is driven by mission requirements calling for the use of language for special purposes. The two are not mutually exclusive…" (Child, 1998a, p. 390)

If the language needed to perform on the job underlaps general proficiency to a great extent, it could be that language professionals do experience growth in language that is not assessed by the DLPT. Further research is necessary to better understand these possible differences between language proficiency and performance, and whether using the DLPT to measure language proficiency growth is appropriate.

This research adds to the number of longitudinal studies in second language acquisition called for by Ortega & Iberri-Shea (2015) and Ross & Masters (2023). Language training at DLIFLC is long-term (up to 67 weeks in length), intensive (six hours of classroom time plus homework) and demanding (graduates must reach ILR Level 2 to be successful). It is not easy to be selected into a language billet, and the selection criteria are such that seats are quite competitive. Only those with high general aptitude scores are given the language aptitude test, and only those with high language aptitude are offered seats at DLIFLC. Drawing on a large dataset comprising over 9,500 service members who completed basic language training

between 2008 and 2018, and who then went on to test in a range of foreign languages, this study found that there was very little language growth after graduation from DLIFLC. Individuals in this sample followed a drop-and-recover pattern, but their trajectories, on average, were almost flat. Given the emphasis placed on graduation levels, and the need for ever-higher levels of language proficiency, these findings should concern the military services. Using the language on the job in itself does not guarantee increased language proficiency. The findings point to a need for continued, intensive language training if meaningful growth in proficiency is to be achieved.

The third research question asked to what extent language aptitude predicted language proficiency growth outcomes across languages, beyond what is predicted by general aptitude. A finding that aptitude predicted growth would mean that existing measures could be used to predict not only the proficiency level at graduation, but also language growth after intensive language training stops. Before discussing the findings, several figures are provided to illustrate the estimates in a model and interpret their meaning. These figures are for demonstration purposes only. A random sample (0.01) was drawn from the final model in Study 1 and illustrated with the 25% and 75% percentiles from each subtest. Figure 8 illustrates *ASVAB-MC* (the mechanical comprehension subtest, coded *ZA_MC*), as a predictor of initial growth, or *Slope12* in the model. The first two bars represent mean scores at the time of graduation (coded 0), and the second two bars represent the mean scores at the second test occasion (coded 1). The white bar represents those in the 25th percentile of *ASVAB-MC* scores, and the diagonally cross-hatched bar those in the 75% percentile. The two groups are quite similar at the time of graduation (the Intercept), but those

with higher *ASVAB-MC* scores drop more steeply between the two test occasions. This drop is captured in the model with the negative fixed effect estimate for *Slope12*ZA_MC*.

Figure 7

*Illustration of Slope12*ASVAB-MC*

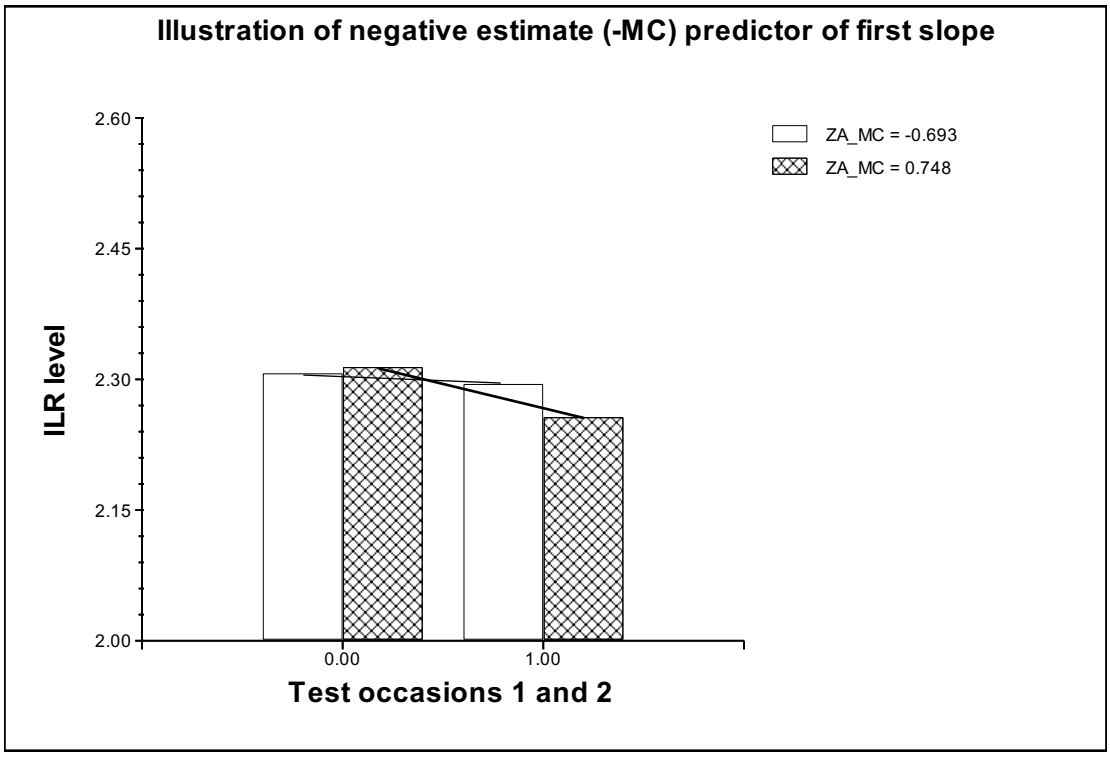
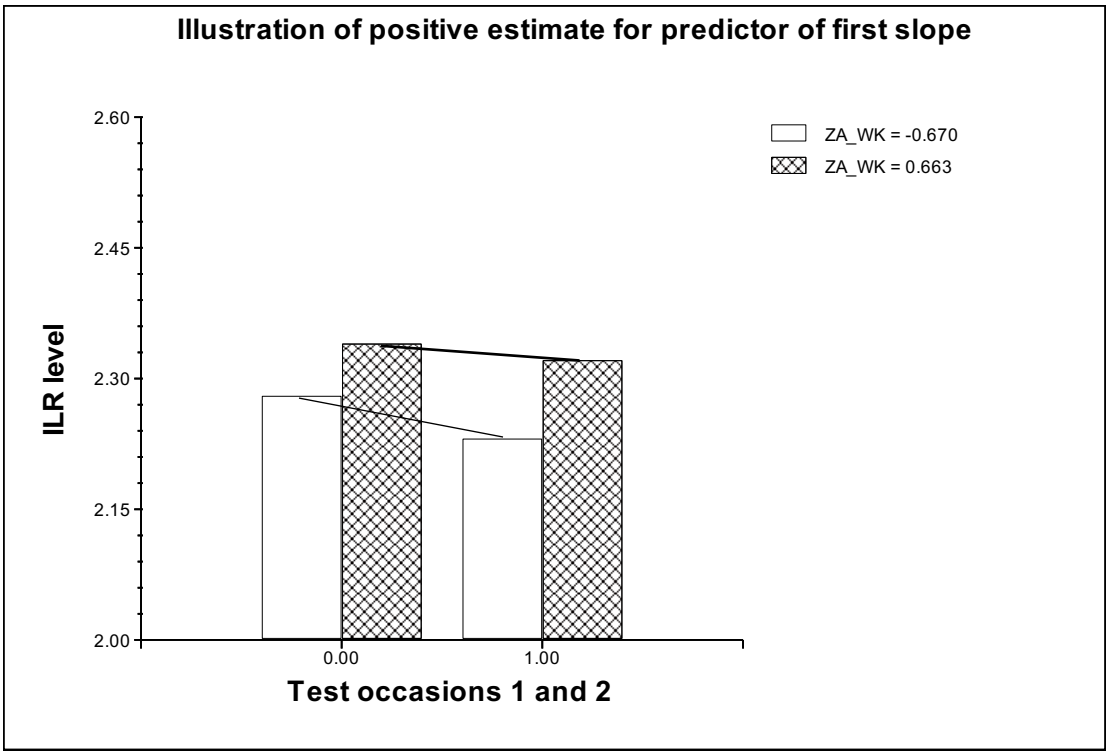


Figure 8 below illustrates the effects of a positive estimate. *ASVAB-WK* (the word knowledge subtest, standardized and coded *ZA_WK*) scores are modeled here as a predictor of growth (*Slope12*). The first two bars represent mean scores at the time of graduation (coded 0), and the second two bars represent the mean scores at the second test occasion (coded 1). The white bar represents those in the 25th percentile of *ASVAB-WK* scores, and the diagonally cross-hatched bar those in the 75% percentile. The two groups differ in their mean ILR level at the time of graduation, and those with lower *ASVAB-WK* scores drop more steeply between the two test occasions.

Overall, the mean slope between the first two test scores is still negative, as scores for both groups fall, but as can be seen with the slope lines between the low/low and high/high bars were drawn, those with higher -*WK* scores have a drop that is less steep. This means that all else being equal, those with higher scores in Word Knowledge have higher proficiency scores when they test a year after graduation.

Figure 8

*Listening HLM Slope12*ASVAB-WK*



Returning now to the question whether aptitude predicts growth, there was limited support in the data for the hypothesis that language aptitude contributes incremental prediction to growth outcomes, after accounting for general aptitude. Table 48 below summarizes the findings for the effects of aptitude on graduation outcomes, initial growth, and subsequent growth in each study, highlighting first the significant ASVAB subtests and then the significant DLAB subtests, by modality

(listening and reading) and methodology (continuous or ordinal outcome). The only subtests in the final models with a significant effect on growth were found in listening, where the mechanical comprehension measure (*ASVAB-MC*), the aural grammatical measure (*DLAB Part 3*) and the concept formation measure (*DLAB Part 4*) subtests were significantly related to growth. The direction of these relationships was negative, meaning that those with higher scores on these subtests had a steeper decrease in their proficiency levels in listening scores in the initial growth period (*Slope12*). There were no significant relationships for any ASVAB or DLAB subtests with subsequent growth (*Slope26*), and no significant relationships were found in reading.

It is not clear why the effects of these subtests were found only in listening and not in reading. The *ASVAB-MC* subtest assesses mechanical comprehension and is one of the components of the technical knowledge factor on ASVAB. This subtest assesses one's knowledge of mechanical concepts such as mass and velocity, and one's ability to apply that knowledge to solve problems. It is likely a measure of explicit inductive learning, and as such, would indicate that those with higher inductive ability lose their language proficiency more quickly in the initial growth period after graduation. As Li (2016) explained, traditional aptitude measures may be more important for preliminary language learning, and for learning in an explicit environment. While *ASVAB-MC* was not a significant predictor of graduation outcomes, the fact that it was a significant, negative predictor of listening growth may be attributed to the environment after graduation, where individuals are using the foreign language, rather than studying it in an intensive program. It is less clear why

there was no finding of aptitude-growth relationships in reading, though it may be that the reading DLPT encourages explicit inductive processing, thereby cancelling out the effect.

The DLAB results that also showed a negative relationship, and only for listening initial growth are also somewhat puzzling. Part 3 is an aural grammatical test, in which the test taker is given linguistic rules that they apply, and Part 4 is a written grammatical test in which the test taker derives the rules from examples using pictures (Bunting et al, 2011, p. 3-25). Those with higher scores on these two parts experienced a steeper drop in their scores in initial growth. It could be that there are other constructs involved in these two DLAB measures, such as rote memory or inductive language learning, that are more strongly called upon during training, and that are no longer called upon by the time the second test was administered. *DLAB Part 3* "builds grammatical learning as it progresses, so that performance on later [parts] is dependent on successfully learning the grammatical rules from earlier [parts]" (Bunting et. al, 2011, p. 7-5). Success on *DLAB Part 3* may call more upon rote learning ability, while *DLAB Part 4* may draw on inductive language learning ability. These skills might be more related to listening than to reading skills, and more important to language acquired in the intensive basic training program. The negative direction of these two estimates suggested that the abilities tapped by DLAB were not as important for growth as they were for demonstrating language proficiency at graduation. There has been little language-related research that included all of the aptitude subtests, so further research on the relationship among the cognitive abilities subtests by modality and growth are called for.

None of the aptitude variables (as main effects or in the interaction terms with FSI0) were predictors of subsequent growth. Researchers have suggested that aptitude may have different effects at early and late acquisition (Doughty, 2019). It is possible that any predictive ability of ASVAB and DLAB is limited to the year after intensive training, captured in this research by the *Slope12* variable. An aim for Hi-LAB was to measure aptitude for implicit learning (Doughty, 2019), and as such it may prove to be a better predictor of growth. More research on other cognitive ability predictors of growth, such as working memory and inductive learning, is called for.

This research did not account for subsequent language training, but it is reasonable to expect that additional formal training, as well as language use on the job, contributed to language growth in the years following the basic course at DLIFLC. The DLPT, as a measure of language proficiency, may not capture these language gains and other measures might be more relevant. The lack of meaningful growth, as well as the limited range of DLAB scores in the sample, could also be contributing to the findings in this research.

Table 48

Comparison of HLM and HGLM significant findings on aptitude in listening and reading in final models

	Listening HLM ASVAB	Listening HGLM ASVAB
Graduation	Intercept: PC, WK	Intercept: PC, WK
Initial Growth	Slope12: MC (neg)	Slope12: MC (pos)
Subsequent Growth	Slope26: None	Slope26: None
	Listening HLM add DLAB	**Listening HGLM add DLAB**
Graduation	Intercept: Part I, Part II, Part III, Part IV	Intercept: Part I, Part II, Part III, Part IV
Initial Growth	Slope12: Part III, Part IV	Slope12: Part III, Part IV
Subsequent Growth	Slope26: None	Slope26: None
	Reading HLM ASVAB	**Reading HGLM ASVAB**
Graduation	Intercept: AR, PC, WK	Intercept: AR, PC, WK
Initial Growth	Slope12: None	Slope12: None
Subsequent Growth	Slope26: None	Slope26: None
	Reading HLM add DLAB	**Reading HGLM add DLAB**
Graduation	Intercept: Part I, Part III, Part IV	Intercept: Part I, Part II, Part III, Part IV
Initial Growth	Slope12: None	Slope12: None
Subsequent Growth	Slope26: None	Slope26: None

The findings in this research replicated those in Silva & White (1993), Bunting et al. (2011) and Wagener (2016) which showed that language aptitude, as measured by DLAB, added incremental validity to the prediction of DLI outcomes beyond those predicted by ASVAB alone. In the present research, the ASVAB and DLAB subtests themselves, rather than composite scores, were included in the models and all nine subtest scores were modeled. This allowed for a more nuanced examination of how aptitude influenced the estimates of language proficiency at the time of graduation. In the final models for listening and reading, with the exception of Study 3 (HLM reading), all four DLAB subtests were significant, positive predictors. Even in the HLM reading study, three of the four DLAB subtests were significant and

DLABPt2 was significant (p = .004), but just under the B-H critical value. The estimates for the intercept were all in a direction that meant that higher aptitude scores were associated with higher ILR levels at graduation. The addition of the language aptitude variables to a model with general aptitude accounted for (ASVAB) reduced the unexplained variance in all four studies, providing additional support for the use of both batteries. The continued relevance of DLAB is an important finding of this research. As explained in the first chapter, even small increases in predictive validity lead to millions of dollars in cost savings. This research supports the continued use of both general and language aptitude tests to select students for the basic course at DLIFLC.

Perhaps unsurprisingly, given that selection into DLIFLC was based on high ASVAB scores, the estimated coefficients for two subtests in the Verbal domain, (*ASVAB-PC* and *ASVAB-WK*) were consistently significant across the studies and had the highest estimates as predictors of the graduation outcomes even in the final model that included the interaction terms as well as previous language learning, motivation, and education. Recall that these two subtests form the Verbal Expression Composite (*ASVAB-VE*) score and also contribute to the composite *AFQT* score, which has often been used by the military services for selection in language training, among other professions. The arithmetic reasoning (*ASVAB-AR)* subtest, which along with math knowledge (-*MK*) and -*VE* comprise *AFQT*, also had a significant effect for outcomes in both final reading models. As Wagener (2016) suggested in his study after finding similar effects, -*AR* scores might indicate a sort of "symbolic assembly" (p. 212) ability that would differentiate learners by language, which might explain why the

variable was significant for reading but not for listening. In all four studies, the *ASVAB-MK* measure was also significant in earlier models but dropped out of significance in the final model when the B-H approach was applied to the data. The findings in the current study seem to confirm the utility of AFQT as a selection measure for prediction of success at DLI graduation, although this could be due to the selection bias inherent in this data.

Given that the effects of motivation, education, prior proficiency and first language on language learning are well established in the literature, these variables were added following the aptitude measures so that their potential influence could be accounted for. While there was no effect found for motivation on graduation (intercept) or growth (slopes), this was not surprising given the inadequacy of the measure used. The source of the motivation measure was one question on a survey that was administered to students prior to their language training. The education variable was found to be a significant, negative predictor of graduation outcomes in listening, even after accounting for general aptitude and language aptitude, but its estimate was almost zero in reading and therefore of little practical value. The direction of the estimate meant that those with more education had lower scores at graduation, all else being equal, which was a counterintuitive effect. It may be that the intensive approach used at DLIFLC differed enough from typical formal education, negating its positive effect. Prior proficiency was found to be a significant, positive predictor of graduation outcomes in both listening models, but not in reading. It did not have a relationship with growth in any of the models. The reliability of this variable should also be questioned, however, given that it was a self-report on a rather

vague scale (poor, fair, good, excellent). The fourth survey variable, English as a first language, had a strong, positive, and significant effect on graduation outcomes, but not growth, in all four studies. Almost all of the individuals in the sample were first language speakers of English, however; accordingly, the use of this variable in the models should be reconsidered.

To improve understanding of how these individual differences relate to language proficiency growth, future studies would benefit greatly from improved variables. For example, to better measure motivation, a new valid and reliable measure of motivation collected over time would improve understanding of how motivation relates to proficiency in a longitudinal study.

The fourth research question examined the effects of language difficulty. While the data was not sufficient to model language at level-3 as originally proposed, six language distance measures were incorporated at level-2 to test whether any individual measure had a significant effect on growth. There was consistent support for the hypothesis that growth is constrained by language distance, after controlling for ASVAB, DLAB and the survey variables. In all four studies, at least three of the measures were significantly related to graduation (intercept) or growth (two slopes). Findings may have been influenced by the fact that the majority of languages in the dataset belonged to harder languages, as sixty percent of the languages in the study were in the hardest DLI category (Cat IV).

Table 48 below depicts a summary of the significant findings. *FSI, TypeRev, NotIndo* were significant predictors of the graduation outcome in all four studies, and *FSI, GateRev, TypeRev, NotLatin*, and *DLI* were always significant predictors of

initial growth. There was less overlap among the various predictors of subsequent growth. What all of the measures did have in common, however, was that with one exception, the direction of the effect (negative) indicated that the harder the language, the lower the ILR level at graduation, the steeper the drop in initial growth, or the shallower the rise in subsequent growth. The one exception to this was the Typology measure in three of four studies, where estimates were in the opposite direction, but because the *TypeRev* estimates themselves were small relative to the intercept or slopes, they had little practical significance.

There are no obvious theoretical reasons for the differences found in the models, but there are some common themes. The FSI and DLI categories were both developed to predict language learning difficulty in intensive, adult language learning settings, and, therefore, it might be expected that they would behave in a similar manner. In most of the models, that did indeed occur, in that if one was significant, the other was as well. The two continuous measures, Gateway and Typology were not always consistent in their effects, despite their high correlation. These measures were approaching zero even when significant, however, and given the lack of reliability information for the measures, it is difficult to draw conclusions.

The non-Latin script measure might have been expected to have a stronger effect in reading, given that its focus is on the writing system of the language, and its effect more likely on graduation. However, that was not the case in these studies, as it was never a significant predictor of graduation. It may be that in language learning, different writing systems are acquired at an early stage, so even though the intercept represents the first test chronologically, DLI students are already well beyond the

stage when another writing system causes learning difficulty. By the time a DLI student graduates and takes their first DLPT, they are already at a proficiency level where the writing system is no longer a novelty, and therefore reading a non-Latin script would not constrain proficiency. However, this theory does not explain why the *NotLatin* measure was a significant predictor of initial growth and subsequent growth in all four studies, listening and reading. This measure may tap into other language features that contribute to language difficulty and therefore constrain growth.

The findings of this research do support the hypothesis that language difficulty constrains growth, though as seen with the other significant variables, the estimates themselves were often quite small, which would have little practical impact. There are policy implications, though, as the results suggest that it is more difficult for individuals in the harder languages to maintain or improve their language, and additional incentive programs should be considered to meet the goal to raise language proficiency across all languages.

The final research question looked at the interaction of language difficulty and aptitude as they related to growth. A finding of a significant effect of such an interaction would be interpreted to mean that the aptitude measures differentially predicted growth depending on language difficulty. No interaction terms were found to be significant in any of the four studies.

The figures shown below illustrate the differences between main effects and interaction terms, drawing on data from the main effects for *DLAB Part 3* (language aptitude) and *FSI0* (language difficulty) for the first slope (growth) from the final model. Figure 10 below is used to illustrate how the main effects of aptitude and

distance from English could be significant, while the interaction is not. In this figure, DLAB Part 3 (*ZD_3*) and the language difficulty measure *FSI0* (0=easier, 1 = hardest languages) are modeled as predictors of the first two test occasions (Slope12). Figure 11 displays the interaction term DLAB Part 3 * FSI0 (coded *D3_FSI0*) for Slope12. These figures are for illustrative purposes only as they only reflect a small portion of the sample to describe the model's effects.

Figure 10 displays four combinations of the variables. The first bar for each test occasion are individuals with low *DLAB Part 3* scores who test in easy languages; the second bar are those with low *DLAB Part 3* scores who test in hard languages; the third bar are those with high *DLAB Part 3* scores who test in an easy language and the fourth bar are those with high *DLAB Part 3* scores who test in a hard language. Individuals with high *DLAB Part 3* scores in the easiest languages (3rd bar) have the highest scores at the time of the first test and experience a drop in scores at the second test. Those with high *DLAB Part 3* scores in the harder languages (4th bar) also experience a drop in score. Those with low DLAB scores, whether in an easy language or hard language, seem to be relatively flat. These effects, in combination with the sample distribution, likely explain why the average slope of the interactions is flat, while the average slope of the individual main effects is not.

Figure 9

*Listening HLM Slope12*DLAB Part 3*FSI0*

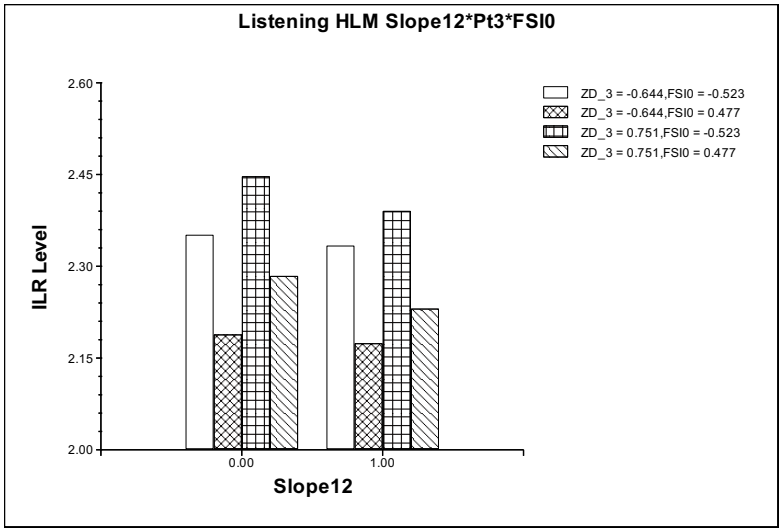
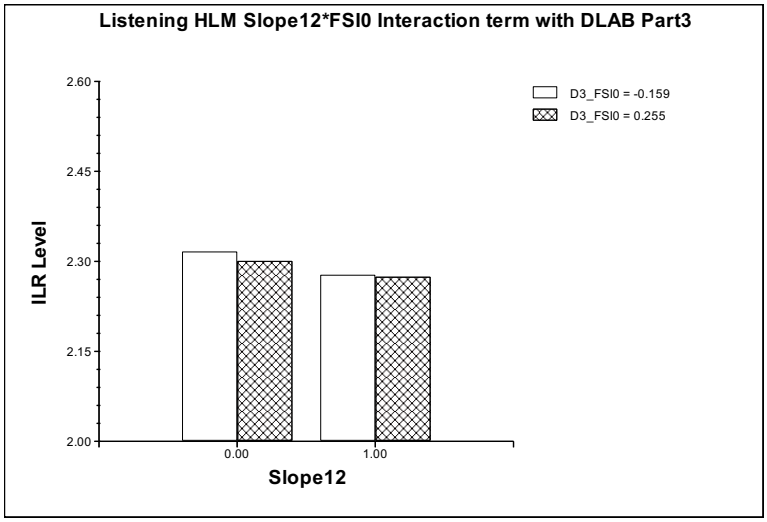


Figure 10

Listening HLM DLAB Part 3 - FSI0 Interaction Term



None of the interaction terms met the criteria for significance and as stated earlier, this might be attributed to the difficulty in trying to predict small amounts of growth with language distance measures that are not reliable, or to the distribution of languages in the sample.

Listening and reading modalities

Listening and reading were modeled separately in this research. For the most part, the results showed very similar patterns in listening and reading. In both modalities, the shape and rate of growth was similar, with a drop-and-recover pattern and very little growth. Average listening scores at the time of graduation were slightly lower than average reading scores. As mentioned above, slight differences were seen in terms of which general and language aptitude variables were significant: the *ASVAB-AR* subtest was a significant predictor (positive) of graduation in reading, but not in listening, while the *ASVAB-MC* subtest was a significant (negative) predictor of initial growth in listening but not in reading. DLAB Parts 3 and 4 were found to be significant as (negative) predictors of initial growth in listening, but once again, not in reading. Negative estimates indicated a decrease in average growth for those with higher scores on the particular subtest, meaning in a steeper decline between the first two tests. There are no obvious reasons for the different findings by modality and these results are difficult to explain. Further research is called for to explore whether these results are idiosyncratic for this dataset.

Two methodologies: HLM and HGLM

One of the contributions of this research to the field is its modeling of the ILR outcome as a continuous measure (Studies 1 and 2 with HLM) and as an ordinal measure (Studies 3 and 4 with HGLM). The analyses showed very similar findings in terms of the shape of growth, as well as in the statistical significance of the covariates. Both methodologies likely violate assumptions central to their analysis: HLM assumes a linear outcome, with equal spacing; HGLM assumes that the effect

of any predictor variable remains constant regardless of the response level. It is perhaps surprising, then, that the results were so similar. Further research on the use of ordinal and numeric scales to represent the ILR scores is called for. In practical terms, it was easier to interpret and graph the results from using the linear model, as the estimates could be more easily understood in the context of ILR levels.

The warrant for a multilevel approach was common across all four studies, as the ICC in both listening studies was 56% and in reading the two studies' ICCs were quite similar, 52-53%. Piecewise slopes were chosen as the best fitting model to reflect time, and the significant fixed effects were similar: the mean level at graduation was in the ILR Level 2 range (or in the ordinal models, more likely to be in the ILR Level 2 range); scores were lower between the first and second test occasions, on average, and rose thereafter, again on average. In all four studies the parameter estimates even for the significant variables were quite small, which means there was little overall growth.

The significance of the variation in growth was not consistent across the studies. In the HGLM reading analysis, the lack of convergence of a random slope model suggested that the slopes be fixed and not allowed to vary, while in the HLM analyses the models converged to allow for a random first slope. Given the somewhat low reliability estimates of the random level-1 coefficients in the three studies with a random first slope, however, it may have been prudent to fix initial growth in the linear models.

In the final model explored in each study, a consistent picture emerged in both approaches in which the main effects of the *ASVAB-PC*, *ASVAB-WK* and all four

DLAB subtests were all found to be significant for graduation, while accounting for education, prior proficiency, first language and interactions between aptitude and a dichotomous FSI variable. The *-AR* subtest was also found to be a significant predictor of reading scores at the time of graduation in HLM and HGLM. It was hypothesized that ASVAB and DLAB would also have an effect on growth, and that hypothesis was only supported by the data in listening, where *DLABPt3* and *DLABPt4* were found to constrain initial growth in both the HLM and HGLM methods. The general and language aptitude variables did not predict subsequent growth. In all, the two methodologies resulted in very similar findings, indicating that for this population, the approach did not make a difference.

The research design in this dissertation contributes to the literature in the language assessment field. The use of a multilevel model is no longer as unique as it once was, but there are very few papers in the second language acquisition context that model repeated measures of language learning using hierarchical linear modeling. The treatment of the ILR scale as a continuous and ordinal measure is also a unique contribution, and the similar findings reported here are somewhat reassuring for those researchers who use a linear conversion of the ILR scale.

After the present research was completed, Rhoades (2023) published her research in which she used another methodology, latent growth curve modeling, to analyze very similar data. She had access to a broader set of influences on language proficiency growth and found that cognitive and non-cognitive factors influenced language growth in this population. Unlike the current research, Rhoades (2023) conducted separate analyses by language and modality. Using only the *-AR, -MK, -*

PC and -*WK* subtests from *ASVAB-AFQT*, all four DLAB subtests, other variables (motivation, age, sex, GPA, education level, training hours, type of billet) and the DLI language difficulty categories, she found that differing patterns of significance emerged in each language and modality, and she concluded that her results highlighted the importance of considering the impact of the individual components of aptitude and limiting generalizations based on composite scores or on more than one language in a difficulty category (p. 284). Additional studies comparing these different approaches to longitudinal data, latent growth curve modeling and multilevel modeling, should be carried out to investigate growth from a variety of perspectives.

Limitations

Only those recruits who performed well on the ASVAB and DLAB are assigned to attend DLIFLC, and only those who successfully completed basic language training served as language professionals who continued to test in their language. This led to range restriction in the general and language aptitude subtests scores, as those who did not qualify for training, in addition to those who did not complete training, never tested and therefore were not in the dataset. Reducing the original dataset to only those who graduated also further truncated the scores in the dependent and independent variables. To minimize the impact of the distribution of ASVAB and DLAB scores, they were standardized. While there are other methods to correct for range restriction by simulating the full population (see Bunting et al, 2011), the data needed for such methods were not available to this researcher. The distribution of languages in this research, which heavily favored the harder languages,

was also a potential issue, but the reality of language training at DLI is that this is the context for military language professionals and will likely be the case for future researchers in this field.

The quality and distribution of the survey variables was a limitation. The four variables that were used in the study were from a self-report questionnaire given to students prior to their basic language training. No reliability information was available for these four variables. A concerted effort on DLI's part to design and administer a better survey, with a higher completion rate, would help future researchers. As mentioned above, it would also be ideal to have a longitudinal measure that better captured motivation across time.

The current study was also unable to account for predictor variables from information available in on-the-job databases, such as job assignments and training information. Significant variation remained unexplained between individuals, and additional sources to explain this variation in proficiency should be investigated. Rhoades (2023), as a government researcher, was able to add variables for language use on-the-job and language training, and she found that in several languages, these job-related variables were significantly related to growth. Such information is not available to the general public and therefore could not be considered in this study.

There is always a danger in over-fitting the data. Attempts were made to correct for the potential for Type I errors. The final models included up to 88 variables, when the main effects of the general aptitude and language aptitude subtests were reintroduced with the interaction terms. The results were corrected with the Benjamini-Hochberg procedure following Thissen et al. (2002). With a large

enough sample size across a variety of languages, future studies could use a cross-validation design that might lead to more generalizable findings.

The HLM8 (Raudenbush & Congdon, 2021) software also had its disadvantages and limitations. Future studies should take advantage of advances in the packages available in R (R Core Team, 2023) for multilevel modeling, as the packages are updated regularly and offer a variety of model improvements not available in HLM, especially in graphing and the ability to model additional variance-covariance matrices to explore assumptions. Improved graphical representations of these complex relationships would improve understanding.

The final limitation to be mentioned here, especially considering the original research questions, was the use of a two-level model. Earlier research indicated that growth rates and trajectories might differ inter-individually as well as by language (Mackey, B., 2014), and indeed those findings motivated the current study which had planned to use a three-level model. A three-level model would have allowed for the individual test scores to be nested by person and then by language, which would have led to a better understanding of the role language played in explaining variation. Fixed and random effects could be modeled at each level: repeated measures, individual, and language. Predictors could be modeled at each level: individual differences in aptitude at level-2, the individual level, and language distance measures at level-3. Cross-level interactions would also be possible to model, such as language distance and aptitude. The inability to use a three-level model was a serious limitation of this study, given the original research questions and the focus on how language moderates growth. However, given that the number of languages taught in the basic

course at the time of this writing was only fourteen (DLIFLC, n.d.), it is not likely that a sufficient sample size will be found in the near future. One possible direction may be to leverage a three-level, cross-classified design, in which individuals are nested under more than one language, which might result in a sufficient number of languages at level-3.

Chapter 8: Summary

In conclusion, this research used multilevel modeling to examine the effects of general aptitude, language aptitude and language distance on language proficiency growth of military language professionals. The findings have implications for DoD policymakers, as they point to a need for further refinement of selection criteria to include a focus on language proficiency after graduation from DLIFLC. While the costs of language training at DLIFLC have led to an emphasis on maximizing outcomes from the basic program, the majority of graduates do not reach the levels needed for successful job performance. The findings in this research indicate that there needs to be a new focus on how to best select those individuals who will be successful in the long run. While findings confirmed the differential validity contributions of the four subtests in DLAB beyond what is predicted by the nine ASVAB subtests for selection, there was only limited support for contributions of cognitive ability components, whether general or language-specific aptitude, to the prediction of growth. There are also a number of non-cognitive variables that could be leveraged to better predict outcomes. The government has already funded research on new aptitude measures designed to improve on existing language aptitude batteries, specifically DLAB2 (Bunting et al., 2011) and Hi-LAB (Linck et al., 2013; Doughty, 2019), which could be leveraged to support further longitudinal research. Neither measure has been deployed operationally in the military services, but a number of past and current language professionals participated in earlier validation studies, and their scores should be available for further longitudinal research. These batteries included new measures such as working memory and personality traits that

are not currently assessed by ASVAB or DLAB, and it may be that language growth post-DLI is better predicted by these different constructs.

The rate and shape of growth after graduation from the basic program should also be considered by policymakers. Proficiency over time is quite flat, and given the language demands on the job, there is a need to revisit language training programs post-DLIFLC if the expectation remains for language professionals to reach and maintain ILR Level 3. Given the overall flat slope of growth in this research, it is clear that additional training interventions are needed to raise the average language proficiency level post-DLIFLC to meet the needs of the nation.

Given the difference between language proficiency and performance mentioned above, the possibility also exists for the research and development of performance measures. The potential relationship between proficiency and performance in the military language environment has not been studied in any depth, and new measures might better capture language growth on the job, and studies could then investigate predictors of proficiency and performance over time.

Drawing on the number of studies that have now shown the constraining effect of language distance on graduation as well as growth, the government should also consider the length of its training programs. Basic language training at DLI currently offers longer courses for languages deemed more difficult for an English native speaker, but nonetheless, differences in graduation outcomes still persist. Course length post-DLIFLC rarely accounts for language difficulty, and given that language distance continued to constrain initial and subsequent growth in this

research, this practice, too, should be reconsidered. Longer courses for harder languages may be needed to maintain and improve language proficiency.

The scope and scale of the language profession in the military is a fertile ground for research if similar data could be made more widely available to researchers. The field of second language acquisition would greatly benefit from continued longitudinal research on these issues, including the treatment of the ILR scale, language growth, predictors of growth, and language distance.

Appendix A

Correlation table for aptitude variables

Correlations

		AO (Assembling Objects)	AR (Arithmetic Reasoning)	AS (Auto / Shop)	EI (Electronic Information)	GS (General Science)	MC (Mechanical Comprehension)	MK (Math Knowledge)	PC (Paragr Comp)	WK (Word Knowledge)	Win-DLAB Part 1 NC	Win-DLAB Part 2 NC	Win-DLAB Part 3 NC	Win-DLAB Part 4 NC
AO (Assembling Objects)	Pearson Correlation	--												
	N	9564												
AR (Arithmetic Reasoning)	Pearson Correlation	.309**	--											
	Sig. (2-tailed)	<.001												
	N	9564	9564											
AS (Auto / Shop)	Pearson Correlation	.192**	.248**	--										
	Sig. (2-tailed)	<.001	<.001											
	N	9564	9564	9564										
EI (Electronic Information)	Pearson Correlation	.225**	.346**	.594**	--									
	Sig. (2-tailed)	<.001	<.001	.000										
	N	9564	9564	9564	9564									
GS (General Science)	Pearson Correlation	.176**	.361**	.423**	.555**	--								
	Sig. (2-tailed)	<.001	<.001	.000	.000									
	N	9564	9564	9564	9564	9564								
MC (Mechanical Comprehension)	Pearson Correlation	.371**	.450**	.567**	.588**	.522**	--							
	Sig. (2-tailed)	.000	.000	.000	.000	.000								
	N	9564	9564	9564	9564	9564	9564							
MK (Math Knowledge)	Pearson Correlation	.225**	.557**	.095**	.264**	.319**	.317**	--						
	Sig. (2-tailed)	<.001	.000	<.001	<.001	<.001	<.001							
	N	9564	9564	9564	9564	9564	9564	9564						
PC (Paragr Comp)	Pearson Correlation	.154**	.314**	.242**	.328**	.383**	.315**	.189**	--					
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	.000	<.001	<.001						
	N	9564	9564	9564	9564	9564	9564	9564	9564					
WK (Word Knowledge)	Pearson Correlation	.099**	.241**	.293**	.403**	.517**	.328**	.132**	.467**	--				
	Sig. (2-tailed)	<.001	<.001	<.001	.000	.000	<.001	<.001	.000					
	N	9564	9564	9564	9564	9564	9564	9564	9564	9564				
Win-DLAB Part 1 NC	Pearson Correlation	-.025*	.088**	-.007	.013	.047**	-.017	.088**	.096**	.140**	--			
	Sig. (2-tailed)	.015	<.001	.500	.215	<.001	.087	<.001	<.001	<.001				
	N	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564			
Win-DLAB Part 2 NC	Pearson Correlation	.131**	.112**	.037**	.055**	.054**	.127**	.086**	.066**	.076**	.021*	--		
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.037			
	N	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564		
Win-DLAB Part 3 NC	Pearson Correlation	.178**	.267**	.053**	.142**	.224**	.183**	.275**	.207**	.257**	.097**	.235**	--	
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001		
	N	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	
Win-DLAB Part 4 NC	Pearson Correlation	.253**	.318**	.110**	.250**	.278**	.267**	.255**	.266**	.275**	.079**	.030**	.252**	--
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.003	<.001	
	N	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564	9564

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Appendix B

Comparison of slope variations in final model

	Slope24	BH sig	Slope25	BH Sig	Slope26	BH Sig
INTRCPT2, β_{00}	2.26 (2.26)	†	2.26 (0.01)	†	2.26 (0.01)	†
ZA_AO, β_{01}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
ZA_AR, β_{02}	0.03 (0.03)		0.03 (0.01)		0.03 (0.01)	
ZA_AS, β_{03}	0.01 (0.01)		0.01 (0.01)		0.01 (0.01)	
ZA_EI, β_{04}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
ZA_GS, β_{05}	0.02 (0.02)		0.02 (0.01)		0.02 (0.01)	
ZA_MC, β_{06}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
ZA_MK, β_{07}	0.02 (0.02)		0.02 (0.01)		0.02 (0.01)	
ZA_PC, β_{08}	0.04 (0.04)	†	0.04 (0.01)	†	0.04 (0.01)	†
ZA_WK, β_{09}	0.05 (0.05)	†	0.05 (0.01)	†	0.05 (0.01)	†
ZD_1, β_{010}	0.04 (0.04)	†	0.03 (0.01)	†	0.04 (0.01)	†
ZD_2, β_{011}	0.02 (0.02)		0.02 (0.01)	†	0.02 (0.01)	†
ZD_3, β_{012}	0.07 (0.07)		0.07 (0.01)	†	0.07 (0.01)	†
ZD_4, β_{013}	0.03 (0.03)	†	0.03 (0.01)	†	0.03 (0.01)	†
EDUC, β_{014}	-0.02 (-0.02)	†	-0.02 (0.00)	†	-0.02 (0.00)	†
PRIORPRO, β_{015}	0.02 (0.02)	†	0.03 (0.00)	†	0.03 (0.00)	†
ENGY, β_{016}	-0.14 (-0.14)	†	-0.14 (0.03)	†	-0.14 (0.03)	†
FSI0, β_{017}	-0.16 (-0.16)	†	-0.16 (0.01)	†	-0.16 (0.01)	†

	Slope24	BH sig	Slope25	BH Sig	Slope26	BH Sig
AO_FSI0, β_{018}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
AR_FSI0, β_{019}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
AS_FSI0, β_{020}	-0.04 (-0.04)		-0.04 (0.01)		-0.04 (0.01)	
EI_FSI0, β_{021}	0.01 (0.01)		0.01 (0.02)		0.01 (0.02)	
GS_FSI0, β_{022}	0.01 (0.01)		0.01 (0.02)		0.01 (0.02)	
MC_FSI0, β_{023}	-0.02 (-0.02)		-0.02 (0.02)		-0.02 (0.02)	
MK_FSI0, β_{024}	0.01 (0.01)		0.01 (0.01)		0.01 (0.01)	
PC_FSI0, β_{025}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
WK_FSI0, β_{026}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
D1_FSI0, β_{027}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
D2_FSI0, β_{028}	0.03 (0.03)	†	0.03 (0.01)		0.03 (0.01)	
D3_FSI0, β_{029}	-0.03 (-0.03)	†	-0.03 (0.01)		-0.03 (0.01)	
D4_FSI0, β_{030}	-0.02 (-0.02)		-0.02 (0.01)		-0.02 (0.01)	
For SLOPE12 slope, π_1						
INTRCPT2, β_{10}	-0.04 (-0.04)	†	-0.04 (0.01)	†	-0.03 (0.01)	†
ZA_AO, β_{11}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
ZA_AR, β_{12}	0.01 (0.01)		0.01 (0.01)		0.01 (0.01)	
ZA_AS, β_{13}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
ZA_EI, β_{14}	0.01 (0.01)		0.01 (0.01)		0.01 (0.01)	
ZA_GS, β_{15}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
ZA_MC, β_{16}	-0.03 (-0.03)	†	-0.03 (0.01)	†	-0.03 (0.01)	†
ZA_MK, β_{17}	0.00 (0.00)		0.01 (0.01)		0.00 (0.01)	

	Slope24	BH sig	Slope25	BH Sig	Slope26	BH Sig
ZA_PC, β_{18}	-0.01 (-0.01)		-0.01 (0.01)		-0.01 (0.01)	
ZA_WK, β_{19}	0.03 (0.03)		0.02 (0.01)		0.02 (0.01)	
ZD_1, β_{110}	0.00 (0.00)		0.01 (0.01)		0.01 (0.01)	
ZD_3, β_{111}	-0.03 (-0.03)	†	-0.01 (0.01)	†	-0.01 (0.01)	†
ZD_4, β_{112}	-0.03 (-0.03)	†	-0.03 (0.01)	†	-0.03 (0.01)	†
PRIORPRO, β_{113}	0.01 (0.01)		-0.03 (0.01)		(not modeled)	
FSI0, β_{114}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
AO_FSI0, β_{115}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
AR_FSI0, β_{116}	-0.01 (-0.01)		-0.01 (0.01)		-0.02 (0.01)	
AS_FSI0, β_{117}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
EI_FSI0, β_{118}	-0.02 (-0.02)		-0.01 (0.02)		-0.01 (0.02)	
GS_FSI0, β_{119}	0.01 (0.01)		0.00 (0.02)		0.00 (0.01)	
MC_FSI0, β_{120}	0.03 (0.03)		0.03 (0.02)		0.03 (0.02)	
MK_FSI0, β_{121}	0.00 (0.00)		0.00 (0.01)		0.00 (0.01)	
PC_FSI0, β_{122}	0.01 (0.01)		0.01 (0.01)		0.01 (0.01)	
WK_FSI0, β_{123}	-0.04 (-0.04)		-0.03 (0.01)		-0.03 (0.01)	
D1_FSI0, β_{124}	0.03 (0.03)		0.03 (0.01)		0.02 (0.01)	
D2_FSI0, β_{125}	-0.01 (-0.01)		0.01 (0.01)		0.00 (0.01)	
D3_FSI0, β_{126}	0.03 (0.03)		0.03 (0.01)		0.03 (0.01)	
D4_FSI0, β_{127}	0.02 (0.02)		0.03 (0.01)		0.03 (0.01)	
For SLOPE24 slope, π_2						
INTRCPT2, β_{20}	0.06 (0.06)	†	0.06 (0.00)	†	0.05 (0.00)	†

	Slope24	BH sig	Slope25	BH Sig	Slope26	BH Sig
ZA_AO, β_{21}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_AR, β_{22}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_AS, β_{23}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_EI, β_{24}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_GS, β_{25}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_MC, β_{26}	0.01 (0.01)		0.01 (0.00)		0.01 (0.00)	
ZA_MK, β_{27}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_PC, β_{28}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZA_WK, β_{29}	-0.01 (-0.01)		-0.01 (0.00)		0.00 (0.00)	
ZD_1, β_{210}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZD_2, β_{211}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
ZD_3, β_{212}	-0.01 (-0.01)		0.00 (0.00)		0.00 (0.00)	
ZD_4, β_{213}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
PRIORPRO, β_{214}	-0.01 (-0.01)		0.00 (0.00)			
FSI0, β_{215}	0.00 (0.00)		0.00 (0.01)		0.00 (0.00)	
AO_FSI0, β_{216}	0.00 (0.00)		0.00 (0.01)		0.00 (0.00)	
AR_FSI0, β_{217}	-0.01 (-0.01)		-0.01 (0.01)		0.00 (0.01)	
AS_FSI0, β_{218}	0.00 (0.00)		-0.01 (0.01)		-0.01 (0.01)	
EI_FSI0, β_{219}	0.00 (0.00)		0.01 (0.01)		0.00 (0.01)	
GS_FSI0, β_{220}	-0.01 (-0.01)		0.00 (0.01)		0.00 (0.01)	
MC_FSI0, β_{221}	0.00 (0.00)		-0.01 (0.01)		0.00 (0.01)	
MK_FSI0, β_{222}	0.00 (0.00)		0.01 (0.01)		0.00 (0.00)	
PC_FSI0, β_{223}	0.00 (0.00)		0.00 (0.01)		0.00 (0.00)	

	Slope24	BH sig	Slope25	BH Sig	Slope26	BH Sig
WK_FSI0, β_{224}	0.01 (0.01)		0.01 (0.01)		0.00 (0.01)	
D1_FSI0, β_{225}	0.00 (0.00)		0.00 (0.00)		0.00 (0.00)	
D2_FSI0, β_{226}	0.00 (0.00)		-0.01 (0.00)		0.00 (0.00)	
D3_FSI0, β_{227}	0.00 (0.00)		0.00 (0.01)		-0.01 (0.00)	
D4_FSI0, β_{228}	-0.01 (-0.01)		0.00 (0.01)		0.00 (0.00)	

References

ACTFL. (2012). ACTFL proficiency guidelines. American Council for the Teaching of Foreign Languages. https://www.actfl.org/educator-resources/actfl-proficiency-guidelines

Anderson, D. (2012). Hierarchical linear modeling (HLM): An introduction to key concepts within cross-sectional and growth modeling frameworks. Behavioral Research and Teaching. Eugene, Oregon.

Army Regulation 350-20/OPNAVINST 1550.13/AFI 35-4004/MCO 1550.4E (2018). Management of defense foreign language training. Washington, D.C. https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB_ID=1001794

Asch, B. J., & Winkler, J. D. (2013). Ensuring language capability in the intelligence community: What factors affect the best mix of military, civilians, and contractors? RAND Corporation.

ASVAB | History of Military Testing. (n.d.). http://official-asvab.com/history_res.htm

Barkaoui, K. (2013). Using multilevel modeling in language assessment research: a conceptual introduction. *Language Assessment Quarterly*, *10*(3), 241–273. https://doi.org/10.1080/15434303.2013.769546

Bauer, D.J. & Sterba, S.K. (2011). Fitting multilevel models with ordinal outcomes: performance of alternative specifications and methods of estimation. *Psychological Methods*. 16(4):373-90. https://doi.org/10.1037/a0025813

Bermudez-Mendez, J. (2020). Student success factors at defense language institute foreign language center. Naval Postgraduate School. Monterey, California. https://apps.dtic.mil/sti/citations/AD1114144

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, *78*(3), 387–409. https://doi.org/10.1348/096317905X26994

Bloomfield, A., Ross, S. J., Masters, M., Gynther, K., & O"Connell, S.Bush, B. J. (1987). The language skill change project (LSCP): background, procedures, and preliminary findings. U.S. Army Research Institute for the Behavioral and Social Sciences. https://apps.dtic.mil/dtic/tr/fulltext/u2/a193081.pdf

Bunting, M., Bowles, A.R., Campbell, S.G., Linck, J.A., Mislevy, M.A., Jackson, S.R. Tare, M, Silbert, N.H., Koeth, J.T., Blake III, C.C., Smith, B.K., Corbett,

R., Willis, R.G., Doughty, C.J. Reinventing DLAB: Potential new predictors of success at DLIFLC. Univerrsity of Maryland Center for Advanced Study of Language.

Bush, B. (1987). The language skill change project (LSCP): Background, procedures, and preliminary findings. Defense Technical Information Center.

Carretta, T. (2014) Predictive validity of the armed services vocational aptitude battery for several US air force enlisted training specialties. AFRL-RH-WP-TP-2014-0046. Air Force Research Laboratory.

Carroll, J.B. & Sapon, S.M. (1959). Modern Language Aptitude Test: MLAT; manual. New York: Psychological Corporation.

Cheslock, J. J., & Rios-Aguilar, C. (2011). Multilevel analysis in higher education research: a multidisciplinary approach. In J. C. Smart & M. B. Paulsen (Eds.), *Higher Education: Handbook of Theory and Research: Volume 26* (pp. 85–123). Springer Netherlands. https://doi.org/10.1007/978-94-007-0702-3_3

Child, J. (1994). List of language distance levels. [unpublished copy].

Child, J. (1998a). Language aptitude testing: learners and applications. *Applied Language Learning*, 9, 1-10.

Child, J. (1998b) Language skill levels, textual modes, and the rating process. *Modern Language Journal*, 31, 3.

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.

Clark, M., O'Rourke, P., Jackson, S., Bloomfield, A., Aghajanian, K., & Kim, S. (2016). The development of the language difficulty categorization framework. Technical paper. , University of Maryland Center for the Advanced Study of Language.

Coakley, T. (2016). Language at the point of need—the defense language institute foreign language center. In S. Berbeco (Ed.), *Foreign Language Education in America: Perspectives from K-12, University, Government, and International Learning* (pp. 190–209). Palgrave Macmillan UK. https://doi.org/10.1057/9781137528506_10

Culver, V. (n.d.). *ASVAB*. Retrieved December 10, 2020, from https://www.officialasvab.com/researchers/references-documentation/

Culver, V. (n.d). *ASVAB Fact Sheet*. https://www.officialasvab.com/researchers/asvab-fact-sheet/

Cysouw, M. (2013). Predicting language-learning difficulty. In L. Boren & A. Saxena (Eds.), *Approaches to Measuring Linguistic Differences* (Vol. 265, pp. 35–58). De Gruyter Mouton. https://doi.org/10.1515/9783110305258

Defense Language Institute. (n.d.) *Languages offered*. Defense Language Institute Foreign Language Center. Retrieved 4 January 2021 from https://www.dliflc.edu/about/languages-at-dliflc/

Defense Language Institute. (n.d.) *Mission and vision*. Defense Language Institute Foreign Language Center. Retrieved 4 January 2021 from https://www.dliflc.edu/about/mission-vision/

Department of Defense/Plans and Resources. (2020). DOD Directive 5160.41E: Defense language, regional expertise, and culture (LREC) program. https://www.hsdl.org/?abstract&did=786667

Department of Defense/Plans and Resources. (2019). DOD Instruction 5160.70: Management of the defense language, regional expertise, and culture (LREC) program. https://dlnseo.org/sites/default/files/DoDI_5160.70.pdf

Department of Defense/Plans and Resources. (2019). DOD Instruction 5160.71: DOD language testing program. https://dlnseo.org/sites/default/files/DoDI_5160.71.pdf

Department of Defense. (2005). Defense language transformation roadmap. https://www.globalsecurity.org/military/library/policy/dod/d20050330roadmap.pdf

Doughty, C. (2019). Cognitive language aptitude. *Language Learning*, 69:S1 pp. 101-126.

Dryer, M. S. & Haspelmath, M., Eds. (2013). WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

Duncan, T.E. and Duncan, S.C. (2004). An introduction to latent growth curve modeling. *Behavior Therapy*, 35, 333-363. Retrieved from http://dx.doi.org/10.1016/j.bbr.2011.03.031

Elder, C., & Davies, A. (1998). Performance on ESL examinations: Is there a language distance effect? *Language and Education*, 12(1), 1–17. https://doi.org/10.1080/09500789808666736

Finch, W. H., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R*. https://ebookcentral.proquest.com/lib/ub-ebooks/detail.action?docID=5829540

Gamallo, P., Pichel, J. R., & Alegria, I. (2017). From language identification to language distance. *Physica A: Statistical Mechanics and Its Applications*, *484*, 152–162. https://doi.org/10.1016/j.physa.2017.05.011

Glisan, E. & Swender, E. & Surface, E. (2013). Oral proficiency standards and foreign language teacher candidates: Current findings and future research directions. *Foreign Language Annals.* 46. https://doi.org/10.1111/flan.12030

Gnanadeskikan, A., & van Rossum, J. (2016). *Gateway Languages Database*. University of Maryland Center for Advanced Study of Language.

Hair Jr., J. F., & Fávero, L. P. (2019). Multilevel modeling for longitudinal data: Concepts and applications. *RAUSP Management Journal*, *54*(4), 459–489. https://doi.org/10.1108/RAUSP-04-2019-0059

Hedeker, D. (2015). Methods for multilevel ordinal data in prevention research. *Prevention Science*, *16*, 997-1006

Held, J. D., Carretta, T. R., Hezlett, S. A., & Mendoza, J. L. (2015). Technical guidance for conducting ASVAB validation/standards studies in the U.S. Navy. Navy Personnel Research, Studies and Technology. Millington, TN.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in human development*, *6*(2-3), 97-120.

Hoffman, L. (2015). *Longitudinal analysis. Modeling within-person fluctuation and change*. Taylor & Francis. https://doi.org/10.4324/9781315744094

Hox, J. J. (2000). Multilevel analyses of grouped and longitudinal data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 15–32, 269–281). Lawrence Erlbaum Associates Publishers.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*(3), 340–362. https://doi.org/10.1016/0001-8791(86)90013-8

Interagency language roundtable skill level descriptions. (1985). https://www.govtilr.org/Skills/ILRscale1.htm

Jackson, S. R., Tare, M., Vatz, K., Silbert, N., Campbell, S. G., Mislevy, M. A., Smith, B. K., Linck, J. A., & Doughty, C. J. (2011). Hi-LAB scoring rubric and aptitude profiles support language personnel selection and training. University of Maryland Center for Advanced Study of Language

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: a meta-analysis: l2 reading and its correlates. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034

Lang, J.W.B., Kersting, M., Hülsheger, U.R. & Lang, J. (2010), General mental ability, narrower cognitive abilities, and job performance: the perspective of the nested-factors model of cognitive abilities. *Personnel Psychology*, 63: 595-640. Https://Doi-Org.Proxy-Um.Researchport.Umd.Edu/10.1111/J.1744-6570.2010.01182.X

Lee, Y., & Kim, S.J. (2010). Linguistic and educational factors affecting TOEFL scores: Focusing on three OECD countries in EFL contexts. *International Journal of Contents*, *6*(2), 33–40. https://doi.org/10.5392/IJOC.2010.6.2.033

Lett, J., Thain, J., Keesling, W., & Krol, M. (2003). new directions in foreign language aptitude testing. *Proceedings from the 2003 IMTA*, 734–741.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*, 385-408.

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition, 38*, 801-842.

Li, S. (2019). Six decades of language aptitude research: A comprehensive and critical review. In Wen, Z., Skehan, P., Biedroń, A., Li, S., and Sparks, R.L. (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 78-96*)*. Routledge.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185–207. https://doi.org/10.1111/lang.12117

Lowe, P. (1998). Zero-based language aptitude test design: where's the focus for the test? *Applied Language Learning*, 9, 11-30.

Mackey, B. (2014) Aptitude as a predictor of individual differences in language proficiency growth. [Unpublished qualifying paper]. University of Maryland.

Mackey, W. F. (1971). *Interlingual Distance*. https://eric.ed.gov/?id=ED049467

Martin, J., Masburn, C. A., & Engle, Randall W. (2020). improving the validity of the armed service vocational aptitude battery with measures of attention control. *Journal of Applied Research in Memory and Cognition*.

Masters, M. (2018). Pathways to proficiency: examining the coherence of initial second language acquisition patterns within the language difficulty categorization framework. [Digital Repository at the University of Maryland]. https://doi.org/10.13016/M2WW77331

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: a review and illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Melby-Lervåg, Monica & Lervåg, Arne. (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, 114 - 135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

O'Connell, A.A. (2000) Methods for modeling ordinal outcome variables, measurement and evaluation. *Counseling and Development*, 33:3, 170-193. https://doi.org/10.1080/07481756.2000.12069008

O'Connell, A.A. (2010). An illustration of multilevel models for ordinal response data. ICOTS8 (2010) Invited paper. Retrieved from www.stat.auckland.ac.nz/~iase/

O'Connell, A., Lo, M., Goldstein, J., Rogers, J., & Peng, J. (2022) Multilevel logistic and ordinal models. In O'Connell, A. A., McCoach, D. B., & Bell, B. A., Eds. *Multilevel Modeling Methods with Introductory and Advanced Applications.* Information Age Press.

O'Mara, F.E., Lett, J.A., & Alexander, E.E. (1994) Language skill change project: The prediction of language learning success at DLIFLC. [Unpublished report, Vol II]. Defense Language Institute Foreign Language Center. Monterey, CA.

O'Mara, F.E., Lett, J.A., Alexander, E.E. (1994) Language skill change project: Post-training foreign language skill change. [Unpublished report, Vol IV]. Defense Language Institute Foreign Language Center. Monterey, CA.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: recent trends and future directions. *Annual Review of Applied Linguistics*, 25, 26–45. https://doi.org/10.1017/S0267190505000024

Ostrow, S. A. (2002). *ASVAB: Armed services vocational aptitude battery* (18th ed). Peterson's. Lawrenceville, NJ.

Peterson, C., & Al Haik, A. (1976). The development of the defense language aptitude battery. *Educational and Psychological Measurement*, *36*(2).

Peugh, J. L., & Heck, R. H. (2017). Conducting three-level longitudinal analyses. *The Journal of Early Adolescence*, 37(1), 7–58. https://doi.org/10.1177/0272431616642329

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, Jr., R.T., du Toit, M. (2021). HLM 8 Hierarchical linear and nonlinear modeling. Scientific Software International.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: not much more than *g*. *Personnel Psychology,* 44(2), 321–332. https://doi.org/10.1111/j.1744-6570.1991.tb00961.x

Rhoades, E.R. (2023). Investigating individual differences' prediction of language proficiency outcomes: a latent growth curve modeling approach. Unpublished doctoral dissertation. University of Maryland.

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The armed services vocational aptitude battery (ASVAB) Little more than acculturated learning (Gc)!?*. *Learning and Individual Differences*, 23.

Ross, S. (2000). 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, Macquarie University. http://www.ameprc.mq.edu.au/__data/assets/pdf_file/0019/241435/Research_Series_11.pdf6page=103

Ross, S. & Masters, M. (2023) *Longitudinal studies of second language learning*. Routledge.

Schepens, J., Slik, F. van der, & Hout, R. van. (2013). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In *Approaches to Measuring Linguistic Differences* (pp. 199–230). De Gruyter Mouton. https://doi.org/10.1515/9783110305258.199

Schmidt, F. L. (2012). Cognitive tests used in selection can have content validity as well as criterion validity: a broader research review and implications for practice. *International Journal of Selection and Assessment*, *20(1),* 1–13. https://doi.org/10.1111/j.1468-2389.2012.00573.x

Schmitz, E. J., Stoloff, P. H., Wolfanger, J. S., & Sayala, S. (2009). *Accession screening for language skills and abilities.* Center for Naval Analyses Technical Report.

Segall, D. and Moreno , K. (1999). Development of the CAT-ASVAB. In F. Drasgow & J. B. Olson-Buchanan (Eds.). *Innovations in Computerized Assessment* (pp. 35—65). Lawrence Erlbaum Associates.

Shearer, S. R. (2013). Modeling second language change using skill retention theory. Naval Postgraduate School, Monterey, CA. https://apps.dtic.mil/docs/citations/ADA585780

Skehan, P. (2019). Language aptitude implicates language and cognitive skills. In Wen, Z., Skehan, P., Biedroń, A., Li, S., and Sparks, R.L. (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 56-77*)*. Routledge.

Silva, J. M., & White, L. A. (1993). Relation of cognitive aptitudes to success in foreign language training. *Military Psychology 5*(2), 79. https://doi.org/10.1207/s15327876mp0502_1

Smith, M., & Stansfield, C. W. (2016). Testing aptitude for second language learning. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-02326-7_5-1

Snijders, T. A. & Bosker, R. (2011). Multilevel analysis: An introduction to basic and advanced multilevel modeling. *Multilevel analysis*. Sage.

Snow, M. S. (1998). Economic, statistical, and linguistic factors affecting success on the test of English as a foreign language (TOEFL). *Information Economics and Policy*, *10*(2), 159–172. https://doi.org/10.1016/S0167-6245(97)00018-8

Steele, F. (2007). Multilevel Models for Longitudinal Data. *Journal of the Royal Statistical Society.* DOI: 10.1111/j.1467-985X.2007.00509.x

Surface, E., Ellington, M. & Dierdorff, E. (2005). A multilevel analysis of language proficiency at the completion of USAJFKSWCS language training: assessing the impact of individual, class and instructor differences on DLPT and OPI scores [SWA Technical Report: 20050672.] Raleigh, NC: Surface, Ward, & Associates.

SWA Consulting Inc. (2009, June). Predictors of proficiency on the OPI: Considering the WPT, Army GT, AFQT, DLAB (Technical Report 62009010612). USSOCOM.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in

Multiple Comparisons. *Journal of Educational and Behavioral Statistics*, *27*(1), 77–83. http://www.jstor.org/stable/3648147

Tigchelaar, M. (2019). Exploring the relationship between self-assessments and opic ratings of oral proficiency in French. In: Winke, P., Gass, S.M. (eds) *Foreign Language Proficiency in Higher Education.* Educational Linguistics, vol 37. Springer, Cham. https://doi.org/10.1007/978-3-030-01006-5_9

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals,* 49(2), 201–223. https://doi.org/10.1111/flan.12198

Verhoeven, L., Perfetti, C., & Pugh, K. (2019). Cross-linguistic perspectives on second language reading. *Journal of Neurolinguistics*, *50*, 1–6. https://doi.org/10.1016/j.jneuroling.2019.02.001

Wagener, T. R. (2016). The influences of aptitude, learning context, and language difficulty categorization on foreign language proficiency.[Unpublished doctoral dissertation]. University of Maryland.

Wang, C.H. (2004). An analysis of factors predicting graduation of students at defense language institute foreign language center. Naval Post Graduate School, Monterey, CA. https://apps.dtic.mil/sti/citations/ADA429897

Watson, A., Harman, R. & Surface, E. (2012). Evaluating DLAB as a predictor of foreign language learning. [SWA Technical Report: 2012010609] .Raleigh, NC: Surface, Ward & Associates. Defense Technical Information Center. https://doi.org/10.21236/ADA585073

Welsh, J. R., & Kucinkas, S. K. (1990). Armed services vocational battery (ASVAB): integrative review of validity studies. [AFHRL-TR-90-22]. Air Force Human Resources Laboratory.

Westrick, P. A. (2017). Reliability estimates for undergraduate grade point average. *Educational Assessment,* 22(4), 231–252. https://doi.org/10.1080/10627197.2017.1381554

Winke, P., Zhang, X., & Pierce, S. (2023). A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency. *Studies in Second Language Acquisition, 45*(2), 416-441. doi:10.1017/S0272263122000079

Zhang, X. (2019). Foreign language anxiety and foreign language performance: A Meta-analysis. *The Modern Language Journal*, 103(4), 763–781. https://doi.org/10.1111/modl.12590