ABSTRACT

Title of Dissertation:	MODERATING EFFECTS OF DIFFICULTY ON INDIVIDUAL DIFFERENCES' PREDICTION OF INTENSIVE SECOND LANGUAGE PROFICIENCY ATTAINMENT
	Catherine Maria Pulupa, Doctor of Philosophy 2024
Dissertation directed by:	Doctor Bronson Hui, Second Language Acquisition

The United States government is perennially in need of employees with proficiency in critical foreign languages to communicate with foreign counterparts and maintain relationships worldwide. In order to fulfill this need, the government devotes significant resources training federal employees to advanced levels of language proficiency through intensive courses aimed at developing communicative language skills that reflect the work that employees will perform in their work advancing the interests of the United States abroad. Notable proportions of employees fail to meet proficiency goals at the end of training, and little is known about what learner individual differences drive whether or not employees will meet their proficiency goals in order to perform their work on behalf of the United States. To this aim, the current investigation utilizes multiple analyses to explore and explain the

interrelationships between learner individual differences, language difficulty, and proficiency attainment throughout training.

The investigation constitutes two related analyses. First, a path-analytic approach examines associations between a cognitive (aptitude) measure and non-cognitive (motivation, familiarity with curricula, previous advanced second language learning) measures with student proficiency achievement throughout training. A second analysis builds on the first: the path-analytic model incorporates a measure of difficulty of the language studied by the students to determine how difficulty influences language learning and ultimate attainment within the context of individual differences in L2 speaking and reading. Results demonstrated consistent influence of language aptitude on proficiency attainment, and notable influences of previous L2 acquisition and the alignment of training to individuals' language use goals. L2 difficulty moderated the relationships between individual differences and proficiency assessment scores during several points in training. The findings support an understanding of adult L2 acquisition that more fully considers learners' goals and previous L2 experiences and consideration of the impact that difficulty can have on individual learners' abilities to achieve target proficiency goals.

MODERATING EFFECTS OF DIFFICULTY ON INDIVIDUAL DIFFERENCES' PREDICTION OF INTENSIVE SECOND LANGUAGE PROFICIENCY ATTAINMENT

by

Catherine María Pulupa

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
2024

Advisory Committee:
Professor Bronson Hui, Chair
Professor Kira Gor
Dr. Martyn Clark
Dr. Megan Masters
Professor Tracy Sweet

Dedication

To my family. Without my husband, Ken, none of this would have happened. Thank you for your love, encouragement, and countless cups of coffee while I reached this goal. To my daughter, Feliciana, the most joyful distraction from this goal imaginable.

To my parents, Ellen and René Pulupa, I hope you can see your Spanish teacher and math teacher inspiration in this work. Thank you for instilling in me a life-long love of languages, learning, and education. To my siblings, whose cheer and character have shaped so much of my life: to my big brothers Patrick and Marc for always giving me enough encouragement to keep me going and enough of a hard time to keep me humble, and to my favorite sister Joan, thank you for treading this path before your older and wiser gemela and showing me the way.

Y sobretodo, a Feliciana y Vincente Pulupa.

Acknowledgements

I have been fortunate throughout my life to have encountered many individuals who have inspired and encouraged my love of languages and research. First, I would like to acknowledge and thank the University of Maryland Second Language Acquisition program. Every professor and fellow student in the program has expanded and challenged my thinking over the last decade. I am deeply grateful to my committee, Drs. Kira Gor, Martyn Clark, Megan Masters, Tracy Sweet, and especially Dr. Bronson Hui, who encouraged me and catapulted me over this final challenge. This work was also shaped by the late Drs. Steve Ross, from whom I learned to see every question through the lens of validity, and Mike Long, whose high standards of assessment of instructed second language acquisition shape my work to this day.

I would also like to thank the faculty of the Quantitative Methodology: Measurement and Statistics program for their statistical training and support throughout my graduate studies, especially Dr. Gregory Hancock for giving me a footing in structural equation modeling, Dr. Laura Stapleton for introducing me to causal interference in educational evaluation, and Dr. Sweet for her training in working as a statistical consultant.

Finally, this work would not have happened without the professional encouragement I have received throughout my career. Dr. Charles Stansfield at Second Language Testing, Incorporated introduced me to the world of language assessment and inferential data analysis. Dr. Rebecca Damari and the research team at the National Foreign Language Center gave me my first opportunities to perform language program evaluation research and changed the course of my career. Finally, my gratitude goes to Ben Kloda and the Evaluation and Measurement Unit at the Foreign Service Institute for their constant good humor and daily thought-provoking discussions.

Table of Contents

Contents

List of Tables

List of Figures

List of Abbreviations

CFI	Comparative fit index
DLAB	Defense Language Aptitude Battery
DLI	Defense Language Institute
DoS	Department of State
FAM	Foreign Affairs Manual
FSO	Foreign Service Officer
GAO	Government Accountability Office
ILR	Interagency Language Roundtable
IRB	Institutional Review Board
L1	First or native language
L2	Second language
LDP	Language-designated position
LTF	Language Training Facility
MLAT	Modern Language Aptitude Test
PA	Progress assessment
RMSEA	Root mean square error of approximation
SEM	Structural equation model
SLA	Second language acquisition
SRMR	Standardized root mean squared residual

Chapter 1: Second Language Acquisition in the United States Federal Government Context

United States Federal Government Language Needs

The United States government is perennially in need of employees with proficiency in critical foreign languages to communicate with foreign counterparts and maintain relationships with the international community. Modern era language skill shortages date to World War I and II, during which the federal government and military scrambled for qualified critical language speakers and established language training centers throughout the country to fill positions to meet the demands of the heightened international activity of the day, training largely monolingual English-speaking United States citizens in Japanese, French, Russian, and Chinese, among other languages (Ornstein, 1956, pp. 213-214). Demand for proficient language speakers did not abate in the following midcentury and Cold War periods, with Arabic, Chinese, Russian, and Korean especially in demand (President's Commission, 1979, pp. 20). In 1956, the first efforts at measuring the proficiency of entering DoS Foreign Service Officers found that 25% tested at a "useful" level (Herzog, n.d.). This pattern continued, as a 1986 DoS analyses found that approximately a third of diplomats at posts abroad lacked the critical language proficiencies required to carry out their duties representing the U.S., including vital posts in the Soviet Union (Goshko). Even following the fall of the Berlin Wall and the end of the Cold War, growing economic interdependence and global connectivity required U.S. federal employees worldwide to employ foreign language skills that continued to fall short of demand (Clifford & Fischer, 1990, p. 117). Following the attacks of September 11, 2001, the Government Accounting Office (GAO) identified the U.S. government's perennial language deficit, especially languages from the "Middle East and Asia", as priorities (GAO, 2002, p. 2), leading to government-wide efforts to increase language learning among U.S. citizens. However, the demand for language specialists remained at long-time high through the early part of the 2000s (Edwards, 2004). The State Department attempted to institute and expand programs of public diplomacy during this period, but language proficiency deficits among employees again constrained these efforts (GAO, 2003).

The languages and linguistic abilities that the U.S. government requires are consistently re-evaluated by the respective federal agencies within the context of their missions. The Mission of the U.S. Department of State (DoS) is "to protect and promote U.S. security, prosperity, and democratic values and shape an international environment in which all Americans can thrive" (Department of State, 2023c): the international aspect of this requires that DoS employees are able to interact and perform work on behalf of the U.S. government in multiple languages. The DoS determines the number of employees and the languages that they need to speak through a Triannual Language Review process, which utilizes interviews and surveys of regional bureau officials and Foreign Service Offices currently serving at language posts worldwide. The positions and language proficiency needs identified in this process are called language designated positions (LDPs), and determine the languages and language training that the DoS needs to provide to its employees to support its mission (U.S. DoS, 2023f). The number of LDPs has increased over the last few decades (Office of Resource Management and Organization Analysis, 2019, p. 4), to 5,200 positions across 64 languages in 2017, the last year for which data is publicly available. The Mission of the U.S. Department of Defense is to "[provide] the military forces needed to deter war and ensure our nation's security" (U.S. Department of Defense, 2023). Military special forces language needs are evaluated every two years by U.S. Special Operations Command: in 2020, over 13,000 positions across 80 languages and in every "Geographic Combatant Command area of responsibility" (GAO, 2023, p. 11).

Current demands for proficient speakers of foreign languages remain high and the DoS and U.S. military struggle to maintain language skills in its personnel. In 2016, 23 percent of DoS LDPs were filled by employees who had not met the language proficiency requirements required; while this was an improvement from the 31 percent in 2008, this still represents a large proportion of positions in which advanced language proficiency is required but unmet (GAO, 2017, p. 8). Similarly, "no more than 45%" of Marine Corps Critical Skills Operators who were required to maintain language skills sustained their required language proficiency, while less than half of Army personnel with foreign language training requirements met their required proficiency levels in 2022 (GAO, 2023, p. 24, 27). These deficiencies point to the need for high-quality language training and identification of the factors that can improve language training outcomes and maintenance skills among critical language proficient personnel.

Federal Language Training

The U.S. federal government maintains several language training facilities for civilian and military employees. These training facilities train federal employees in intensive, full-time courses designed to "[provide] language and culture training to U.S. government employees with job-related needs" (U.S. Department of State, 2023a, para. 1). Federal employees in language training may study at military-specific training institutes such as the Defense Language Institute Foreign Language Center (DLI) (Defense Language Institute Foreign Language Center, 2023), the Air Force Academy (U.S. Air Force Academy, 2023), or DLI-Washington (Defense Language Institute Foreign Language Center, 2023a). Further government training institutes that train both civilian and military personnel include the Foreign Service Institute (FSI) (U.S. Department of State, 2023d) and the National Cryptological University (National Security Agency, n.d.). Although these training facilities differ in their student populations, their core populations consist of federal employees training for specific jobs that require language skills to serve the interest of the U.S. government. Language training is considered to be the full-time job of students in these training facilities, with the expectations that they will be able to functionally utilize the L2 they are trained in at the end of a specified training period. Typically language learners are expected to achieve a score equivalent to an Interagency Language Roundtable (ILR) Proficiency Level of 2, 2+, or 3 in reading and/or speaking at their end of training, depending on agency needs.

The Interagency Language Roundtable Proficiency Scale

The ILR proficiency scale is a second language proficiency scale ranging from 0 through 5, describing a range from no proficiency in an L2 (0) to what is termed "Functionally Native Proficiency" (5). Each point of five-point scale, as well as four intermediary "plus" ("+") levels, is described through series of positively-phrased statements which describe what the language learner is able to do at the specified level. When applicable, the scale also describes what types and complexity of texts learners can interact with and interpret at that level or what types of interactions they can handle at the level. Scales exist in the following domains: reading; speaking; listening; writing; translation performance; interpretation performance; competence in intercultural; communication audio; and translation performance (Interagency Language Roundtable, n.d.).

The ILR scale was developed at the FSI, "based on the experience of FSI Washington staff and in-field service officers concerning typical language use requirements and degrees of language performance demonstrated by FSI graduates in the field" (Clark & Clifford, 1988, p. 130). It was developed in response to the international contact brought about by World War II and the Korean War, when the U.S. government recognized that as a whole, it did not have a standard method of measuring language proficiency (Herzog, n.d.). The initial form of the scale, which measured global proficiency, was developed at the FSI and was published in 1955 (ibid.), with six scale points. Initial assessments performed using this scale were not found to be reliable, and more specific descriptions were developed alongside the first FSI language tests, which consisted of a structured interview. In 1968, interagency cooperation produced skill-specific reading, writing, listening, and speaking scales; these were added to the Government Personnel Manual. These scales were adapted and developed into the NATO proficiency scales (ibid.). Finally, in the 1980s, in recognition of the fact that finer measurements of proficiency were needed to more accurately measure language acquisition, the ETS "Common Yardstick" program developed "+" levels to measure proficiency in between the main scale points using a Department of Education grant (Clark & Clifford, 1988, p. 132).

The ILR scale is used throughout the U.S. federal government as the standard measurement of proficiency to determine capability for performing language-related jobs. For example, the Department of State, Department of Defense, Central Intelligence Agency, Federal Bureau of Investigation, and Peace Corps measure their employee's language proficiency along the ILR scale. While each of these agencies manages their proficiency evaluations using their own instruments or methods, notably the Defense Language Proficiency Tests and ACTFL Oral Proficiency Interview assessments, FSI testing procedures formed the initial bases of many of these other testing procedures (Clark & Clifford, 1988).

Description of the Federal Language Training Facility Examined in this Study

The federal language training facility (LTF) discussed in this study teaches federal employees approximately 60 languages per year. Language classes are given by instructors who have "native or near-native" proficiency in the target languages. The target class size of courses at the LFT is 3, to maximize interaction and speaking practice. Students who are identified as not reaching progress assessment benchmarks may be assigned to smaller classes, and languages with fewer than three students enrolled may have smaller class sizes. Curricula at LFT are usage-driven: qualitative needs analyses based on visits to sites where federal employees work worldwide drive curricula and content practices, though a primary goal of LFT courses is to train students up to their ILR goal, the required proficiency level for their future work. Evidence of achievement of this target proficiency level is measured through three formative progress assessments and one end of training (EOT) test in the modality of the target proficiency level.

Research Questions

This dissertation seeks to address the shortcomings in the supply of critical language proficient federal workers by identifying the learner factors that drive language learning success at the LTF within the context L2 language difficulty. To this end, the following research questions will be investigated:

Research Question 1: Do L2 proficiency assessment scores throughout intensive language study differ between adult L2 learners by aptitude, language learning background, motivation, and previous experience with curricula?

Research Question 2: Does the influence of individual differences on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

Research Question 3: Does language difficulty moderate the relationships between individual differences and L2 proficiency assessment scores throughout intensive language study? If so, which points in learning are affected and in what ways?

Research Question 4: Does the influence of individual differences and difficulty on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

Research questions 1 and 2 expand upon previous research. Using datasets with more diverse populations and wider ranges of proficiency outcomes, these questions are expected to support previous findings and provide more generalizable outcomes. The following questions (3-4) are exploratory, investigating the roles that difficulty plays in learners' proficiency attainment in training and its relationship to individual differences.

Chapter 2: Literature Review

Adult Language Learning in Intensive Federal Training

Multiple analyses have examined the patterns of proficiency growth and ultimate learning outcomes in federal intensive language training programs. These programs are unique in second language research, as they typically consist of full-time adult second language training for specific job-related purposes, with relatively advanced goals of ILR 2 or 3 (proficiency levels approximately equivalent to those typically achieved by a four-year undergraduate language degree in the United States, Winke, et al., 2020, pp. 57-58). Learners are typically federal employees whose job placements are reliant on their proficiency gains, and often have monetary incentives to achieve additional proficiency levels. Learners in these programs, the largest of which are the Department of Defense's Defense Language Institute (DLI) and the Department of State's Foreign Service Institute (FSI), typically spend between twenty-four weeks to two years to learn their target languages from a level of zero proficiency to the level identified as necessary for their jobs, spending 5-8 hours per day in classes. DLI students are typically members of the United States armed forces, though the student population includes a few students from other federal agencies, members of other countries' armed forces, and law enforcement personnel, while FSI students are typically DoS employees, most often Foreign Service Officers, though the student population also encompasses a range of other federal agencies, family members of DoS employees, and occasional military servicemembers. DLI students are typically expected to achieve an ILR 2 in reading and speaking, while FSI students are more often expected to achieve an ILR 3 in both modalities. Studies examining adult L2 language learning in these institutions typically utilize standardized formative and end of training proficiency assessments, cognitive tests, and aptitude measures to predict student progress and proficiency outcomes. These studies are unique in the field of language learning due to their highly regulated environments, large stable populations, and standardized proficiency examinations. This section reviews second language research that has taken place in these unique environments.

Multiple research studies have examined student language learning at the DLI, most often examining the influences of language aptitude, motivation, and cognitive measures of intelligence on language proficiency gains in the context of students' target L2 and its distance from English. Research by Wagener (2015), Masters (2018), Bermudez-Mendez (2020), and Rhoades (2023) have each examined student language learning at the DLI. Each of these studies examined individual student differences in language proficiency outcomes and the influences that individual student differences and linguistic factors had on proficiency.

Wagener (2015) focused on differential influences of a series of language aptitude measures on language acquisition in three environments: foreign language classrooms at the United States Naval Academy (USNA), study abroad with the USNA, and intensive language instruction at the DLI. Controlling for SAT writing and math scores, and undergraduate GPA in a linear regression, aptitude as measured by the Defense Language Aptitude Battery (DLAB) was a significant predictor of language learning in the foreign language classroom setting, but it was not predictive of learning in a study aboard environment or the intensive language learning environment of the DLI.

Masters (2018) utilized a wave path analysis model to explore language proficiency growth in DLI students across three levels of coursework and through their final DLPT exams. This analysis explored the relationships between learners' Armed Forces Qualifying Test (AFQT) results, language learning aptitude (measured by the Defense Language Aptitude Battery (DLAB)), and language preference self-assessment to these four endogenous proficiency measures. Masters found that DLAB was strongly positively related to the three levels of coursework during learners' time at DLI, but that the AFQT results and language preference were not related to these outcomes. In the second "wave", Masters found that the third level of coursework consistently predicted end of training DLPT scores, and identified one language (Arabic) in which the second level of coursework predicted DLPT outcomes.

Bermudez-Mendez (2020) performed a series of logistic growth models, examining learner individual differences' impacts on their success or non-success at meeting ILR 2 proficiency levels as measured by the DLPT at two midpoints in their studies and at their end of training. The models at end of training were most strongly predictive when course grades were added as predictors. Bermudez-Mendez found that prior language experience (as defined by having professional experience in the target language), DLAB scores, and course grades were positively predictive of meeting proficiency goals. Students who were repeating in language studies because they had not previously achieved their goals and learners of "easier" (Category I) languages were less likely to meet their goals (p. xvi). Bermudez-Mendez attributed this difference between learners of difference categories to their lowered DLAB requirements, meaning that their lower proficiency was due to their lower aptitude levels.

Rhoades (2023) utilized latent growth curve modeling to examine the influences of aptitude, general cognitive ability, attitude toward their L2, age, grade point average, sex, educational level, time in training, and billet type on DLI students' proficiency trajectories and language proficiency maintenance following language training. Rhoades found that attitude toward the target language was the only predictor to significantly influence the slope modeling language learning growth. Cognitive measures as measured by subtests of the Armed Services Vocational Aptitude Battery and the DLAB varied in their influences on different language's growth trajectories in ways that were not internally consistent. These mixed results became more unclear when Rhoades included the demographic data in the models, as many previously predictive measures became non-significant, and higher scores on several cognitive subtests even predicted slower language growth within the sample. As Rhoades found several unexpected results in her analyses, and could not fully explain differences between languages of different difficulty categories, and modalities of testing and between learners of the same languages, further exploration into the causes of variation in language growth is called for.

Examining language learning in a different light, Mackey's 2023 research examined factors that influence language proficiency growth following successful federal language training at DLI, to explore post-instruction language growth associated with on-the-job language practice and exposure. While this research was constrained to learners who already had a basis of mid-to advanced-proficiency and did not examine language growth during language training, its findings shed light on the complex manners in which individual differences influence federal language learning. Mackey found slight variation in language learning growth patterns after graduation, including a "drop-and-recover" pattern of lowered then raised proficiency which was attributed to attrition from a gap between training and on-the-job language use. Aptitude as a predictor of language growth following graduation did not demonstrate consistent results, and Mackey explained this contradiction using previous research by hypothesizing that aptitude is more influential at earlier stages of language learning and in instructed language learning. Finally, Mackey found several typological factors that significantly influenced language proficiency: several composite measures of language distance from English, whether or not a language were Indo-European, and whether or not a language utilized Roman orthography were predictive of either proficiency at graduation or proceeding language growth.

Fewer studies, and none in the past few decades, have been published regarding language learning at FSI. One major FSI study to examine outcomes by student individual differences is by Ehrman and Oxford (1995). This study examined correlations of student reading and speaking proficiency outcomes by aptitude measures, aptitude, learning strategies, learning styles, personality, motivation, and anxiety. Aptitude measures, including the sum score of the Modern Language Aptitude Test (MLAT), all MLAT subscores, and reading ability in English positively correlated most strongly with both reading and speaking scores, with all correlations at $r^2 \geq .34$. Motivational factors, self-confidence, and affective measures occupied a second “echelon”, with all measures correlating significantly and positively again with at least $r^2 \geq .34$. Finally, personality measures correlated positively and significantly but with lower $r^2 \geq .20$, and anxiety measures correlated negatively and significantly with learning outcomes. While this analysis was wide-ranging, a major weakness rests in the fact that each variable was correlated independently with proficiency outcomes, leaving unopened the question of which would be more influential were they to be accounted for in a single model reflecting the complexity of individual factors in language learning. In addition to this, small ranges of ordinal values of the ILR scale were treated as continuous linear values, possibly violating the assumption of normality in the tests selected and biasing the outcome parameters and overall results (Rhemtulla, et al., 2012, Beauducel & Herzberg, 2006, Robitzsch, 2020).

A second study by Ehrman (1994), also found that MLAT subscores significantly correlated with reading and speaking proficiency scores, most strongly within Category 1 languages. A multiple regression of proficiency scores on age, education level, number of previous languages studied, highest previous speaking and reading ratings, a general motivation rating, two self-efficacy ratings, two anxiety ratings, and the MLAT Index Score (a MLAT sum score centered for FSI purposes) found that MLAT and previous highest reading scores significantly and positively predicted proficiency. Again, while this research shed light on the student population and factors affecting language proficiency gains at FSI, the same violations of normality assumptions of the models likely biased results.

Finally, a 1990 study by Ehrman & Oxford examined differences between FSI learners with varied self-identified learning styles, Myers-Brigges personality test results, genders, occupations, and ages on the learning strategies employed and instructor evaluations of their study efforts during the course of their language studies (Ehrman & Oxford, 1990). No differences were found by gender, but instructors did evaluate students from different agencies differently, with DoS Foreign Service Officers (FSOs) demonstrating median study skills (p. 316-317). Older students demonstrated poorer study skills (p. 317). Not surprisingly, students who the Myers-Briggs test identified as Extroverts and Feelers used more social study methods, while Introverts and Thinkers used more metacognitive methods (p.317-320). While this study did not examine proficiency outcomes, it did demonstrate that even within the FSI student population, wide ranges of student study methods and language learning beliefs existed despite their similar goals and training experiences.

Influential Factors in Adult Second Language Learning

The following literature review provides summaries of the factors found by these previous federal language training studies to be influential in intensive adult language learning, as well as additional factors that are known to be related to adult L2 acquisition.

Aptitude in Second Language Learning

Language learning aptitude is one of the most studied yet controversial individual differences considered to influence SLA. Defined as "as a special talent for learning languages" (Doughty, 2019, p. 101), aptitude is most frequently described in terms of cognitive abilities or characteristics that contribute to an individual's ability to learn a second language, and determine the ultimate limit of second language learning that a learner can achieve (Doughty, 2013, p. 154). Language learning aptitude is a characteristic of adult language learners, rather than younger learners who have not yet passed out of the biological window of the critical period for language acquisition (Granena & Long, 2013), and is considered a stable individual attribute in adults (Carroll, 1973). Aptitude is also considered to be a separate trait from general intelligence: while both may be predictive of the ability to learn a language, aptitude explains variance in language learning unique from and beyond what general intelligence does (Sasaki, 1993, p. 313; Silva and White, 1993, p. 79). Thus, with all other variables such as general intelligence, motivation, and language learning background held steady, language aptitude would be the key determinant of how successful a different learners would be in their L2 acquisition and their ultimate level of achievement. This holds true in both instructed and non-instructed SLA (see Robinson, 2002, for review) as well as explicit and implicit language instruction (Granena, 2016). Several analyses have found that aptitude is especially impactful on language learning at earlier stages (Li, 2015), rather than at advanced levels (Mackey, 2023, p. 209).

Numerous studies have examined the impact of aptitude on adult L2 learning in intensive federal language learning environments. As previously discussed, Silva and White (1993), Erhman and Oxford (1995), Waegner (2015), and Masters (2018) found strong positive relationships between aptitude measures and L2 acquisition, while Mackey (2023) and Rhoades (2023) found mixed results. Each of these studies utilized index, or sum scores of either MLAT or the DLAB, both of which test components of language aptitude separately but are whose sum scores are more frequently utilized than their subtest results. Further research by Doughty (2019) found that both MLAT and Hi-LAB, an aptitude battery designed to be complementary to MLAT, predicted language learning among FSI students.

The Modern Language Aptitude Test

The Modern Language Aptitude Test (MLAT) was the first assessment instrument designed to measure "an individual's ability to learn a foreign language" (Carroll et al., 2010, p. 2), and has been called the "gold-standard" for aptitude tests (Doughty, 2019, p. 161). MLAT was developed by Caroll and Sapon to predict L2 learning among adult native English speaker language learners under intensive government training. Caroll and Sapon identified four components of language aptitude: phonetic coding ability; grammatical sensitivity; rote learning ability for foreign language materials; and inductive language learning ability, which became the subcomponents of the MLAT (see Sasaki, 2012 for review). MLAT was originally published in 1959, as "Form A"; this is the single test form still used as the entirety of the MLAT test, as no parallel forms were developed. MLAT requires approximately 70 minutes to administer and is scored dichotomously (Language Learning and Testing Foundation, 2021). MLAT has been utilized consistently by government agencies since its publication, and is currently utilized by the LTF, among other organizations. It is considered the "benchmark" of language aptitude assessment, with other language aptitude assessments often designed to measure constructs that are complementary to those already assessed on MLAT (Grigorenko, et al., 2000,, p. 390; Doughty, 2019, p. 161).

Although the MLAT was designed, piloted, and normed with the goal of predicting language learning during the period of behaviorist audio-lingual SLA instruction, its predictive power has remained consistently through transitions to todays' more communicative and task-oriented instructional methodologies (Sasaki, 2012). Initially developed and normed using data from Chinese language students in the U.S. Army and Air Force, high school and college foreign language students, and FSI language students, initial MLAT validation studies found predictive validities between .42 and .62, with reliability ranging between .92 and .97, depending on the population studied (Carroll et al., 2010, p. 8). More recently, Ehrman (1994) examined MLAT's predictive ability among 343 FSI language students, and found a significant correlation of .44 between MLAT and EOT proficiency scores, some variation between languages of different difficulty categories and between speaking and reading modalities (p. 83). Ehrman examined the correlations of both MLAT index (sum) scores and subtest scores, and found that the index was the strongest predictor, but that predictive capabilities of both the index and subscores were less reliably predictive among the strongest language learners (p. 84).

Previous Second Language Learning Experiences

While third-language (L3) acquisition is an underdeveloped area of study, in the field of SLA, a review of literature reveals strong indications that the acquisition of an L2 then a subsequent L3 influences the acquisition of the L3 (Amaro et al., 2012). Numerous studies have examined the effects of positive and negative transfer from the L2 interlanguage to the L3. Interlanguage transfer has been found on lexical, phonological, and syntactic levels. For example, Bardel & Fark argued that there is a "qualitative" difference between L2 and L3 acquisition (2007, p. 459). They utilized a cross-linguistic dataset to argued that observations of L2's influence on L3 syntax in early stages of L3 acquisition demonstrated that L3 Swedish and Dutch learners relied more on their knowledge of negation placement in their L2 than in their native L1 to process L3 input, noting especially strong patterns in beginner L3 learners. Another study by Fark and Bardel (2011) found that intermediate learners' performance on grammaticality judgements regarding object pronoun placement German L3 was more strongly influenced by their L2s (English or French) than by their L1s (French or English) (p. 72-73), demonstrating the L2 influence on L3 learning continues past the beginner levels. Furthermore, De Angelis (1999) demonstrated that lexical interlanguage transfer could occur fully or partially, by observing instances of L2 Spanish words into L3 Italian as either complete words or by influencing the morphosyntactic production of the L3 word. De Angelis noted that the L2 was more strongly influential on the L3 production than the L1 of English.

A larger question than specific interlanguage feature transfer, however, is whether or not it is the act of previous language learning and knowledge of an L2 that changes a learner's overall ability to gain any L3. Klein (1995) hypothesized that acquisition of any interlanguage taught learners skills to improve future language learning by enabling them to more easily access and Universal Grammar parameters by demonstrating that L3 learners had stronger lexical and syntactic learning overall than L2 learners. Competing theories as to why L2s influence L3 acquisition include

Flynn et al.'s Cumulative Enhancement Model, which proposed that "language acquisition is accumulative, i.e. the prior language can be neutral or enhance subsequent language acquisition." (2004, p. 14). Flynn et al. based this theory on an elicited imitation study of L3 learners of English among L1 Kazakh/L2 Russian speakers producing varied relative clauses, concluding that the L1 is not "privileged" in its influence (p. 13). Finally, Rothman's Typological Primacy Model (2011) proposed that L2 transfer to L3 is selective, dependent on the L2's typological similarity to the L3. Rothman examined L3 Spanish among L1 Italian/L2 English speakers and L3 Portuguese among L1 English/L2 Spanish, using semantic interpretation and collocation judgment tasks testing understanding of adjective locations within determiner phases: Rothman found that L3 learners utilized their L1 or L2 by transferring the more typologically similar (L1 Italian or L2 Spanish, respectively).

Prior interlanguages do not only have strictly linguistic effects on subsequent language learning: prior interlanguages have an affective impact on subsequent L2 acquisition. Negative experiences in language learning are understood to accumulate over time, creating permanent negative associations to language study (MacIntyre, 2017). Furthermore, learner affect at the start of language training is especially impacted by the memories and experiences in previous language learning experiences (Saito et al, 2018, pp. 732-733). Both of these have the potential of impacting learner affect or motivation at the start of a new language study period, creating differences in learner affect and motivation between learners who have L2 study experience and those to whom L2 study is a novel experience.

Prior Instructed Second Language Experience within a Single Setting

An understudied area of prior L2 experience is experience with the specific L2 curricula and learning setting. For example, do learners who have previously studied an L2 utilizing a particular curriculum have any advantage or disadvantage compared to L2 learners who have not studied an L2 under the curriculum? Differences in performance by learners with and without prior specific curriculum experience may indicate metalinguistic understanding of L2 as a "system" (Thomas, 1988, p. 235). A study by Pérez (1995) found a greater variety of study methods among students experiencing an intensive university L2 English curriculum for the first time, compared to returning students' study methods (pp. 282-283). Furthermore, Pérez found that new students were concerned with course content (ibid., p. 292), while returning students were not, reflecting their familiarity with the curriculum. In this study, new students also described nearly all English skills taught to them as useful, while returning learners were more likely to describe some skills as unnecessary, and opt out of courses teaching these skills (ibid., p. 292), indicating that their experiences since their first experience in the program had shaped their opinions of the skills they needed. Together, these studies provide a conceptualization of new students as more open to wide varieties of instruction and motivated to study, while returning learners, who may have metalinguistic advantages due to their familiarity with a curriculum, are focused on specific goals and have identified methods that they prefer for language study.

The influence of prior experience with a particular curriculum in a specific setting is likely related to the influence of prior interlanguage acquisition, as described above. The influence of prior interlanguage features, cognitive characteristics of individuals with interlanguages, and potential affective influence of prior L2 study experiences all apply to the situation of an individual repeating a course of study but in a separate L2.

In federal intensive language learning, a study by Ehrman and Oxford (1995) of students at FSI studying L2s found significant correlations between the number of languages that the student had previously studied at FSI and their reading scores at EOT (.34, $p \leq .01$, p. 77) and speaking scores at EOT (.32, $p \leq .01$, p. 76). Ehrman and Oxford noted that the EOT scores from previous language tests were not related to subsequent language test scores, concluding that it was the "breadth of exposure" that made the difference (p. 81).

Language Difficulty

Some languages are more difficult to learn than others. However, difficulty itself as a feature of language acquisition is understudied and ill-defined (Housen & Simoens, 2016, pp.163-164). Studies vary between defining difficulty without specifying *what* makes the languages difficult, and identifying a specific grammatical form or language function as "hard" and compare learner's success acquitting it in implicit versus explicit instruction but generalizability of these studies' findings is low (ibid., p. 165). Furthermore, the many definitions of difficulty used in literature makes understanding a general concept of difficulty a "moving target" (DeKeyser, 2016, p. 353). In a review of difficulty, DeKeyser illustrates this by describing the various manners of defining difficulty, from language characteristics or feature-based, to L1-L2 distance, to the differing degrees to difficulty in that individual learners experience while learning (2016, pp. 354-356)

Housen & Simoens propose a definition of difficulty with a cognitive perspective, framing it as a "how costly, demanding, or difficult a given language feature is for a given language learner in a given learning context, particularly in terms of the mental resources allocated and cognitive mechanisms deployed in processing and internalizing the feature" (ibid., p. 166). Defined in this manner, difficulty encompasses a great many facets of SLA, including individual differences, learning environments, and features specific to the L2s themselves. Housen & Simoens acknowledge that this definition of difficulty itself is two-dimensional. The first dimension is language difficulty that is measurable by *subjective (learner related)* measures, which they define as not just as measures more traditionally considered subjective such as learner personality and motivation but also individual differences considered more objective such as working memory (ibid., p. 167). Housen & Simoens's second dimension of difficulty is *objective, feature-related difficulty*, such as language forms, functions, and their regularity, transparency, and optionality (169).

Multiple scaled measures of overall language difficulty have been proposed, typically with the goal of predicting language learning progress and ultimate attainment. These measures align with language forms that affect Housen & Simoen's *objective* dimension. As noted by Cysouw, typological differences between languages often correlate with continuous difficulty scales (2013, p. 52), making these difficulty scales summative measures of overall differences between an L1 and respective L2s.

Perhaps most prominently among measures of language difficulty, FSI classified languages into four categories of difficulty for L1 English learners of other languages. The FSI categories are used throughout the U.S. federal government and were initially based on mean and standard deviations of the weeks of instruction that FSI students were observed to require to reach goals (Hart-Gonzalez & Lindemann, 1993).

Table 1 provides a list of languages taught at FSI categorized by difficulty, and lists how many weeks of instruction FSI considers necessary for students to learn the language to an ILR 3.

Table 1. Department of State Language Difficulty Categories and Training Periods (U.S. Department of State, 2023)

<table>
<tr><td colspan="3">Category I Languages: 24-30 weeks (600-750 class hours)
Languages similar to English.</td></tr>
<tr><td>Danish (24 weeks)
Dutch (24 weeks)
French (30 weeks)</td><td>Italian (24 weeks)
Norwegian (24 weeks)
Portuguese (24 weeks)</td><td>Romanian (24 weeks)
Spanish (24 weeks)
Swedish (24 weeks)</td></tr>
<tr><td colspan="3">Category II Languages: Approximately 36 weeks (900 class hours)*</td></tr>
<tr><td>German
Haitian Creole</td><td>Indonesian</td><td>Malay</td></tr>
<tr><td colspan="3">Category III Languages: Approximately 44 weeks (1100 class hours)
“Hard languages” – Languages with significant linguistic and/or cultural differences from English. This list is not exhaustive.</td></tr>
<tr><td>Albanian
Amharic
Armenian
Azerbaijani
Bengali
Bulgarian
Burmese
Czech
Dari
Estonian
Farsi
Finnish
Georgian
Greek
Hebrew
Hindi</td><td>Hungarian
Icelandic
Kazakh
Khmer
Kurdish
Kyrgyz
Lao
Latvian
Lithuanian
Macedonian
Mongolian
Nepali
Polish
Russian
Serbo-Croatian
Sinhala</td><td>Slovak
Slovenian
Somali
Tagalog
Tajiki
Tamil
Telugu
Thai
Tibetan
Turkish
Turkmen
Ukrainian
Urdu
Uzbek
Vietnamese</td></tr>
<tr><td colspan="3">Category IV Languages: 88 weeks (2200 class hours)
“Super-hard languages” – Languages which are exceptionally difficult for native English speakers.</td></tr>
<tr><td>Arabic
Chinese – Cantonese</td><td>Chinese – Mandarin
Japanese</td><td>Korean</td></tr>
</table>

* No summative definition of Category II languages is given on the Department of State’s public Foreign Language Training site.

Several other methods of defining difficulty have been proposed and utilized in SLA research. Miller and Chiswick, for example, created a continuous scale of difficulty similar to that of the FSI scale, but more fine-grained, defining the difficulty of a language as the average score on the ILR scale that an adult English L1 learner typically attains after 24 weeks of intensive, full-time language instruction (2004, p. 7). Miller and Chiswick obtained these measures from an FSI publication by Hart-Gonzalez and Lindemann (1993). Cysouw (2013), utilized a variety of measures of linguistic distance, including a seven-point scale based on a now-defunct website by the Office of the Director of National Intelligence which listed the four categories, with asterisks for some languages indicating that they were "typically somewhat more difficult for native English speakers to learn than other languages in the same category" (National Virtual Translation Center, 2007). Cysouw used these asterisks to establish three additional points on the difficulty scale for a total of seven points. Cysouw also created scales that were composite continuous variables representing orthographic distance from English and typological distances from English, respectively, and found that both of these were strongly related to the seven-point FSI scales and the 24-week ILR proficiency scores (2013, p. 72), providing evidence relating the global difficulty measures to typographic characteristics of languages.

Several studies have examined the impact of difficulty on intensive federal language learning. Wagener (2015) examined and compared the effects of individual difference variables on DLI students' language proficiency gains over time in the four federal language difficulty categories. Aptitude as measured by DLAB was predictive of language proficiency gains as measured by students' overall grade point average

(GPA) and arithmetic reasoning was found to be predictive of GPA in categories I, II, and IV but not III (p. 196). Wagener found that the influences of cognitive individual differences in reading and listening proficiency scores (as measured by DLPT), however, differed much more between language categories. In listening, arithmetic reasoning was predictive of proficiency growth in categories I and IV over time only, mathematical knowledge (as a proxy for orthographic abilities) was predictive over time for learners of category IV languages, and for II and III learners in the last two assessments given during training, and aptitude was predictive only for learners of II and IV languages (pp. 196-199). In reading, arithmetic reasoning was positively predictive of growth in category I but negatively in category II, mathematical knowledge was predictive of learning across all categories across time, and DLAB was predictive across time in all four categories but less strongly in later assessments (pp. 199-200). Wagener noted that these differences between categories supported the concept that the language categories, while aligned with the time period necessary to train in a language to a specified proficiency level, were not necessarily aligned with language characteristics and that the different cognitive measures examined represented different aptitude constructs which would support learning of different types of languages, crediting this as support of differential aptitude theory (p. 201).

Three other studies were found to relate language difficulty to proficiency outcomes, although with limited ability to infer differences between languages of different difficulty levels. First, Rhoades examined differences in the influence of individual differences on language proficiency between FSI language difficulty categories, looking at language maintenance following training. Rhoades found that similar individual difference variables predicted language growth in different FSI difficulty categories (2023, p. 278). Second, Mackey utilized a measure of language distance in examining the impact of language difficulty on proficiency gains. She found "consistent support" for the impact of language distance while controlling for general cognitive ability, aptitude, and a survey of motivational variables: language distance restrained both initial growth and long-term acquisition (2023, p. 202). However, this finding was limited due a lack of variability in the languages' difficulty categories (ibid), and the small coefficients, which Mackey noted "would have little practical impact" (p. 204). Finally, Masters found internal "coherence" in development patterns between Arabic, Chinese, and Korean, and considered this to be support for their categorization within a single FSI language difficulty level (2018, p.92). The absence of a parallel analysis in a contrasting proficiency level or levels, however, limits the ability of this claim to support the validity of the difficulty scale.

Alignment of Curricula Tasks to Learner Goals

Language programs such as federal training institutes which train their students for specific job duties develop their training to reflect ultimate language uses and incorporate tasks that the learners will perform in the language post-training into their curricula. For example, the DLI focuses on "mission readiness" in language training (DLI, 2023c, para. 1), while the FSI focuses on "job-related needs" (U.S. Department of State, 2023a). These curricula are thus designed with the primary goal of training L2 learners in completing "tasks", or "the real-world activities people think of when planning, conducting, or recalling their day" in their respective target languages (Long, 2015, p. 6). [1] The use of tasks in curricula development sets standard “expectations of language ability (or aspirations for language learning) that are held for particular individuals, groups, or communities” (Norris, 2016, p. 232). In selection of these tasks, emphasis should be on a “comprehensive range of representation (ibid., p. 234). This relevance to learner goals is accomplished through alignment of tasks to learner goals through needs analysis of “communicative needs to achieve these goals” (Long, 2015, p. 11). Task-centered language instruction that serves these learner needs is considered “effective pedagogy” (Bryfonski & McKay, 2019, p. 622) that satisfies stakeholders (ibid., p 603, Chua & Lin, 2020; NamazianDost, 2017; Kim et al., 2000). The inverse is also true: without aligning instructional tasks to learner needs and goals, language instruction becomes “irrelevant” to students (Long, 2015, p. 14), leading to lessened motivation and engagement.

It is this motivation that may be especially influential to the proficiency outcomes of adult L2 learners who will soon be using their target language to perform their jobs. Extrinsic utility value motivation is defined as “how a task fits into an individual’s future plans”, (Wigfield & Eccles, 2000, p. 72). This is tied to the learners’ “desired end state” (ibid., p 73). Utility value is closely tied to incentives given for motivation and the cost that learners put on their learning efforts, but is driven by the end results of the learning, rather than unrelated rewards or the burdens

[1] While federal language training facilities center language training around target tasks that personnel need to perform in the course of their work (DLI, 2023c, para. 1; U.S. Department of State, 2023a) and have publicly trained instructional staff in Task Based language teaching (U.S. Department of State Foreign Service Institute, 2020), the current research does not evaluate the extent to which LTF syllabi follow or do not follow the philosophical principles (Long, 2015, pp. 63-66) or methodological principles (ibid., pp. 305-325) of Task-Based Language Teaching and does not claim to evaluate the curricula methods themselves.

that learning places upon a learner. This value drives achievement-related choices such as time dedicated to study and the methods and resources used by the learner (ibid., p. 73). In the context of adult language learning for job-related purposes, this is hypothesized to be a major aspect of language learning motivation and thus of proficiency gains. Extrinsic utility value is not well-studied in adult SLA research, but one example by (Meyer, et al., 2019) found that even controlling for gender, socio-economic status, cognitive ability, and learner college major, university students in second-language English courses who assigned higher extrinsic utility value to their English language learning demonstrated higher grades, final exam scores, and overall scores.

This student characteristic of assigning an extrinsic utility value to their language learning is likely associated on a smaller scale to *task motivation* on a day-to-day basis in class. Task motivation, or "learners' motivational disposition while completing a task" (Dörnyei, 2019, p. 59) describes a student's motivation to engage in individual tasks in instructed SLA (p. 60). Without this engagement, Dörnyei argues that students will not "seal the deal", or actually acquire the language (ibid). Student engagement with a task is dependent on several elements of the tasks' relevance to students' ultimate goals. Among these are *task presentation*, or the presentation of a task specifically within the context of how it related to their overall L2 goals; *task goals*, or how related the task is to students themselves and their "personal learning"; and *task content,* which must be "relevant and real" to students (pp. 62-63). According to Dörnyei, these elements are necessary elements to student task motivation. The daily effect of students' motivation and engagement in language learning may compound over time to affect ultimate gains in language proficiency.

Chapter 3: Study Motivation

Second Language Acquisition

This study is motivated by the desire to understand the impacts of the learner individual differences on proficiency achievement throughout long term, intensive instructed adult second language acquisition. The current investigation provides insight into which individual differences drive L2 proficiency at different points in training, within the context of L2 difficulty. This is a clear gap in SLA literature: while vast amounts of research into individual differences in SLA exist, crosslinguistic analysis on this scale has proven largely elusive.

While the impacts of several of the individual differences included in this study have been examined individually, the unique opportunity of being able to explore these factors within a single analysis will grant understanding of these student factors on different stages of language growth and ultimate achievement. Many of these individual differences such as previous L2 study and aptitude, although well-studied independently, have not previously been examined systematically under such well-rounded and longitudinal circumstances, while controlling for other individual differences such as language learning background. Nor have they been examined cross-linguistically, under the controlled circumstance of a single population, curriculum, and proficiency measure.

Impact and Application in Training

This research will not only add to the body of knowledge regarding adult second language acquisition, but will also include practical findings applicable to adult language learning at the LTF. As previously described, the U.S. federal government is continually at need for employees skilled in critical languages. This need for language skills is especially critical in filling employee jobs around the world in a timely manner: delays in reaching language proficiency goals have “in some cases, affected [the government’s] ability to properly adjudicate visa applications; effectively communicate with foreign audiences, address security concerns, and perform other critical diplomatic duties.” (GAO, 2017). Findings from the proposed research will be applicable and actionable in several manners. First, the identification of learner individual differences that are associated with differences in language growth or proficiency outcomes will allow for more accurate planning by determining which learners will likely need additional support in terms of language learning resources such as language learning materials, tutoring or conversation partners, or overall counseling and guidance. Second, the understanding how L2 difficulty drives variance in language learning outcomes may allow the LTF to understand the language-level factors that influence language acquisition to aid in continual training reform, including curricular and instructional adjustments directly addressing linguistic factors and greater understanding of expected proficiency outcomes for planning future adjustments to training period lengths. Finally, utilization the present analysis will enable identification of the times at which students who are likely to not meet EOT proficiency goals can be identified, allowing for targeted student support once struggling students are identified and enabling stronger communication with future job sites, enabling the government to support continuing international relationships.

Research Questions

This dissertation seeks to address the theoretical and practical motivations described above by utilizing student language learner data to determine the individual differences that determine patterns of proficiency growth and ultimate success in the intensive adult second language learning context across languages of different difficulty levels. In order to do so, the following research questions will be addressed:

Research Question 1 (RQ1): Do L2 proficiency assessment scores throughout intensive language study differ between adult L2 learners by aptitude, language learning background, motivation, and previous experience with curricula?

Research Question 2 (RQ2): Does the influence of individual differences on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

Research Question 3 (RQ3): Does language difficulty moderate the relationships between individual differences and L2 proficiency assessment scores throughout intensive language study? If so, which points in learning are affected and in what ways?

Research Question 4 (RQ4): Does the influence of individual differences and difficulty on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

Aptness of Data to Address Gaps in Research

The data selected for this research provide an apt dataset to investigate patterns and influential factors in adult intensive language acquisition. The environment of the LTF provides a uniquely controlled environment with a set population of language learners who experience similar curricula, training policies, and proficiency tests across over sixty languages. While languages follow the weeks of instruction determined by average weeks of study required to meet proficiency requirements (Hart-Gonzalez & Lindemann, 1993), language students vary widely in the ILR reading and speaking abilities that they achieve within the provided time frames, indicating that individual differences or other exogenous variables play key roles in language acquisition.

In addition to this unique language learning environment and population, the proposed research will expand on previous research in several key ways. First, the LTF population differs from federal language learner populations examined in previous research in several ways. The majority of LTF students are expected to learn their target languages to ILR 3 in reading and speaking, as opposed to the more common expectation of an ILR 2 in reading and speaking, giving a wider range of proficiency outcomes to draw conclusions with. Second, DLI students are "re-languaged", meaning that they are removed from language training if they do not meet progress benchmarks, limiting the population of students taking the end of training exams to those who were successful during earlier stages of training. This may limit the accuracy and reliability of measures of language growth by creating a missing-not-at-random dataset in which attrition systematically biases analyses. Finally, the LTF populations is not selected for their language learning aptitude: the DLI population goes through several series of cognitive tests and must meet minimum aptitude measures in order to even study at DLI, with higher standards required to study more difficult languages. Aptitude measures are not used at all in selection of LTF students, nor of the languages they study. Thus, the LTF population demonstrates a wider distribution language aptitudes, enabling more generalizable findings regarding the influence of aptitude on intensive language learning growth.

In addition to utilizing a population with wider distributions of predictive and outcome variables, the proposed research expands on previous intensive language learning research in several key areas. First, the analyses will utilize a cross-linguistic data set in which multiple L2s are incorporated to fully explore the influence of L2 difficulty on language acquisition in the context of individual learner differences. This research also controls for and examines the impact of previous L2 study and knowledge, something which is known to influence L2 acquisition but unexplored in a crosslinguistic setting. Secondly, while previous research relied on survey data to capture student motivation, the proposed research will examine motivation as an externally driven characteristic of the learner's relationship to the language, rather than a self-reported value which may fluctuate on a daily basis or be dependent on their language study progress.

Chapter 4: Data

Data

Data Request

The data used in this study were requested through an Action Memo from the U.S. Department of State following the "[LTF] Guidance Requesting Permission to Conduct Extracurricular Professional or Academic Research Projects" and following all DoS data policies as per Accessing and Using Department of State Information (Foreign Affairs Manual, 2023b).

Exemption from University of Maryland Institutional Review Board

The data for the research was determined to be exempt from Institutional Review Board (IRB) review according to federal regulations by the University of Maryland IRB (University of Maryland Institutional Review Board, 2023). All individual level data is previously-collected and anonymized, and is absent of personal identifiers that could be used to ascertain individual identities (Exemption category #45CFR46.104(d)(4)(ii)).

Language Training Facility Students Population Data

This data consists of demographic records, students L2 proficiency test records at a large-scale U.S. federal language training facility. Each individual case in the data set obtained through the request consists of a student who studied a language at the LTF and took a standardized reading and speaking test EOT assessment on or after January 1, 2022. Each case is an individual student and no student appears more than once in the dataset. Each student in the study population is a beginner language student, meaning they either have no history of exposure or education in the target language, or that they have tested as having an ILR 0 or 0+ in the target language. ILR 0+ is considered a beginner by the LTF because students with this level of proficiency demonstrate similar language growth and achievement patterns to students with no proficiency. Each student in the dataset has an achievement goal of ILR 3 speaking or reading 3 by the end of their training. Because students may have different ILR achievement goals in speaking and reading, the population size between the two modalities varies slightly. As seen in Table 2, there are data for 618 students in the speaking modality, with 354 students (57.28%) having complete data. As seen in Table 3, there are data for 649 students in the reading modality, with 350 students (53.93%) having complete data.

Students in these datasets studied 42 languages. Although the populations between speaking and reading datasets nearly completely overlap, a few LDPs require different ILR levels for reading and speaking, leading to variation in the total counts between reading and speaking dataset populations.

A notable variation between reading and speaking is the distribution of students between languages. For example, in the speaking dataset, the most common languages are Spanish, Portuguese, French, Italian, and Romanian, all Indo-European, Romance languages. Alternately, in reading, the most common are Spanish, Portuguese, French, Indonesian, and Romanian: Indonesian is a non-Indo-European language, and thus a non-Romance language. The size-order of the languages between speaking and reading is varied in this way between the two modalities, with very roughly similar distributions but some difference in the most common target L2. This presents a more homogenous distribution of languages among reading data than speaking. A brief quantitative analysis of five characteristics of languages known to influence language acquisition (Cysouw, 2013, p. 75) found no significant differences between reading and speaking in the proportions of languages studied that were Indo-European ($\chi^2(1, n = 1266) = .544, p = .456$), ergative ($\chi^2(1, n = 1266) = .863, p = .353$), use postpositions ($\chi^2(1, n = 1266) = 1.277, p = .258$), obligatorily use nominative plural marking ($\chi^2(1, n = 1266) = .054, p = .816$), or use Roman orthography ($\chi^2(1, n = 1266) = .545, p = .455$). The reading and speaking datasets also did not differ in terms of average difficulty as measured by the expected score after 24 weeks of study ($t(1264) = .332, p = .740$). Although these analysis do not fully mitigate the effects of the different distributions of target languages between the two datasets, this does at least indicate that differences in results between the two datasets is likely not attributable to these typological differences. However, differences between speaking and reading results may be influenced by the proportions of individual languages, for example, in reading, Indonesian cases are 4.62% of the data, while in speaking Indonesian makes up less than 0.97% of the data (6 cases). Thus the particular typological characteristics of Indonesian or characteristics of the Indonesian program may be particularly influential on the reading outcomes in a way they may not be in speaking. This potential is true across all of the languages in the datasets, and is not controlled for in the current study.

Table 2. Student Sample Population by Language and Completeness: Speaking

Language	Total N	Percent of Sample	No Missing Data		Some Data Missing	
			n	%	n	%
Arabic (Formal Std)	12	1.94%		-	12	100.00%
Azerbaijani	8	1.29%	5	62.50%	3	37.50%
Czech	5	0.81%	2	40.00%	3	60.00%
French	94	15.21%	71	75.53%	23	24.47%
Georgian	5	0.81%	4	80.00%	1	20.00%
German	17	2.75%	15	88.24%	2	11.76%
Greek	5	0.81%	5	100.00%		-
Indonesian	6	0.97%		-	6	100.00%
Italian	34	5.50%		-	34	100.00%
Portuguese	95	15.37%	25	26.32%	70	73.68%
Romanian	23	3.72%		-	23	100.00%
Russian	18	2.91%	14	77.78%	4	22.22%
Spanish	217	35.11%	190	87.56%	27	12.44%
Thai	8	1.29%		-	8	100.00%
Turkish	12	1.94%		-	12	100.00%
Other*	59	9.54%	23	38.98%	36	61.02%

* "Other" languages consist of those with fewer than 5 learners in the dataset. These languages include: Albanian, Armenian, Bulgarian, Burmese, Chinese (Mandarin), Danish, Dutch/Flemish, Estonian, Hebrew, Hindi, Hungarian, Kazakh, Latvian, Macedonian, Malay, Nepali/Nepalese, Norwegian, Persian/Farsi/Irania, Persian/Tajiki, Polish, Slovak, Slovenian, Swahili/Kiswahili, Turkmen, Ukrainian, Urdu, and Vietnamese.

Table 3. Student Sample Population by Language and Completeness: Reading

Language	Total N	Percent of Sample	No Missing Data		Some Data Missing	
			n	%	n	%
Arabic (Formal Std)	11	1.69%		-	11	100.00%
Azerbaijani	8	1.23%	5	62.50%	3	37.50%
Chinese (Mandarin)	4	0.62%	2	50.00%	2	50.00%
Danish	4	0.62%	1	25.00%	3	75.00%
Estonian	5	0.77%	4	80.00%	1	20.00%
French	89	13.71%	69	77.53%	20	22.47%
Georgian	5	0.77%	5	100.00%		-
German	17	2.62%	15	88.24%	2	11.76%
Hungarian	5	0.77%		-	5	100.00%
Indonesian	30	4.62%		-	30	100.00%
Nepali/Nepalese	4	0.62%		-	4	100.00%
Portuguese	93	14.33%	23	24.73%	70	75.27%
Romanian	22	3.39%		-	22	100.00%
Russian	19	2.93%	15	78.95%	4	21.05%
Spanish	266	40.99%	189	71.05%	77	28.95%
Swahili/Kiswahili	4	0.62%	3	75.00%	1	25.00%
Turkish	12	1.85%		-	12	100.00%
Other*	51	7.83%	19	37.25%	32	62.75%

* “Other” languages consist of those with fewer than 5 learners in the dataset. These languages include: Albanian, Armenian (Eastern), Bulgarian, Burmese, Czech, Dutch/Flemish, Greek, Hebrew, Hindi, Italian, Kazakh, Latvian, Macedonian, Malay, Norwegian, Persian/Farsi/Irania, Persian/Tajiki, Polish, Slovak, Slovenian, Thai, Turkmen, Ukrainian, Urdu, and Vietnamese.

The LTF language student population is apt for a study examining the research questions described above. The LTF presents a highly controlled environment with standardized goals and training methods in which a wide range of languages (approximately 70 on average per federal fiscal year) are taught to a large population of language learners. LTF language programs follows analytic language syllabi, rather than synthetic or form-focused language syllabi (Long & Crookes, 1992, p. 28), with the goal of training their students to perform work for the federal government, with a wide range of interlocutors and professional tasks. Nearly all language learners at the LTF are federal employees, with few exceptions made for close family members of federal employees serving abroad.

The LTF provides instruction in a wider range of languages than its comparable government institute, the Defense Language Institute, whose students and courses have been well-examined in several recent studies as described in the preceding literature review. Indeed, the LTF presents a unique opportunity to study intensive adult language learning in a controlled environment, as these previous studies focused on military service member adult learners who were selected for their language learning aptitudes through personality type assessments, motivational questionnaires, and whose target languages were selected using language aptitude measures. The LTF primarily serves civilian students who serve abroad. While a great many variables may play a part in the country that a learner wishes to serve in, as civilians, aptitude measures cannot be used in the selection process as a deciding factor for what employee serves in what country, and thus does not play a part in whether they study a language or the selection of the target language that they learn. Thus the low variance in aptitude measures observed in DLI students is not a factor in LTF student population, enabling a clearer understanding of the impact of aptitude on language acquisition.

Furthermore, LTF employees do not typically serve in posts for more than 2-3 years, and the majority return multiple times to the LTF to gain proficiency in a third or fourth L2. This further enriches the dataset in question and its ability to shed light on the language learning process by allowing the observation of the influence of prior language learning on subsequent L2 acquisition, as well as the impact of learning a language with or without previous knowledge of the curriculum and tasks the curriculum is based on to scaffold the novel L2 learning.

Individual Difference Data

Modern Language Aptitude Test Scores

MLAT index sum scores will be used as a measure of aptitude to learn a second language. MLAT scores at LFT are "indexed", or centered at the mean sum score for the LTF. Table 4 and Table 5 describe the descriptive statistics for MLAT in each modality.

Table 4. MLAT Descriptive Statistics (Speaking Data)

	n	Mean	Std. Deviation	Minimum	Maximum	Skew	Kurtosis
Dataset with Complete Cases Only	354	59.97	9.06	29	80	-.53	.07
Dataset with Incomplete Cases	191	60.97	8.48	36	80	-.39	.07
Total	545	60.32	8.87	29	80	-.49	.09

Table 5. MLAT Descriptive Statistics (Reading Data)

	n	Mean	Std. Deviation	Minimum	Maximum	Skew	Kurtosis
Dataset with Complete Cases Only	350	60.05	9.03	29	80	-.53	.09
Dataset with Incomplete Cases	213	60.63	8.88	27	80	-.58	.36
Total	563	60.27	8.98	27	80	-.54	.17

Role at Post

LTF students are categorized in this study as following one of two career paths: *specialized* and *non-specialized,* and it is hypothesized that the differing language demands of the roles they are training for may lead to differences in their language learning processes. Most are considered non-specialized: these are employees who select one of five career paths related to international relations that they will likely follow for their entire career. Non-specialized students train in their target languages knowing that they will most likely use their language in their day-to-day work at their work abroad. The other category of LTF students typically "have a professional or a functional skill for which there is a continuing need" (Foreign Affairs Manual, 2023a, para. 2218.2-1.a). Many of these employees with specialized roles perform jobs that are classified into one of six specific professional categories and employees enter these jobs as already-highly-trained professionals in their field and undergo language training with the understanding that their work for the U.S. government will be to provide these specialized skills, and that their work abroad may be performed in English. Other federal employees study languages at the LTF as well, ranging from military language specialists to United States Agency for International Development officers. These employees are similar to specialized employees in that they are not training for more general positions, but for specific positions utilizing their professional skills.

The language tasks taught during training are typically focused on the needs of non-specialized learners, so specialized learners' studies are focused on tasks that differ from their ultimate language use goals at post. These differences in the alignment of tasks studied and language use goals between the two employee types are hypothesized to create differences in the extrinsic utility value which LTF students give their language learning. Extrinsic utility value is the aspect of integrative language learning motivation that describes the value learner gives to their potential use of the language (Wigfield & Eccles, 1992, p. 16). In this case, the potential use of the language is different between nonspecialists, who know that they will rely on the language learned to carry out their future tasks, versus employees with specialized roles and other federal employees, who will rely more heavily on their pre-existing professional skills or who are training for other professional positions.

For example, a learner with a non-specialized position likely studies under a curriculum developed specifically around their future work: they may be trained in specific negotiation practices or tasks to support US citizens abroad, the tasks they will perform in their future jobs. However, their specialized classmates may be facilities management professionals, whose future work tasks are not specifically targeted by the curriculum. For example, a facilities manager may need to work at an advanced L2 level in a professional setting, for example, arranging for municipal permits to close a road and renting construction vehicles for repairs to U.S. property abroad, but such L2 tasks are not in the more generalized coursework since they are designed for the general LTF population. This difference between the alignment versus misalignment between students' future work and the tasks they are trained to perform in their L2 is hypothesized to drive differences in the extrinsic utility value they assign to their language training, creating differences in motivation between learners with specialized versus non-specialized jobs.

While employees with nonspecialized and specialized jobs have the same target level of proficiency, and are required to meet these proficiency levels, these differing values that they give to their language skills is hypothesized to impact language learning. This variable is dichotomous, with 0 representing students with nonspecialized roles and 1 representing students with specialized roles and other employees. Table 6 and Table 7 describe the distributions of employee roles in both modalities, respectively.

Table 6. Employee Roles: Speaking

	Nonspecialized		Specialized	
	n	%	n	%
Dataset with Complete Cases Only	279	78.81%	75	21.19%
Dataset with Incomplete Cases	218	82.58%	46	17.42%
Total	497	80.42%	121	19.58%

Table 7. Employee Roles: Reading

	Nonspecialized		Specialized	
	n	%	n	%
Dataset with Complete Cases Only	282	80.57%	68	19.43%
Dataset with Incomplete Cases	258	86.29%	41	13.71%
Total	540	83.20%	109	16.80%

Prior Test

Student language learning background is known to impact second language acquisition, especially the previous acquisition of a second language. *Prior Test* is a dichotomous variable indicating whether (1) or not (0) a learner has a language test on record prior to the start of study of their current target language study at the LTF. This variable represents whether the learner either has studied a language at the Institute and took an end of training test, or reported that they had extensive understanding of a language other than English as an L2, native, or heritage speaker and tested in this language for professional purposes. LTF students are motivated to report extensive language experience, due to the increased pay that is given to those who have proficiency in select languages (U.S. Department of State, 2023h), the "bonus points" that are granted to applicants who can prove they have prior language proficiency (U.S. Department of State, 2023i), or a desire to "test out" of required language training by proving pre-existing proficiency. This variable demonstrates prior proficiency in a non-English language, whether through study at the LTF, other L2 learning experience, or as a heritage learner, although it does not specify the level of proficiency of this prior language experience nor the specific language the learner tested in. This variable represents the linguistic and affective effects of prior interlanguages, whether positive or negative. Table 8 and Table 9 describe the distributions of students who had taken previous LTF tests.

Table 8. Previous Language Tests and Language Tests in Language from Same Family: Speaking

	Prior Test			
	Prior Test		No Prior Test	
	n	%	n	%
Dataset with Complete Cases Only	253	71.47%	101	28.53%
Dataset with Incomplete Cases	202	76.52%	62	23.48%
Total	455	73.62%	163	26.38%

Table 9. Previous Language Tests and Language Tests in Language from Same Family: Reading

	Prior Test			
	Prior Test		No Prior Test	
	n	%	n	%
Dataset with Complete Cases Only	248	70.86%	102	29.14%
Dataset with Incomplete Cases	209	69.90%	90	30.10%
Total	457	70.42%	192	29.58%

First Tour

Students at the LTF are defined as being on their "first tour" [of duty] at the LTF if they are studying a language at the LTF for the first time. This is operationalized by identifying students who were enrolled in the LTF orientation courses immediately prior to their enrollment in language study. Learners who are returning to the LTF for language study are not on their "first tour" of duty at the LTF. This variable is measured dichotomously and stands as a proxy for student understanding of language learning goals and experience with the LTF curriculum, policies, and proficiency goals, and the impact that this knowledge and experience has on their subsequent learning experience. Notably, a learner who is a "first tour" learner at the LTF may have prior instructed or non-instructed language learning experience or they may have native or heritage language abilities in another language, limiting interpretation of this variable to the influence of experience with this specific curriculum in the LTF setting. Table 10 Table 11 describe the distributions of cases in of first tour learners in speaking and reading.

Table 10. First Tour of Language Learning: Speaking

	First Tour		Not First Tour	
	n	%	n	%
Dataset with Complete Cases Only	87	24.58%	267	75.42%
Dataset with Incomplete Cases	57	21.59%	207	78.41%
Total	144	23.30%	474	76.70%

Table 11. First Tour of Language Learning: Reading

	First Tour		Not First Tour	
	n	%	n	%
Dataset with Complete Cases Only	89	25.43%	261	74.57%
Dataset with Incomplete Cases	70	23.41%	229	76.59%
Total	159	24.50%	490	75.50%

<u>*Language Data*</u>

Difficulty

Languages included in the analysis are described by the continuous variable *Difficulty* as defined by Miller and Chiswick (2004). This variable represents the average tested ILR ability of a full-time language adult language learning in federal language training following 24 weeks of training (p. 7). Miller and Chiswick utilized values from an FSI report by Hart-Gonzalez and Lindemann (1993) to create a seven-point scale of language difficulty. When the values were not available from Miller and Chiswick, data was supplemented by utilizing the LTF benchmark for training at 24 weeks. Three languages did not have an available 24-week benchmark value, and the training benchmark for 21 weeks was utilized. The data here were treated as continuous, as the values represent unequally spaced scores along the ILR proficiency scale. Difficulty values by language are listed in Appendix A.

Table 12. Difficulty Descriptive Statistics (Speaking Data)

	n	Mean	Std. Deviation	Minimum	Maximum	Skew	Kurtosis
Dataset with Complete Cases Only	354	2.30	0.16	1.75	2.75	-0.26	1.08
Dataset with Incomplete Cases	264	2.34	0.37	1.50	3.00	-0.47	0.19

Table 13. Difficulty Descriptive Statistics (Reading Data)

	n	Mean	Std. Deviation	Minimum	Maximum	Skew	Kurtosis
Dataset with Complete Cases Only	350	2.29	0.16	1.75	2.75	-0.25	1.14
Dataset with Incomplete Cases	298	2.33	0.37	1.50	3.00	-0.41	0.73

Language Proficiency Measures

End of Training Speaking and Reading Scores and LDP Achievement

Students with ILR proficiency goals to fulfill job requirements the LTF are given assessments in reading and speaking in their target L2s at their scheduled EOT. These exams are scored on the ILR scales for reading and speaking. The speaking test consists of “scenario-based, topical conversation that flows through a series of listening and speaking tasks” (Sawyer, 2022, para. 16). This scenario is a professional meeting or conferences, similar to that which employees attend as part of their duties, and conversation consists of both concrete conference details from a conference program and professional abstract discussion reflecting the examinee’s career track (ibid.). Reading scores are based on examinee’s ability to report what they read in their L2 in English in terms of both gist and details. Students are given a set period of time to read an ILR-3 leveled passage and then report its content to an interlocuter, and are scored on reporting both accurate gist and details.

For the purposes of the quantitative analyses in this study, these scores will be converted from ordinal scales, i.e. 0, 0+, 1, to a continuous scale, with “+” scores considered as .6, for example, a 0+ is a .6. This is possible because ordinal scales that consist of more than 4 points produce unbiased parameter estimates when converted to continuous scales (Rhemtulla, et al., 2012, Beauducel & Herzberg, 2006, Robitzsch, 2020). The use of .6 rather than .5, a midpoint between the integer values of the ILR scales is recommended by the Interagency Language Roundtable for “when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level.” (Interagency Language Roundtable, n.d., para.

6). This converted value is used for quantitative research utilizing the ILR scale, especially in crosslinguistic research to compare scores between languages (see Rhoades, 2023; Wagener, 2015).

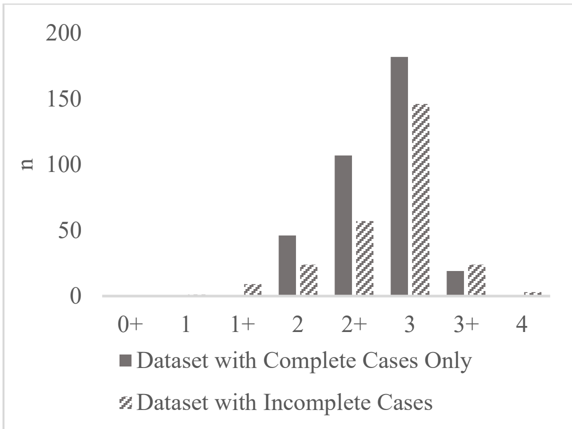

Figure 1. End of Training ILR Speaking Score Distribution

Table 14. End of Training ILR Speaking Score Descriptive Statistics

	n	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
Complete Cases Only	354	2.78	0.39	2.00	3.60	-0.50	.32
Incomplete Cases	264	2.83	0.47	1.0	4.00	-0.77	1.43
Total	618	2.77	0.43	1.0	4.00	-0.65	0.63

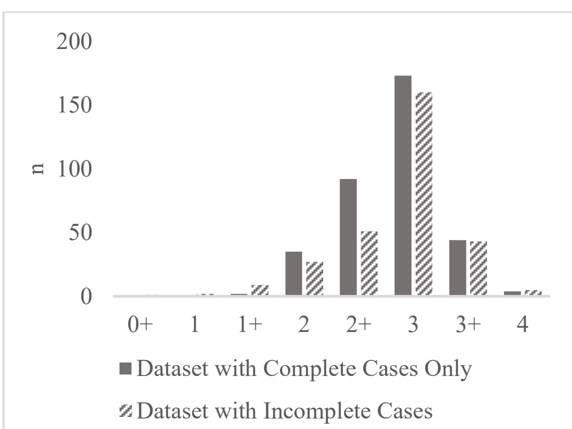

Figure 2. End of Training ILR Reading Score Distribution

Table 15. End of Training ILR Reading Score Descriptive Statistics

	n	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
Complete Cases Only	350	2.87	0.44	1.60	4.00	-0.22	0.43
Incomplete Cases	299	2.88	0.53	.60	4.00	-0.92	1.86
Total	649	0.5	4.0	2.84	0.48	-0.73	1.38

Formative Progress Assessments

LTF students are given three progress assessments, which are formative proficiency exams, at quarterly points in their studies. While these assessments are not officially standardized, the traditional format of the progress assessments is a simulated EOT exam. This practice is followed by all languages in this study. In theory, this enables the assessment to have a high degree of predictive validity in terms of its association with EOT test scores. This also grants the progress assessments a high degree of face validity for students and instructors, who are focused on language use in terms of passing EOT tests. As student proficiency grows throughout the course of study at the LTF, progress assessments more closely resemble the end of training assessment, as they are more able to perform the tasks on the final EOT assessment, especially at the third assessment, which is the final formal proficiency measure taken before the EOT assessment.

These assessments are scored on the ILR scale, and they are used to determine which students are progressing adequately in their language studies and which may need additional support. In some language programs, students who score significantly below the expected ILR benchmark at particular assessment points are given additional training materials, support, or provided smaller class sizes. Furthermore, students who are identified as not meeting progress assessment proficiency benchmarks may be identified to their future workplaces as not likely to arrive on time, i.e., that they are likely to not meet their EOT proficiency goals and will require additional training time at the LFT. Students are aware of these consequences, and may experience increased test anxiety (Horwitz et al., 1986, p. 127) with each subsequent assessment due to the greater consequences and the approaching EOT test.

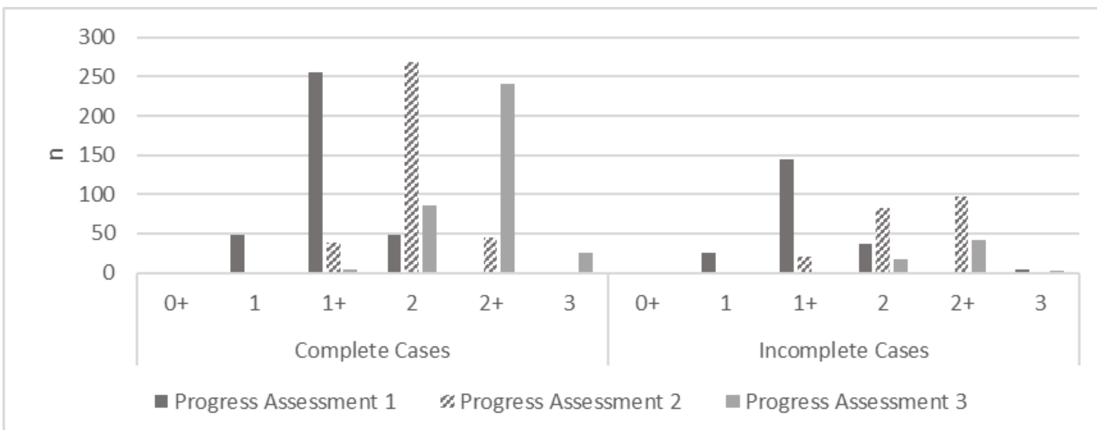

Figure 3. Speaking Progress Assessment ILR Scores

Table 16. Progress Assessment ILR Speaking Scores Descriptive Statistics

	Complete Cases						
	n	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
Progress Assessment 1	354	1.57	0.27	1.0	2.6	-0.63	1.20
Progress Assessment 2	354	2.04	0.26	1.6	3.0	0.99	1.62
Progress Assessment 3	354	2.47	0.31	1.6	3.0	-0.75	-0.18
	Incomplete Cases						
	n	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
Progress Assessment 1	212	1.63	0.33	1.0	3.0	0.87	4.62
Progress Assessment 2	201	2.25	0.37	1.0	3.0	-0.41	-0.85
Progress Assessment 3	61	2.45	0.29	2.0	3.0	-0.73	-0.80

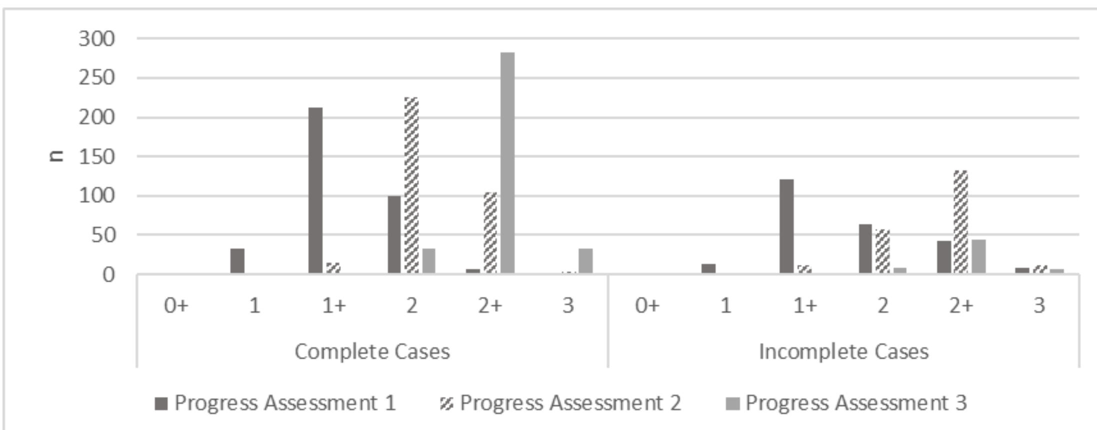

Figure 4. Reading Progress Assessment ILR Scores

Table 17. Progress Assessment ILR Reading Scores Descriptive Statistics

	Complete Cases						
	n	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
Progress Assessment 1	350	1.68	0.30	1.0	2.6	-0.13	1.39
Progress Assessment 2	350	2.17	0.31	1.0	3.0	0.44	-0.40
Progress Assessment 3	350	2.57	0.23	1.6	3.0	-1.30	3.47
	Incomplete Cases						
	n	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
Progress Assessment 1	250	1.88	0.47	0.6	3.0	0.46	0.10
Progress Assessment 2	215	2.40	0.37	1.0	3.0	-0.99	0.77
Progress Assessment 3	58	2.56	0.26	2.0	3.0	-0.95	1.39

Chapter 5: Methods

Structural Equation Models

Structural equation models (SEMs), also known as causal models or covariance structure models, simultaneously evaluate the relationships between multiple measured and estimated latent variables in *a priori* conceptualized causal systems. Structural equation models extend the capacities of simpler regression and factor analytical methods by integrating relationships and interrelationships between exogenous and endogenous variables. The relationships describing the data, or the model structure, are specified by the researcher and dependent on hypotheses made regarding the relationships between the data. Structural equation models estimate the strength and direction of the relationships between observed or latent variables hypothesized by the specified model.

The defining aspect and strength of structural equation modeling is its reliance on the variances and covariances of the observed data to estimate these relationships, rather than the data itself. Structural equation models are evaluated on their ability to find parameters that maximize the likelihood of these data (Hoyle, 2023, pp. 5-6). The closer that a covariance matrix produced using the model estimated by the specified model is to the observed matrix, the better the model is said to fit the data, and therefore the more accurately the hypothesized structure describes the actual relationships between the variables. Via an iterative process, the model fitting process estimates specified model parameters repeatedly to determine the parameter values that most closely resemble the variance covariance matrix of the observed data. This iterative process stops when the difference between the observed matrix and model-implied matrix no longer differs from one iterations to the next (typically when the difference is <.0001 of the calculated discrepancy value) (Finney et al., 2016, p. 137.).

Structural Equation Models and Causality

Due to the directional relationships and hypothesis-based nature of structural equation models, the output parameters from SEM methods may interpreted as "causal effects", thereby enabling researchers to corroborate claims of causal relationships between variables. This ability to make claims of causality allows individual differences and language's typological factors to be described as driver and barriers to student language learning and attainment in training, not just associations.

The tenability of these causal relationships rely not only on the standard H_0 assumption that a hypothesized relationship must be falsifiable, but on a system of empirical assumptions, model claims, and data which substantiate the system to some degree (Hoyle & Pearl, 2023, 49-50). In building an argument for the causal nature of SEM, Pearl (2009, pp. 157-158) describes it as an "inference method" which includes three inputs and produces three outputs. The three inputs consist of: a set of *A, assumptions*, which are qualitative causal claims which compose a model *M* (typically a path model); a set of calculable *queries, Q,* pertaining to "causal and counterfactual" relationships between variables; and a set *D* of *data* which are collected through methodology congruent with the scientific basis upon which *M* is built. Pearl describes the outputs: *A*,* a set of statements that are the logical but non numerical implications of *A*; a set of *C claims* supported by the data which denote the parameters of the relationships, such as variance or the magnitude of the effect of an intervention; and *T*, a set of testable implications of A. As described through this model, all queries and outcomes are dependent on the validity of the assumptions *A*. When building a model *M* of queries *Q* based on *A*, results are reported as *if A then C* for all hypothesized *C*. Thus, according to Pearl, the willingness to accept *A* indicates that *C* must also be accepted, allowing for the specified relationships to be described as causal.

Under Pearl's view, two other points must be noted: first, that the an overall goodness of fit test may not be fully informative regarding the validity of any specific *C*, due to inconsistencies of model sections not related to *C*; and second, that, especially in the social sciences, that unmeasured variables and relationships inevitably make up a proportion of the variance in the data, and that future models may supersede the model under observation. Both of these points only add support to the necessity of an evidence-based model with assumptions based in previously validated research.

Data-Model Fit Evaluation and Parameter Estimation

Overall fit of SEM analyses are estimated through the use of multiple output parameters. These measures are used to determine whether or how well the hypothesized structural model fits the observed data: i.e. if the hypotheses modeled in the structure were true about the relationships between the observed and latent variables, how well do the parameter estimates align with the actual observed variances, covariances, and means of the observed data? Three measures of overall fit of the specified models will be used in these analyses:

1) Chi-squared (χ^2) test of overall fit: this test evaluates "badness of fit", or whether the specified model does not significantly misfit the data more than a fully-saturated model. Contrary to the more commonly used χ^2 test of

independence, if the observed chi-squared test value does *not* exceed the critical value at the degrees of freedom, then the null hypothesis that the specified model fits as poorly as a fully-saturated model is rejected. The typical cutoff criteria for a chi-squared test is $p > .05$.

2) The root mean square error of approximation (RMSEA, Steiger & Lind, 1980): this test also estimates badness of fit. RMSEA estimates a noncentral chi-squared distribution, incorporating the degrees of freedom, normed, and taken as a square root. This value has a lower limit of 0, and no theoretical upper range. RMSEA at or below .06 indicates a "close" fit, and values of .08 or .10 are considered unacceptable (West et al., 2023, p. 190).

3) The comparative fit index (CFI, Bentler, 1990): this test measures "goodness" of fit, i.e. how well the specified model fits the data. This model relies on the same theory of a noncentral chi-squared distribution as RMSEA, but norms the formula to 0. CFI ranges from 0 to 1. Values >.95 indicate strong model fit, and >.90 are acceptable.

4) The standardized root mean square residual (SRMR, Bentler, 1995): this test measures "badness" of fit, and is calculated by converting residuals into a standardized value. SRMR has a lower limit of 0, and 1 is considered an unacceptable value. Values under 0.08 are considered acceptable.

Path Analysis Illustration

Path analyses are illustrated using diagrams that demonstrate the interrelated natures of the variables modeled. Variables that represent collected or measured data, also known as indicators, are represented by rectangles. Latent variables, which are hypothesized and estimated through the model rather than measured, are represented by circles. Relationships between these variables are indicated by arrows. Single-headed arrow indicate a unidirectional relationship, in which one variable is regressed on another (the source of the arrow). Two-headed arrows represent bidirectional relationships, in which the two specified variables are allowed to covary, and the estimated parameter represents the magnitude of their joint variability.

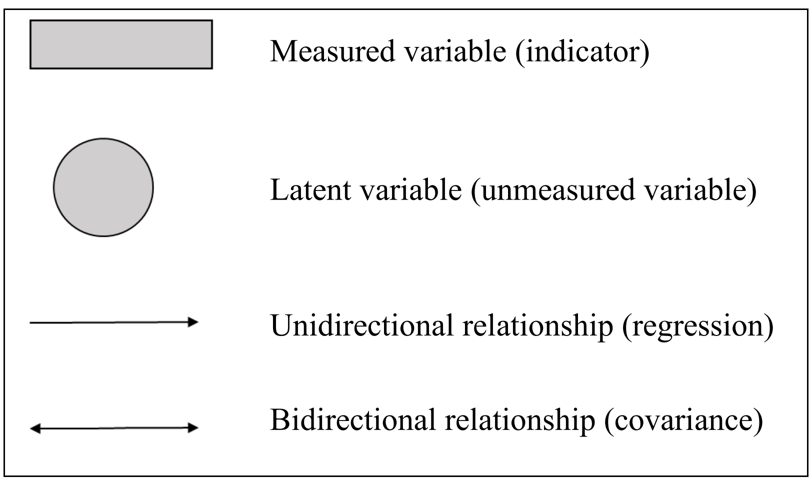

Figure 5 Conventional Path Analysis Illustration

Software

MPlus version 8 (Muthén & Muthén, 2017), will be used to fit the structural equation models below. MPlus is a code-based statistics software that allows for structural equation modeling using continuous and categorical variables such as those in this model and estimates and exports all parameter estimations and statistical test values necessary for determining whether findings support or fail to support the hypothesis described in the research questions. SPSS version 28 (IBM Corp, 2022) will be used when developing descriptive statistics and non-parametric measurements of variables such as means, standard deviations, medians, or reliability measures. Microsoft Excel (Microsoft Corporation, 2018) will be used for basic data cleaning and variable formatting and coding.

Missing Data: An Overview

Missing data occurs when values for variables in a dataset are not recorded in a dataset. Missing data can lead to biased parameter estimates, greater standard errors, and decreased generalizability of findings. The treatment of missing data in order to account for these issues depends on the nature of the “missingness”. Missing data is characterized in one of three ways: missing completely at random (MCAR); missing at random (MAR); and not missing at random (MNAR) (Rubin, 1976; Little & Rubin, 1989). MCAR is characterized by the missingness of the individual data values being unrelated to other values in the case and to the values of that the datapoint would have if the data were not missing. Provided there are enough complete cases of data, MCAR models can be fit using only complete data (i.e. listwise deletion of all incomplete cases) without affecting parameter estimates. An example of data MCAR would be if random pages in a series of test booklets stuck together, causing different students to miss different test items, which were unrelated to the students themselves or to whether they would have gotten the missed items correct. MAR data is characterized by its relationship to the case, and its missingness can be explained by other variables but isn't necessarily related to the actual values that are missing. For example, an undocumented person may skip items asking for their address: this data is related to the person's demographic characteristics and possibly to other survey responses such as what restaurant they most often eat at but not to the specific value of the missing datapoint (the address). MAR data is often considered *ignorable* due to the fact that if the variables that it is related to (such as immigration status) are controlled for, the missing data is explainable, and methods such as maximum likelihood estimation or multiple imputation can be used to produce reliable parameter estimates (Enders, 2013, p. 497).

In the most complex type of missing data, MNAR, the missingness is related to the values that the missing data would take on, and the missingness is related to the other variables. For example, in the current dataset, if the missing progress assessment scores were related to individual differences because some kinds of students were more likely to miss assessments, the data would be considered MNAR. MNAR data is capable of biasing parameters when the missing data is related to other variables in the model (Collins, Schafer, & Kam, 2001).

Methods of Accounting for Missing Data

MNAR data can be accounted for through the use of one of two processes: maximum likelihood estimation or imputation. These two approaches differ greatly in their treatment of missingness.

Maximum likelihood (ML) allows the model to be fit with incomplete data by allowing the size of the data matrix to vary based on the parameter being modeled: essentially every estimated parameter is based on all available data for that specific relationship, accounting for missing data by maximizing the ability of the model to use all available data rather than only complete cases (Allison, 2003, p. 548; Enders, 2013, p. 499-501). Multiple imputation does include several limitations. First, although ML methods do not impute data, by “marginalizing over” the missing data, it does create a situation Widaman (2006) referred to as “implicit imputation”, describing how the estimator’s use of a normal distribution assigned weights establishes a pseudo-imputed dataset to draw results from (Widaman, 2006, p. 49). Second, ML assumes that all incomplete values have a normal distribution, although simulations have demonstrated that this does not typically introduce bias (Enders, 2013, p. 227). Despite these limitations, ML is the most widely-used (Allison, 2001, p. 548) and “pragmatic” (Newson, 2023, p. 2) method of estimation for SEMs with missing data. Simulation studies have demonstrated that in MAR and MCAR datasets, ML generates unbiased parameters estimates and standard errors compared to deletion methods or imputation “[ML] estimation was superior across all conditions of the design” (Enders & Bandalos, 2001, p. 430).

Conversely, multiple imputation is a three-step process consisting of imputing multiple complete datasets, fitting the structural equation model on all complete imputed datasets, and then pooling the findings between the datasets to determine overall parameters and standard errors (Allison, 2003, p. 550; Enders, 2013, p. 507). Numerous other imputation methods exist: some simpler methods such as hot-deck or mean imputation utilize values from other cases or mean values of variables to complete the dataset for analysis, but these typically underestimate standard errors, biasing parameter estimates toward the non-missing data (Enders, 2013, pp. 499).

Multiple imputation has several notable limitations. First, imputation is by its very nature simulating a full dataset by utilizing only existing data to determine what the missing data would hypothetically look like. Thus, the imputation data depends on the accuracy, reliability, and representativeness of the non-missing data. The imputation relies heavily not only on the non-missing data quality but also on the assumption that the missing data values are related to the non-missing data in the relationships established in the I-phase, and are not better explained by excluded exogenous variables. Second, there is no manner of estimating whether the shape of the actual values of the missing data is accurately captured by the imputation process. Finally, no standard method of empirically assessing imputation results: any use of CFI or RMSEA fit statistics with imputed data are "ad hoc" (Enders, 2013, p. 513).

Characteristics of Missingness in the Study Data

The patterns of missingness in the data in this study are hypothesized to originate in differences in adherence record keeping practices between language programs, thus consisting of MNAR data. Figure 6 and Figure 7 illustrate the distribution of missing data among language programs: as seen, in both speaking and reading, the data follows a roughly bimodal distribution. The greatest number of language programs demonstrate complete or near complete data across all student records (<95% fully complete cases), while many programs have complete data records for 80% or fewer of their students. This indicates that programs tend to either record all or nearly all of their progress score data, or they record data irregularly, indicating that the missingness is likely related to the language program, and thus missingness in the dataset is associated with the L2 studied. A chi squared test indicated that the proportions of cases in the speaking and reading datasets that were complete and incomplete differed between languages ($\chi^2(41, N = 618) = 327.77, p \leq .01; \chi^2(41, N = 649) = 250.71, p \leq .01$). Due to the number of languages, no post-hoc test was performed to determine which languages differed from each other in respect to missingness of data.

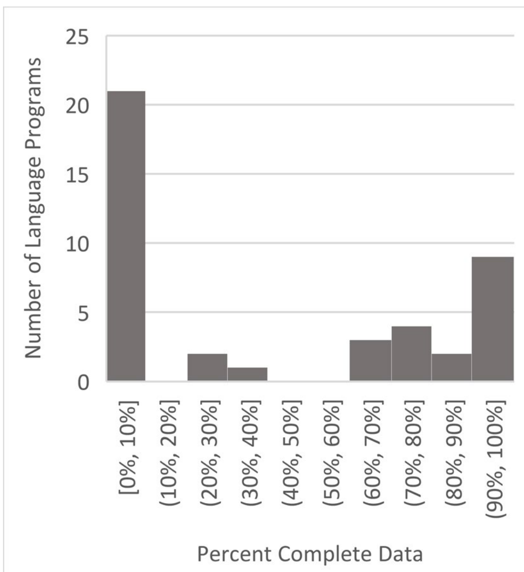

Figure 6. Frequency of Percents of Missing Data in Speaking Data by Language Program

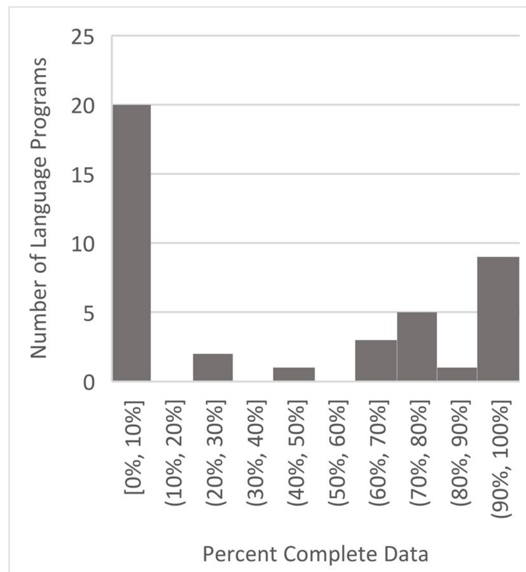

Figure 7. Frequency of Percents of Missing Data in Reading Data by Language Program

In an analysis of the representativeness of the cases by individual differences, no relationships were found between individual differences and missingness. Across the individual difference variables in the reading dataset, the proportions of cases in the speaking dataset that were complete and incomplete did not differ by MLAT ($t(543) = -1.26, p = .21$), previous test ($\chi^2(1, N = 618) = 1.98, p = .15$), first tour ($\chi^2(1, N = 618) = .75, p = .39$)), or role at post ($\chi^2\ (1, N = 618) = 1.36, p =$

.24)). Similarly, the proportions of cases that were complete and incomplete in the reading dataset did not differ by MLAT ($t(562) = -.74, p = .46$), previous test ($\chi^2(1, N = 649) = .07, p = .79$)), first tour ($\chi^2(1, N = 649) = .35, p = .55$)), or role at post ($\chi^2(1, N = 618) = 3.77, p = .06$)). This indicates that the missingness is likely not associated with the individual differences of the learners.

In the speaking dataset, 42.71% (264 of 618 cases) had at least one missing progress assessment score; in the reading data set 46.10% (299 of 649 cases) had at least one missing progress assessment score. The variables missing from the data consist of progress assessment scores: all other data (individual differences, typological characteristics, and end of training scores) are complete. Figure 8 displays the patterns of missingness in the three progress assessments across three progress assessments in speaking and reading.

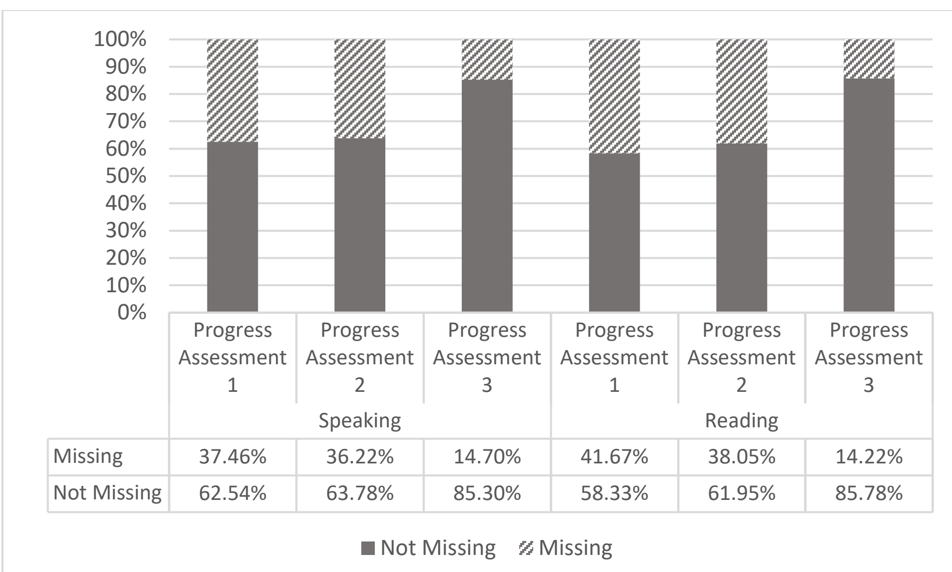

Figure 8. Distribution of Missing Data in Speaking and Reading Progress Assessment Scores

Individual Differences in Language Proficiency Gains

Structural Equation Path Analytic Model

In order to respond to the research questions, multiple path-analysis models will relate student-level cognitive and language learning background data to the end of training speaking and reading proficiency results, respectively.

The first model, shown in Figure 9. Model 1: Individual Differences Path-Analysis Diagram, will utilize direct causal relationships from each of the student individual difference variables indicators to the three formative progress assessments and the EOT proficiency. As described in the previous literature review and study motivation chapters, learner aptitude, previous study of languages, and specialization of job (as a proxy for motivation), have all been shown by previous research to have causal relationships with EOT language learning outcomes. The significance of each of these paths will be considered evidence of a causal relationship between the exogenous individual differences and the endogenous proficiency assessment scores. The model will be conditioned on all individual differences.[2] Each of these exogenous individual difference variables is modeled to have a unidirectional causal relationship with the three progress assessments and the final score, the learners' ultimate achievement during training to demonstrate the individual differences' impact throughout the language learning process.

[2] An initial model proposed to fit this data covaried *MLAT* with *Previous Test* and *Previous Test* with *First Tour.* Consultation with a structural equation model specialist during the initial stages of model estimation and additional reviews of the literature revealed that covarying exogenous variables alters model calculation, and that they are thus considered endogenous with additional distributional assumptions (G. Hancock, personal communication, May 28, 2024; Muthén, 2011). In order to maintain the assumption that the individual variables are exogenous variables (Muthén, 2011) and enable independent estimation of the paths of interest, the covariances were not specified in this manner (G. Hancock, personal communication, May 28, 2024; Muthén, 2011).

As each subsequent measure of proficiency depends on the progress made at the times of previous assessments, progress assessment 1 (PA 1) is specified to predict progress assessment 2 (PA 2), and progress assessment 2 (PA 3) is specified to predict progress assessment 3. Because each of the individual formative progress assessments build toward the final end of training score, each is hypothesized to have a unidirectional causal relationship to the final assessment. This structure, in which the individual differences are specified to have a unidirectional causal path to both progress assessment scores and to the end of training scores and the progress assessment scores are specified to have unidirectional paths to the end of training scores, establishes a mediating relationship in which the progress assessment scores mediate the relationship between individual differences and the end of training score. Often the establishment of a mediating model is to determine whether the mediating relationship explains enough of the variance of the endogenous variable that a previously significant direct relationship between the exogenous variable and the endogenous variable is no longer significant (Baron & Kenny, 1986, p. 1176). However, in this case, the advantage of the use of a mediating model is to be able to estimate the total effect, direct and indirect, of the individual difference over both the proficiency growth throughout training and the ultimate proficiency attainment while taking into account the cumulative nature of language acquisition, in which previous stages of language learning impact future stages.

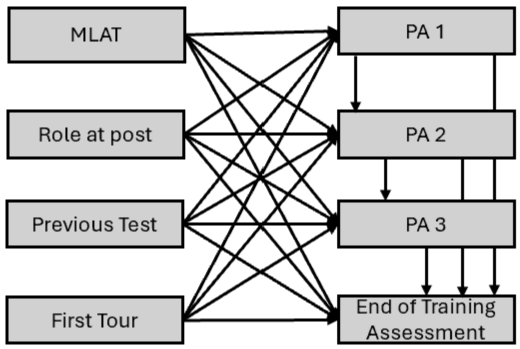

Figure 9. Model 1: Individual Differences Path-Analysis Diagram

Model Estimation

Model 1 will be fit in both reading and speaking modalities. In both modalities, a rough sensitivity analysis will be performed after fitting the specified model using both the dataset with missing data with an ML estimator, and the dataset without missing data with an ML estimator. An analysis of the fit indices will be performed to determine whether there is evidence as to whether the model with or without incomplete cases fits the data more closely. While the two models cannot be directly compared using fit indices due to their non-nested nature (Kline, 2023, p. 182), differences in the acceptability of the two models will be noted. For example, if both estimation methods demonstrate acceptable fit statistics across both modalities both reading and speaking will be fit including the missing cases to utilize the stronger estimation method (Allison, 2001, p. 548; Newson, 2023, p. 2; Enders & Bandalos, 2001, p. 430). However, if in either modality the casewise deletion method demonstrates acceptable fit indices and the ML estimator demonstrates unacceptable indices, casewise deletion will be used.

Model Identification and Power

Path analytic models are defined as "identified" if the specified model produces a single set of estimates for each parameter given the known values, i.e. the variances and covariances of the variables in the input matrix (Brown, 2023; Alhija, 2010). Once a model is specified, model identification can be confirmed *a priori* by confirming that the number of parameters in the model is not greater than the degrees of freedom of the model. A model is overidentified if the number of elements of the input matrix, i.e. the number of known values prior to model estimation, is greater than the number of unknown elements, i.e. the estimated variances, covariances, error variances, and error covariances.

The number of known values in a model, i.e. the number of observations, is calculated by multiplying the number of observed variables (v) times the number of variables plus one, and dividing by two, i.e. $(\mathrm{v}(\mathrm{v}+1)/2)$. Equation 1, below, illustrates the calculation of this value for the model in Figure 9, above. The total number of observations for Model 1 is 36.

$$\boxed{\frac{v(v+1)}{2} = \frac{8(8+1)}{2} = 36}$$

Equation 1. Number of Observations in the Path Analytic Model 1

The number of estimated parameters is calculated through summing the number of exogenous variables (4), unidirectional (29) and bidirectional (6) relationships (not shown: model-implied covariances of exogenous variables, (Muthén, 2011)), and endogenous variables (4) in the model in Figure 9, above. This sum of 35 describes the number of unknown values to be calculated by the model estimation.

The degrees of freedom in a path analytic model is the difference between the number of observations and the number of estimated parameters in the model. As seen in equation 2, 1 degrees of freedom is associated with Model 1. Because the value is greater than 0, Model 1 is *overidentified*, enabling estimation of model goodness of fit and determination of how well the model is able to estimate the relationships between the observed variables in the data set (Brown, 2023, pp. 264-265).

$$\boxed{number\ of\ observations - number\ of\ estimated\ parameters = 36 - 35 = 1}$$

Equation 2. Degrees of Freedom in the Path Analytic Model 1

The overall model fit will be defined using multiple fit indices. First, overall model fit will be assessed by determining whether the specified mode produces a chi squared value of over $p = .05$, an RMSEA at or below .06, a CFI of .90 or above, and an SRMR of less than 0.08.

No single standard for determining sample sizes in SEM exists in the literature. Standard "rule of thumb" recommendations range from a minimum of 100-200 cases (Boomsma, 1985) to 10 cases per observed variable (Nunnally, 1967). As this research makes use of a predetermined data set with 354 complete speaking cases and 350 complete reading cases, it easily meets these standards. Although these estimates are widely utilized, Wolf et al. (2013) provides a more accurate and evidence-based estimate of impact of SEM sample sizes on power to detect effects of different sizes. Through the use of Monte Carlo modeling with an alpha level of .05, Wolf et al. estimated that with the given sample sizes, a SEM would have the power of .86 to detect small direct effects (r =.25) and power of .62 to detect small indirect effects. The given sample sizes would have the power of .99 to detect large direct and indirect effects (r=.85) (Wolf et al., 2013, Data Supplement, Table 4).

Model Interpretation: Research Question 1

Research Question 1 (RQ1): Do L2 proficiency assessment scores throughout intensive language study differ between adult L2 learners by aptitude, language learning background, motivation, and previous experience with curricula?

Statistically significant path coefficients from exogenous individual differences variables to endogenous progress assessment scores will be considered evidence of direct effects of individual differences on language proficiency attainment at points throughout training. Statistically significant total effects of each individual difference on the end of training scores will be considered evidence of the influence of individual difference variables on the ultimate attainment in language training: the total effect is sum of the total of the path coefficients from a single exogenous variable to an endogenous variable, including path coefficients from the exogenous variable to mediating variables and from these mediating variables to the specified endogenous variable. For example, the total effect from MLAT to the end of training score includes the path directly from MLAT to the end of training score, the paths from MLAT to each of the progress assessments, the paths from these progress assessments to the end of training score, and the paths from progress assessment one to two and two to three. This value describes the total effect MLAT on ultimate achievement, taking into account the mediation caused by differences in language learning patterns. Reporting this in addition to the direct effect from the individual differences explains the expected influence of the individual differences within the context of a student's growth throughout the training period.

Hypothesis 1

The model is expected to converge, or fit the data to a statically acceptable degree, using the model fit indices described previously. The individual difference of MLAT score is expected to significantly positively predict language learning at all assessments, given previous similar findings in comparable analyses (Masters, 2018; Rhoades, 2023, inter alia). MLAT may be especially predictive of progress in the first progress assessments (1 and 2), given its association with predicting intensive language learning in the early stages (Li, 2015; Linck et al., 2013). A previous test in a language is expected relate most strongly to the first progress assessment, and not to subsequent assessments, as L2 interference in L3 development lessens. Significant negative directionality of this relationship will be explained as indicating L2 interference. Job specialization is hypothesized to relate negatively to the later progress assessments and the end of training assessment, as the language functions and tasks in the curriculum advance from general greetings to specific content less related to non-specialized students' future work. Finally, first tour learners are expected to perform more poorly on their progress assessments, as their unfamiliarity with the LTF curricula, future work, and progress assessment specifications leave them at a disadvantage compared to students who have studied at the LTF previously, but they are expected to "catch up" and demonstrate no significant relationship to end of training scores.

Model Interpretation: Research Question 2

Research Question 2 (RQ2): Does the influence of individual differences on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

A qualitative analysis will compare the significance of the path coefficients found in the respective reading and speaking models responding to research question 1. This will first consist of a comparison of the significance of the direct effects of each exogenous individual difference variable on speaking versus reading progress assessment scores. Differences between the significance of the path coefficients in the reading and speaking models will be considered differences in the patterns of L2 proficiency growth due to individual differences. Furthermore, if both speaking and reading paths are significant, the signs of the paths will be compared to observe whether they are similar or different.

Similarly, a qualitative analysis will compare the significance of the total indirect effects from individual differences to end of training scores found in the respective reading and speaking models responding to research question 1. Differences between the significance of the total indirect effect in the reading and speaking models will be considered differences in the relationships between individual differences and ultimate achievement. Furthermore, if both speaking and reading paths are significant, the signs of the paths will be compared to observe whether they are similar or different.

Due to the separate models fit for reading and speaking, direct comparison of the parameters through inferential tests is not possible. Observed differences will rely on comparison of whether or not parallel causal paths between the two models are both significant, if both are not, or if there is a difference in their significance. If both are significant, comparison of the parameter values will consist of comparison of their signs, for example, if both are positive, both are negative, or if one is positive and one is negative.

Hypothesis 2

No significant differences between speaking and reading Model 1 path coefficients' significance nor directionality of significant coefficients is expected between reading and speaking. In all complete datasets in the study, reading and speaking LTF benchmarks expectations for speaking and reading skills are the same for progress assessments 1, 2, and 3, and all learners in the datasets have end of training proficiency goals of ILR 3 in reading and speaking, indicating that students have historically gained proficiency in both modalities at similar rates and in similar patterns at the LTF. Three reasons are hypothesized to account for this observed parallel pattern in LTF assessments scores. First, reading is scored based on an oral description of the examinee's understanding of written passages, and the oral production skills used in this recitation may overlap with the oral production skills used in the speaking assessments. Second, the progress assessments and end of training assessments are administered at the same time and often by the same staff members. This opens the possibility for a "halo effect" (Thorndike, 1920, p. 27), a phenomenon in which raters score an individual on scales that are supposed to measure separate constructs, but the measurement given on one scale affects the other, often in a positive direction. Halo effects have been observed in L2 evaluation, for example, O'Grady found halo effects between complexity, accuracy, and fluency ratings of English language learners when the constructs were scored by individual raters versus by a single rater (O'Grady, 2023, p. 6-7), and Bechger et al. found halo effects in raters' scoring of aspects of L2 Dutch speech production that were not present when different raters scored the test, calling the scores of different constructs scored by the same raters so similar that they were "exchangeable" (Bechger, 2010, p. 613). This effect could sway rater's opinions of the student's ability on the two scales, causing them to be scored more similarly than they would be if scored independently, potentially without their even realizing it (Wu et al., 2013, p.281). Third, the ILR scale has relatively broad "bands" of proficiency. Thus, a single score band encompasses a range of proficiency levels.

Individual Differences and Difficulty in Language Proficiency Gains

The second path analysis, Model 2, will build on the first to investigate L2 difficulty's influence on proficiency attainment throughout training within the context of individual language learner differences. This analysis is exploratory in nature, as previous research has not included these individual differences and difficulty in a simultaneous model. Model 2 is a more complex version of Model 1, as it adds variables and directional relationships between variables to Model 1 but does not change any of the specifications already in place in Model 1. This makes it comparable through the use of omnibus model fit statistics.

The relationships between student exogenous individual differences and endogenous assessment scores will be specified to be the same as in Model 1. The variable describing the difficulty of the language studied by each student will be used as a moderating variable on the relationships between exogenous student individual differences and endogenous proficiency scores in both the speaking and reading models. A moderating variable is "a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (Baron & Kenny, 1986, p. 1174). The use of a variable in this manner establishes a relationship in which the relationship between the independent exogenous variables and the endogenous variables is moderated, or altered by a third variable. In this case, the moderating variable is the language's difficulty. For example, Figure 10 illustrates a model in which a language's difficulty moderates the relationship between MLAT score and progress assessment 1. In SEM, a moderating relationship is modeled by multiplying the moderating variable by the exogenous variable, and using the exogenous variable, moderating variable, and calculated product variables as predictors of the endogenous variables (Hayes, 213, p. 209-213). Figure 11 illustrates the SEM specification of the relationship in theorized in Figure 10.

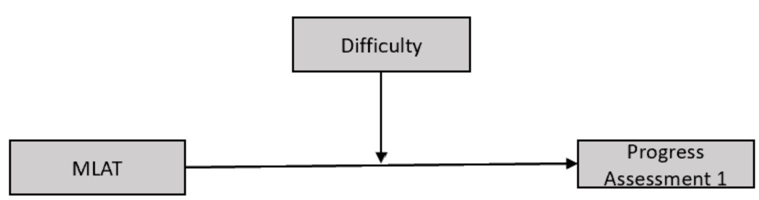

Figure 10. Theoretical Visualization of a Moderating Relationship

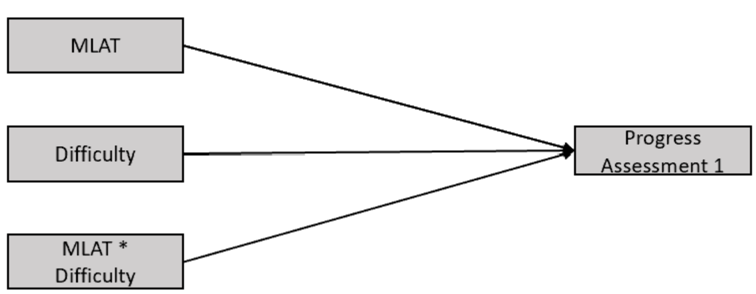

Figure 11. Structural Equation Model of a Moderating Relationship

Figure 12 illustrates Model 2, the structural equation model utilizing the individual different variables, L2 difficulty, and interaction variables as exogenous variables predicting the progress assessment scores and end of training proficiency score. As in Model 1, the significance of each unidirectional path from an endogenous individual difference variable to a progress assessment score or end of training score will indicate a significant relationship between the endogenous and exogenous variables. Significance of the path from a product variable to an exogenous variable indicates the presence of a moderating effect of the specified typological variable on the relationship between the individual difference and the endogenous assessment score (Hayes, 2013, p. 253). A moderating effect may be present with or without significance of either of the two exogenous variables it involves (the individual difference and difficulty variables, in this case): the significance of these two variables "are not directly related to testing the moderator hypothesis" (Baron & Kenny, 1986, p. 1174). In order to provide context, the significance of the exogenous variables that make up the moderator variable will be noted for any significant moderating effects found.

In order to model a moderating effect with interpretable outcomes, an alteration must be made to the variables. The continuous *MLAT* and *Difficulty* variables will be centered prior to model fit. Continuous exogenous variables without meaningful zeros (such as *MLAT* and *Difficulty*) are centered in order to increase interpretation: when centered, the coefficient of the interaction term describes the effect of the moderating variable on a case with the mean value (Shieh, 2011). For example, if the path from *Previous Test*Difficulty* to *Progress Assessment 2* is significant, then the coefficient is the difference between the expected score of a student who had taken a LTF language test prior to starting their studies and one who had not, with all other individual differences and typological factors held constant.

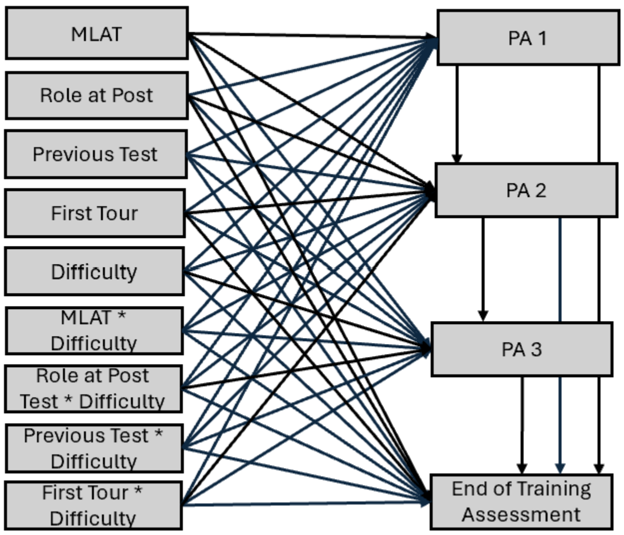

Figure 12. Model 2: Individual Differences and Typological Factors in Path-Analytic Diagram

Model Identification and Power

The number of known values in a model, i.e. the number of observations in is calculated by multiplying the number of observed variables (v) times the number of variables plus one, and dividing by two, i.e. $(v(v+1)/2)$. Equation 3, below, illustrates the calculation of this value for the model in Figure 5, above. The total number of observations for Model 2 is 91.

$$\boxed{\frac{v(v+1)}{2} = \frac{13(13+1)}{2} = 91}$$

Equation 3. Number of Observations in the Path-Analytic Model 2

The number of estimated parameters is calculated through summing the number of exogenous variables (9), unidirectional (41) and bidirectional (2) relationships (illustrated with arrows), and endogenous variables (4) in the model in Figure 12, above. This sum of 56 describes the number of unknown values to be calculated by the model estimation.

The degrees of freedom in a path analytic model is the difference between the number of observations and the number of estimated parameters in the model. As seen in Equation 4, 35 degrees of freedom are associated with Model 2. Because the value is greater than 0, Model 2 is *overidentified*, enabling estimation of model goodness of fit and determination of how well the model is able to estimate the relationships between the observed variables in the data set (Brown, 2023, pp. 264-265).

$$number\ of\ observations - number\ of\ estimated\ parameters = 91 - 56 = 35$$

Equation 4. Degrees of Freedom in the Path Analytic Model 2

As with Model 1, the overall model fit will be defined using multiple fit indices. First, overall model fit will be assessed by determining whether the specified mode produces a chi squared value of over $p = .05$, an RMSEA at or below .06, a CFI of .90 or above, and an SRMR of less than 0.08.

As mentioned previously, no single standard for determining sample sizes in SEM exists in the literature. Standard "rule of thumb" recommendations range from a minimum of 100-200 cases (Boomsma, 1985) to 10 cases per observed variable

(Nunnally, 1967). As this research makes use of a predetermined data set with 354 complete speaking cases and 350 complete reading cases, it easily meets these standards. The Wolfe et al. (2013) study used to describe the ideal sample size for Model 1 does not apply in this case, and Wolfe et al. did not present a similar projection of necessary sample size for the more complex Model 2 used in this model, so a precise estimation of power is not obtainable through this source. However, the data does still fit the minimum cases estimated by Boomsma (1985) and Nunnally, (1967), indicating that the model can be expected to converge with interpretable results, though potentially with less power to detect small effects due to the increased model complexity.

Model Interpretation: Research Question 3

Research Question 3 (RQ3): Does language difficulty moderate the relationships between individual differences and L2 proficiency assessment scores throughout intensive language study? If so, which points in learning are affected and in what ways?

The first step in the evaluation of Model 2 is establishing whether it explains the data above and beyond what was explained in Model 1.

Comparison of the fit indices of Model 1 and Model 2 will determine whether the typological characteristics of the respective L2s explain variance in second language proficiency growth and ultimate attainment above and beyond what individual differences explain. Models 1 and 2 are not "nested" by the SEM definition of utilizing the same datasets and variables with different parameters (Kline, 2023, p. 182), and cannot be traditionally compared using traditional fit statistics, because Model 2 utilizes additional variables (the difficulty and the moderating variables). A method of simulating nested models is to fit a model utilizing all variables of the second fuller model, but fixing all paths that are not included in the base model to 0 (Muthén, 2010). This will produce fit indices which are able to be used to compare Models 1 and 2.

A typical chi-squared difference test of nested models in which the difference of the chi-squared value is tested again the difference in degrees of freedom cannot be used with maximum likelihood with robust standard errors estimated SEMs (Satorra & Bentler, 2010), because the difference in chi-squared values between the two models can be negative. To account for this, Satorra and Bentler established an alternate test relying on the log likelihood of the nested and non-nested models. The test utilizes the loglikelihood of both models ($L_{Model\ 1}$, $L_{Model\ 2}$), the Satorra-Bentler scaling correction factor ($c_{Model\ 1}$, $C_{Model\ 2}$), and the number of free parameters ($p_{Model\ 1}$, $p_{Model\ 2}$) to calculate and test a re-scaled chi-squared test statistic (Satorra & Bentler, 2010, p. 6). The test statistic is calculated using a two-step process, first calculating the scaling correction factor (cd), then the test value (TRd):

Equation 5. Calculation of Satorra Bentler Test Statistic (reproduced from Satorra & Bentler, 2010, p. 7)

$$cd = \frac{(p_{Model\ 1} * c_{Model\ 1} - p_{Model\ 2} * C_{Model\ 2})}{(p_{Model\ 1} - p_{Model\ 2})}$$

$$TRd = -2(L_{Model\ 1} - L_{Model\ 2})/cd$$

The test statistic (TRd) follows a chi-squared distribution and is tested against $p_{Model\ 1} - p_{Model\ 2}$ degrees of freedom.

The difference in Bayesian Information Criterion (BIC) will be considered as a measure of the strength of evidence of the difference between the models: BIC can be used to compare non-nested models with the same endogenous variables (Muthén, 2015), which is the case in this comparison. Utilizing the standards established by Kass and Raffery (1995, p. 777), the model with the smaller BIC value will be considered stronger. BIC is not a single-model fit statistic, but can be used to estimate the "relative probability of competing hypotheses" of two models though analysis of likelihood of models (Preacher & Yaremych, 2023, p. 212). Table 18 describes the strength of evidence that the difference in BIC between two models demonstrates.

Table 18. Evidence Provided by Difference in BIC (reproduced from Depaoli, et al., 2023, p. 711)

BIC difference	Bayes factor	Evidence against M_2
0 to 1	1 to 3	Weak
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Statistically significant path coefficients from exogenous individual differences variables or the difficulty variable to the endogenous progress assessment scores will be considered evidence of direct effects of individual differences or difficulty, respectively, on language proficiency attainment at points throughout training. Statistically significant total effects of each individual difference or difficulty variable on the end of training scores will be considered evidence of the influence of individual differences or difficulty, respectively, on the ultimate attainment in language training: the total effect is sum of the total of the path coefficients from a single exogenous variable to an endogenous variable, including path coefficients from the exogenous variable to mediating variables and from these mediating variables to the specified endogenous variable.

Significance of the path coefficient from the mediating difficulty variable (the product of the seven-point difficulty scale and individual difference variable) to a progress assessment scores will be considered evidence of the mediating effect of the difficulty on the relationship between the individual difference and language proficiency growth at that point in training.

Significance of the path coefficient from a mediating variable to an end of training score will be considered evidence of a mediating effect of difficulty on the relationship between the individual difference and the end of training score.

Model Interpretation: Research Question 4

Research Question 4 (RQ4): Does the influence of individual differences and difficulty on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

A qualitative analysis will compare the significance and direction of the path coefficients found in the respective reading and speaking models responding to research question 4. This will first consist of a comparison of the direct effects of each exogenous individual difference variable on speaking versus reading progress assessment scores. Differences between the significance of the path coefficients in the reading and speaking models will be considered differences in the patterns of L2 proficiency growth due to individual differences: if both are significant, directionality of the relationships will be compared.

Similarly, a qualitative analysis will compare the significance and direction of the total indirect effects from individual differences to end of training scores found in the respective reading and speaking models responding to research question 3. Differences between the significance of the total indirect effect in the reading and speaking models will be considered differences in the relationships between individual differences and ultimate achievement: if both are significant, directionality of the relationships will be compared.

Finally, a qualitative analysis will compare the significance and direction of the path coefficients of the mediating variables between the reading and speaking models. Differences between the significance of the total indirect effect in the reading and speaking models will be considered differences in the relationships between individual differences and ultimate achievement: if both are significant, directionality of the relationships will be compared.

Due to the separate models fit for reading and speaking, direct comparison of the parameters through inferential tests is not possible. Observed differences will rely on comparison of whether or not parallel causal paths between the two models are both significant, if both are not, or if there is a difference in their significance. If both are significant, comparison of the parameter values will consist of comparison of their signs, for example, if both are positive, both are negative, or if one is positive and one is negative.

Hypothesis 3

The specified Model 2 is expected to explain patterns of L2 growth and above and beyond what Model 1 explained in both the speaking and reading models for each language characteristic. This is because the difficulty of the L2s are known to influence language learning (Cysouw, 2013, see Language Difficulty section above for review).

Several path coefficients describing the relationships of the moderating variables to the progress assessments and end of training scores, respectively, are hypothesized to significantly moderate the relationships between individual learner differences. Each is expected to reflect the negative impact language learning at different stages and to relate to reading and speaking skill acquisition differently.

In Model 2, the individual learner differences are expected to have the same relationships to progress assessment scores and EOT assessment scores as in Model 1, although the coefficient values are expected to shift with the addition of the difficulty and moderation variables are added. MLAT is expected to have significant positive effects throughout training, most notably at PA 1 and 2. *Role at Post* is expected to have significant negative effects later in training, at PA 2, 3, and EOT. And *Previous Test* and *First Tour* are expected to have significant negative effects early in training at PA 1, but are expected to have fewer significant effects on later assessment scores.

Language difficulty and the moderating variables representing interactions between individual differences and difficulty are expected to have significant impacts on assessment scores throughout training. As language difficulty is defined by the expected ILR proficiency after 24 weeks in training for a learner in a specific language, lower values are associated with "harder" languages. The expected results throughout training of positive coefficients associated with the path from the difficulty variable to each assessment score indicate that learners studying less difficult languages are expected to score higher on each assessment. Language difficulty is known to negatively relate to achievement of end of training proficiency goals in federal language training (Bermudez-Mendez, 2020; Wagener, 2016). Language difficulty is also significantly related to proficiency assessments given immediately post-training indicating that language training outcomes were related to difficulty (Rhoades, 2023).

The interaction of aptitude as measured by MLAT and language difficulty is not expected to be significant, or is expected to be minimal. A review of literature found several studies concluding that aptitude and difficulty, although both significant predictors of proficiency gains, explained unique variance in L2 assessments (Bokander, 2020, p. 37). In a similar federal language training situation, Mackey found no interaction between aptitude and language distance among federal language students in predicting language proficiency growth (2023, p. 204).

The influence of L2 difficulty on the relationship between *Role at Post* and the assessment scores is considered exploratory in the current investigation. As *Role at Post* represents the alignment of tasks studied during training and language use goals, and is considered to be a rough proxy for motivation, it is hypothesized that the increased stress and complexity of studying a more difficulty language will curtail this motivation among learners who are already less motivated due to lesser alignment between their studied tasks and ultimate L2 use goals. Thus a significant moderation of the causal path between *Role at Post* and the assessments is hypothesized. This effect is most strongly expected in the second and third progress assessments and at EOT, when the tasks studied become more complex and more closely related to the work at post performed by non-specialists.

The interactions of difficulty on the relationships that *Previous Test* and *First Tour* have with assessment scores are particularly exploratory, as a literature review did not find previous research relating previous language learning experiences and subsequent language difficulty to proficiency outcomes at different difficulty levels. It is hypothesized that learners who have taken a test previously at TLF would be less influenced by difficulty, as they would be ablet to rely on the cognitive benefits of previous L2 study, leading to significant moderator coefficients in earlier progress assessments. It is also hypothesized that students who are on their first tour of study at LFT will not have the ability to rely on familiarity with the curricula and assessment methods, and will be more influenced by difficulty, leading to significant coefficients at early progress assessments.

The path coefficients and total path indirect path coefficients from individual difference variables are not hypothesized to differ in significance between reading and speaking.

Chapter 6: Individual Differences in Language Proficiency Gains Results

Estimator Selection

The first step of the analysis was the decision regarding whether or not to incorporate missing data using maximum likelihood estimation, or to exclude them from the analysis. The primary step was a rough sensitivity analysis, determining whether there was a rationale for excluding cases with missing data through listwise deletion or for including them through the use of a ML estimator.

Speaking Estimator Evaluation

The initial ML model utilizing all data, including missing cases, demonstrated fair model fit: ($n = 545, CFI = 0.971, RMSEA = 0.168, \chi^2 = 16.389, \mathrm{p} \leq .001$, $\mathrm{SRMR} = 0.027$). The model utilizing only complete cases also exhibited fair model fit ($n = 354, CFI = 0.970, RMSEA = 0.187, \chi^2 = 13.347, \mathrm{p} \leq .001, \mathrm{SRMR} = 0.029$). Although the χ^2 and RMSEA values were not at ideal levels under either estimation method, CFI and SRMR were at acceptable levels, and were prioritized (Shi et al., 2019, p. 330; Asparouhov & Muthén, 2018, p. 4), and the models were considered acceptable.

Reading Estimator Evaluation

The initial ML model utilizing all data, including missing cases, demonstrated fair model fit: ($n = 563, CFI = 0.998, RMSEA = 0.036, \chi^2 = 1.722, \mathrm{p} = 0.1895, \mathrm{SRMR} = 0.014$). All fit indices were acceptable. The model utilizing only complete cases exhibited stronger model fit ($n = 350, CFI = 0.999, RMSEA = 0.026, \chi^2 = 1.245, \mathrm{p} = 0.2645, \mathrm{SRMR} = 0.010$).

Estimator Decision

In both reading and speaking, both casewise deletion and maximum likelihood methods demonstrated acceptable model fit statistics. The determination was made to not exclude the cases with missing progress assessment data from the set and to utilize the maximum likelihood estimator.

Individual Differences in Language Proficiency Attainment

Research Question 1 (RQ1): Do L2 proficiency assessment scores throughout intensive language study differ between adult L2 learners by aptitude, language learning background, motivation, and previous experience with curricula?

Speaking Model 1 Results

As previously described, speaking Model 1 fit the data acceptably well. Altogether, the model explained 42.7% of the variance in EOT scores ($R^2 = .427$), a higher percent than what Masters found in her similar study explaining speaking end of training exams at DLI using a qualifying exam score, aptitude measure, and language preference (2018, p.89). Figure 13, below displays the causal paths in the Model 1 speaking results using black arrows; gray arrows represent non-significant causal paths.

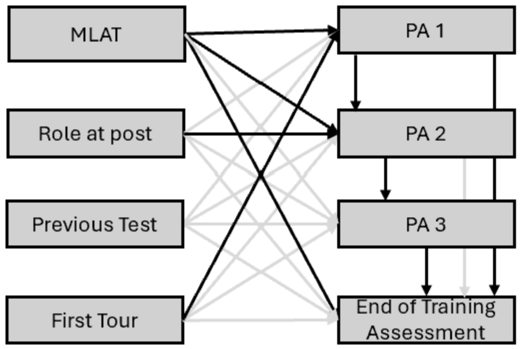

Figure 13. Model 1: Speaking Results

Relationships of Individual Differences to Speaking Proficiency Assessment Scores

The most notable individual differences were MLAT, *Role at Post*, and *First Tour*: MLAT scores significantly predicted the first and second progress assessment scores, as well as the end of training assessment score. *Role at Post* was significantly and negatively associated with scores on the second progress assessment, and *First Tour* was positively associated with scores on the first progress assessment. Table 19-Table 21, below, display the non-standardized and standardized causal path coefficients, the standard errors of the path coefficients, and the significance level of the coefficients for the speaking Model 1 results. The following section discusses these results. A qualitative summary of these results will be presented as part of the response to research question 4 (Table 25).

The most consistent individual difference was aptitude as measured by MLAT: MLAT scores significantly predicted the first and second progress assessment scores, as well at the end of training assessment score. For every standard deviation larger a student scored on the MLAT (8.86 points), the learner could be expected to score 0.174 higher on PA 1, 0.123 higher on PA 2, and 0.124 higher on the EOT assessment, controlling for all other individual differences. By the end of training, the total effect of MLAT on the EOT assessment was an expected 0.261 points higher for every 8.86 points higher on MLAT. This means that, with all other variables held constant, a 38.46 point difference in MLAT scores would predict a proficiency difference of 0.5 on the ILR scale, or a half-level increase on the EOT test, taking into consideration effects patterns in the student's proficiency gains throughout training. This positive effect throughout is as hypothesized and the impact on learning at earlier stages aligns with previous research by Li (2015) and Linck et al. (2013). *Role at Post* was negatively associated with performance on PA 2: students who would be entering jobs with specialized roles could be expected to score 0.078 points lower than students with non-specialized roles, controlling for all other individual differences. The hypothesis associated with *Role at Post* was an expected lower score for specialized students, especially toward the end of training: while the direct effects of role at post on PA 3 and EOT scores were not significant, the total effect of role at post on the EOT score neared significance, with 0.099 lower scores expected for specialized students. Having a *Previous Test* on record was expected to be most strongly associated with PA 1 scores, but was not significantly associated with any assessment score, nor did it exhibit a significant total effect on the EOT assessment. Finally, *First Tour* learners were hypothesized to perform more poorly on their progress assessments, but students in this population who studied at the LTF for language training for the first time were more successful at their first PA than returning students: they could be expected to score .112 points higher on their first EOT test than non-first time learners. However, this pattern did not hold out over the length of training: there was no significant direct effect on any following assessment, nor was the total effect of being a first-tour learner at LTF statistically significant.

The assessment scores over time were closely related to each other (Table 20). PA 1 performance predicted PA 2 performance, with a 0.528 higher score expected on PA 2 for every point on PA 1, and a 0.518 higher score expected on PA 3 for every point on PA 2. Finally, it is notable that PA 1 and PA 3 are both significant predictors of the EOT assessment score: for every point higher on the first progress assessment score, a learner can be expected to score 0.405 of a point higher at EOT, and for every point higher on PA 3, a learner can be expected to score 0.541, over a half point higher at EOT, or half an ILR level. This strong association of the third assessment with EOT scores is not unexpected, but it is notable that the second progress assessment, which takes place at a midpoint in training, was not a significant predictor of EOT scores when controlling for the other assessments and individual differences.

Table 19. Speaking Model 1: Significant Paths of Individual Differences to Assessment Scores

Path	**Parameter Value**	**S.E.**	**Significance Level**	**Standardized Parameter Value**	**Standardized S.E.**	**Significance Level**
MLAT to PA 1	0.006	0.002	0.001	0.174	0.048	≤0.001
MLAT to PA 2	0.004	0.001	0.002	0.123	0.039	0.002
MLAT to EOT	0.006	0.002	0.001	0.124	0.037	0.001
Role at Post to PA 2	-0.078	0.037	0.033	-0.096	0.045	0.032
First Tour to PA 1	0.112	0.037	0.003	0.163	0.055	0.003

Table 20. Speaking Model 1: Significant Paths between Progress Assessments

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
PA 1 to EOT	0.405	0.065	≤0.001	0.275	0.044	≤0.001
PA 3 to EOT	0.541	0.070	≤0.001	0.406	0.053	≤0.001
PA 1 to PA 2	0.528	0.043	≤0.001	0.476	0.036	≤0.001
PA 2 to PA 3	0.518	0.049	≤0.001	0.521	0.038	≤0.001

Table 21. Total Effects from Individual Differences to End of Training Assessment Score: Speaking

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT to EOT	0.013	0.002	≤0.001	0.261	0.045	≤0.001
Role at Post to EOT	-0.099	0.056	0.080†	-0.092	0.052	0.079†

†Approaches significance

Reading Model 1 Results

As previously described, reading Model 1 fit the data acceptably well. Altogether, the model explained 23.1% of the variance in EOT scores ($R^2 = .231$), a lower percent than Masters found in her similar study explaining speaking end of training exams at DLI using a qualifying exam score, aptitude measure, and language preference (2018, p.84). Figure 14, below, displays the causal paths in the Model 1 reading results using black arrows; gray arrows represent non-significant causal paths. Dotted arrows represent causal paths which approach significance ($0.05 \leq p \leq 0.08$).

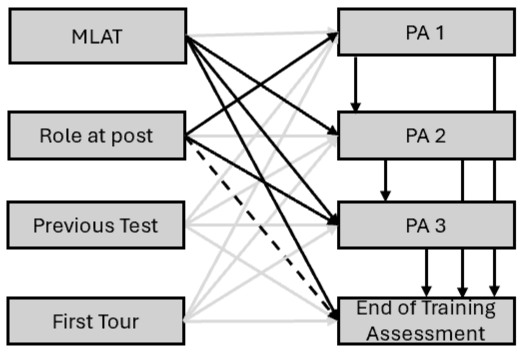

Figure 14. Model 1: Reading Results

Relationships of Individual Differences to Reading Proficiency Assessment Scores

The most notable individual differences were MLAT and *Role at Post*: MLAT scores significantly predicted the second and third progress assessment scores, as well as the end of training assessment score, and *Role at Post* was significantly and negatively associated with scores on the first and third progress assessment scores, but positively associated with the end of training test. Table 22-Table ***24***, below, display the non-standardized and standardized causal path coefficients, the standard errors of the path coefficients, and the significance level of the coefficients for the reading Model 1 results. The following section discusses these results. A qualitative summary of these results will be presented as part of the response to research question 4 (Table 25).

Aptitude as measured by MLAT was the most notable individual difference estimated by the model. For every standard deviation larger a student scored on the MLAT (6.48 points), the learner could be expected to score 0.087 higher on PA 2, 0.173 higher on PA 2, and 0.237 higher on the EOT assessment, controlling for all other individual differences. By the end of training, the total effect of MLAT on the EOT assessment is an expected .301 higher score for every 6.48 points higher on

MLAT. This means that, with all other variables held constant, a 10.76 point difference in MLAT scores would predict a proficiency difference of 0.5 on the ILR scale, or a half-level increase. This overall effect is as hypothesized, except for the lack of relationship between MLAT and PA 1. *Role at Post* was negatively associated with performance on PA 1 and 3: students with specialized jobs could be expected to score 0.115 points lower than students with non-specialized jobs on the first progress assessment and 0.083 points lower on the third. The hypothesized effect of *Role at Post* was an expected lower score for specialized students, especially toward the end of training. The direct effect of role at post to the reading EOT scores neared significance in a positive relationship. Having a *Previous Test* on record was expected to be most strongly associated with PA 1 scores, but was not significantly associated with any assessment, nor did it exhibit a significant total effect on the EOT assessment. Finally, *First Tour* learners were expected to perform more poorly on their progress assessments, but this individual difference was also not significantly related to any progress assessments nor directly with the EOT score, nor did it demonstrate a significant total effect on the EOT score.

The assessment scores over time were closely related: PA 1 performance predicted PA 2 performance, with a 0.505 higher score expected on PA 2 for every point on PA 1, and a 0.193 higher score expected on PA 3 for every point on PA 2. Finally, it is notable that PA 1, PA 2, and PA 3 are all significant predictors of the EOT assessment score: for every point higher on the first progress assessment score, a learner can be expected to score about a quarter of a point higher (0.251) at EOT, for every point higher on PA 2 a learner can be expected to score 0.147 points higher at

EOT, and for every point higher on PA 3, a learner can be expected to score about .457 points higher at EOT.

Table 22. Reading Model 1: Significant Paths of Individual Differences to Assessment Scores

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT to PA 2	0.003	0.002	0.036	0.087	0.041	0.034
MLAT to PA 3	0.004	0.002	0.003	0.173	0.056	0.002
MLAT to EOT	0.013	0.002	≤0.001	0.237	0.044	≤0.001
Role at Post to PA 1	-0.115	0.053	0.030	-0.112	0.051	0.029
Role at Post to PA 3	-0.083	0.039	0.032	-0.133	0.062	0.031
Role at Post to EOT	0.107	0.060	0.075†	0.083	0.046	0.073†

†Approaches significance

Table 23. Reading Model 1: Significant Paths between Progress Assessments

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
PA 1 to EOT	0.251	0.066	≤0.001	0.199	0.052	≤0.001
PA 2 to EOT	0.147	0.075	0.050	0.110	0.056	0.049
PA 3 to EOT	0.457	0.116	≤0.001	0.220	0.054	≤0.001
PA 1 to PA 2	0.505	0.035	≤0.001	0.540	0.035	≤0.001
PA 2 to PA 3	0.193	0.038	≤0.001	0.298	0.051	≤0.001

Table 24. Total Effects from Individual Differences to End of Training Assessment Score: Reading

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT to EOT	0.016	0.002	≤0.001	0.301	0.046	≤0.001

Differences in Influence of Individual Differences between Reading and Speaking

Research Question 2 (RQ2): Does the influence of individual differences on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

The impact of individual differences on progress and end of training assessment scores was similar in reading and speaking, with differences primarily in the times during training when the individual differences predicted proficiency scores. Table 25 presents a summary of the hypothesized causal relationships between exogenous individual differences and endogenous proficiency scores in speaking and reading alongside a summary of the observed causal relationships.

Table 25. Summary of Model 1 Results in Speaking and Reading

Individual Difference	Hypothesized Effect	Speaking Outcomes	Reading Outcomes
MLAT	Positive effect throughout training, most strongly in first and second PAs	• Positive effect at PA 1, PA 2, EOT • Total positive effect on EOT	• Positive effect at PA 2, PA 3, EOT • Total positive effect on EOT
Role at Post	Negative effect in PAs 2 and 3 and at EOT	• Negative effect at PA 2 • Total negative effect at EOT†	• Negative effect at PA 1 and 3 • Positive direct effect at EOT†
Previous Test	Negatively effect in PA 1, non-significant in later training and at EOT	• No effects observed	• No effects observed
First Tour	Negatively effect in PA 1, non-significant in later training and at EOT	• Positive effect at PA 1	• No effects observed

†Approaches significance

In both speaking and reading, aptitude as measured by MLAT significantly and positively predicted performance on progress assessment scores and the EOT assessment scores. However, in speaking, MLAT predicted performance at earlier stages (PA 1 and 2), compared to reading (PA 2 and 3). In both modalities, MLAT was a significant positive direct predictor of EOT scores, with a significant positive total effect.

The effect of the job role that the student was training for (*Role at Post*) and the alignment of the job tasks with the training received was different between speaking and reading. In speaking, students with a specialized job had significantly lower scores on the second progress assessment, and the negative total effect of having a specialized role approached significance. In contrast, students with specialized roles had significantly poorer outcomes on the first and third progress assessments in reading, but a total positive effect on EOT approached significance.

The effect of being on a F*irst Tour* of language learning at LTF differs between speaking and reading. In speaking, learners who are in their first tour of language learning at the LTF perform better on the first progress assessment than non-first time students, but there is no effect on any following assessments, nor a significant total effect at EOT. However, in reading, no effect of being a first tour learner is observable, whether on any of the assessments or in total at EOT.

Also notable is the lack of influence of *Previous Test*. Across reading and speaking, the effect of having taken a previous test, used in this study as a proxy for previous advanced language learning, did not significantly predict any of the assessment scores, nor did it have a significant total effect on the EOT score.

Another similarity between speaking and reading is the relationships between the proficiency assessment scores. In both reading and speaking, the first progress assessment significantly and positively predicted the second, and the second significantly and positively predicted the third. In both modalities, the first and third progress assessment scores significantly and positive predicted the EOT score, but in the speaking modality the second progress assessment was not predictive of the EOT score.

Proportions of Explained Variance in Assessments

In addition to the contributions of each exogenous variable to the endogenous variables, the proportion of variance of each endogenous variable that is explained by the model can shed light on how well the individual differences and prior assessment scores explain proficiency at each assessment point. This is accomplished using an R^2 value which describes explained variance. The proportions of variance explained by the model differed between reading and speaking: a higher proportion of variance in speaking than reading was explained across nearly all assessments (Table 26). A traditional interpretation of R^2 indicates that the effect size of the variance accounted for in this model ranges from insignificant (<0.09), to small (>0.09), to moderate (>0.25) (Cohen, 1988, p. 157). A rule of thumb about R^2 that is often cited is that below 0.6 is negligible: however, this standard is not as applicable to the social sciences, where an R^2 between 0.1 and 0.5 may be considered notable and significant if at least one predictor is significant (Ozili, 2023, p. 8). Using this standard and the traditional effect size established by Cohen, all but the first progress assessment demonstrate at least a small if not moderate effect size with notable percents of variance explained. The percent of variance explained in the first progress assessment scores is notably lower than the proportion of variance explained in subsequent assessments in both modalities. In speaking, the percent of variance explained rose with each assessment, from 4.9% in PA 1 to 42.7% at EOT. However, in reading, the second assessment had the highest percent of explained variance (32.9%), higher than the third (17.2%) or EOT (23.1%) assessments. A notably higher percent of variance in EOT speaking scores (42.7%) was explained than variance in EOT reading scores (23.1%).

Table 26. Model 1: Percent Assessment Score Variance Explained in Reading and Speaking

	Speaking		Reading	
	R^2	% Explained Variance	R^2	% Explained Variance
PA 1	0.049	4.9%	0.017	1.7%
PA 2	0.286	28.6%	0.329	32.9%
PA 3	0.327	32.7%	0.172	17.2%
EOT	0.427	42.7%	0.231	23.1%

Discussion of Individual Differences Models

The models fit explained a fair amount of variance in the patterns of language proficiency growth and end of training proficiency outcomes in both reading and speaking. However, relatively few causal pathways were significant in predicting assessment scores, considering that the model was nearly fully-saturated.

The pattern of significant positive influence of aptitude across individual reading and speaking progress assessments and in total effect on EOT aligns with previous research in intensive SLA (Silva and White, 1993; Erhman and Oxford, 1995, Linck et al., 2013; Waegner, 2015, Masters, 2018; Doughty, 2019). A notable exception to this pattern is the lack of significant causal path between MLAT and the first progress assessment in the reading model. It is possible that reading skills at the beginning of training are less dependent on aptitude and more influenced by another individual difference or exogenous variable: this is discussed below, with the Proportions of Explained Variance in Assessments.

Whether or not a learner was training for language use in a role with specialized or non-specialized skills was associated their progress assessment scores during mid and late training in both modalities. Learners who had more specialized roles were hypothesized to perform more poorly due to lower motivation and engagement because the tasks they practiced and were tested on did not align to their planned language use, and this was observed in their progress assessment scores: a negative effect at PA 2 in speaking and PA 2 and 3 in reading. An unexpected result was the relationship of *Role at Post* to EOT scores: *Role at Post* has a near-significant negative total effect on speaking EOT assessment scores, and a near-significant positive direct effect on reading EOT assessment scores. However, in both cases, the effect on EOT scores was about 0.010 points, or one one-hundredth of an ILR level, and given the non-significant value, in practice *Role at Post* likely has negligible effects on EOT scores.

Another unexpected finding was low numbers of significant causal paths explaining the relationships between having taken a previous test or being on a first tour of study and the progress assessments. Given the influence of a prior interlanguage on third-language acquisition, these individual differences were expected to have a larger role in the language learning process and the EOT results. It appears that, controlling for aptitude and role at post, the effect of previous interlanguage acquisition or study are largely non-significant in the individual differences model.

The only significant path from either of these two exogenous variables was an early positive effect of first tour: being a first time language student at the LTF was found to positively predict PA 1 scores in speaking. The hypothesized effect was a negative association with the first and possibly second progress assessments, as learners adjusted to the curricula, schedule, and goals of the LTF. A possible explanation is that first-time learners were especially motivated to succeed at their studies, which would align prior with longitudinal motivation research. For example, Busse and Walter found that language learner self-efficacy and motivation fall over time (2013, p. 443). Furthermore, learners who have undergone previous tours of language learning at the LTF may have memories of the difficulty of language learning (MacIntyre, 2017), which may negatively impact their language learning motivation and emotions at the start of a subsequent training tour (Saito et al, 2018, pp. 732-733). This finding was not found in reading, and it is hypothesized that this is because the speaking tasks and assessments had heighted face validity with tasks at post, compared to reading tasks: thus reading tasks were not more or less motivating between groups of students. There was no total effect of *First Tour* on EOT scores in either modality, nor direct effect of *First Tour* on any assessment other than PA 1 speaking.

The relationships between assessment scores were largely as expected, except for the finding that the second progress assessment was not predictive of the EOT assessment in speaking. This is especially unusual as the first progress assessment, which occurs at a greater temporal distance from the EOT assessment, was significantly predictive of it, and the finding that the second progress assessment was significantly associated with the first and third progress assessments as predicted. It is possible that at this midpoint time in training, students are so variable in their speaking proficiency that the association with the end of training test drops, but its association with the immediate previous and following tests remains due to their temporal closeness to the assessment, or that the PA assessment methods do not target the intermediate skills reliably.

Discussion of Explained Variance in Individual Differences Model

As noted above, the proportions of variance that were explained by the respective models roughly rose over time: in both reading and speaking, negligible amounts of proficiency as measured by PA 1 assessments were explained by the individual differences in the model. While some degree of construct-irrelevant variance is expected in all language proficiency tests (Messick, 1984, p. 216), this low degree of variance explained points to the existence of at least one unobserved exogenous variable. It is unknown whether this variance would be explained by an individual difference, typological variable, or learning environment factor. In speaking, possible exogenous variables that were not included in this model but that have been shown to explain beginner L2 speaking proficiency include varied levels of classroom participation (Ely, 1986), willingness to communicate (Hernández, 2010), and integrative motivation (Hernández, 2006). It is also possible that the novel situation of an oral interview at a point when they had beginner proficiency caused students to rely on their L1 interpersonal skills, something also unmeasured in this model. In reading, a possible exogenous variable which would explain variance in PA 1 scores which was unmeasured in this model but has been shown to be predictive of early L2 reading is L1 reading skills, particularly L1 word decoding and text comprehension (Sparks et al., 2023, p. 492-494), or L1 and L2 reading attitude (Yamashita, 2004, p. 11), or L1 print exposure (Sparks et al., 2012, p. 497). It is also possible that given the low basis of proficiency that learners have at the first PA, they were unable to draw sufficiently on their interlanguage skills in order to reliably measure L2 reading (Lee & Schallert, 1997, p. 732; Laufer, 1992).

Second, the fact that lower proportions of variance of reading assessment scores were explained of than speaking assessment scores indicates that there is likely some exogenous variable that is more closely related to reading proficiency that is unobserved and excluded from this model. The same variables that were discussed above as potential exogenous variables that could explain variance in PA 1 reading scores would be likely candidates for these: L1 word decoding and text comprehension (Sparks et al., 2023, p. 492-494), or L1 and L2 reading attitude (Yamashita, 2004, p. 11), or L1 print exposure Sparks et al., 2012, p. 497).

Chapter 7: Language Difficulty and Individual Differences in Language Proficiency Gains Results

Research Question 3 (RQ3): Does language difficulty moderate the relationships between individual differences and L2 proficiency assessment scores throughout intensive language study? If so, which points in learning are affected and in what ways?

Contribution of Language Difficulty

To respond to research question 3, difficulty was added as a moderating variable to the speaking and reading models established in Model 1. Table 27 and Table 28, below, describe the fit statistics of the Model 2 analyses. In order to compare Model 2 to Model 1, a pseudo-Model 1 (Model 1^{A}) was fit using all variables in Model 2, but with the additional Model 2 paths set to 0. Model 1^{A} fit statistics that were used for comparison to Model 2 are listed in Table 27 and Table 28, Model 1^{A} fit statistics not used for model comparison are not listed.

Table 27. Speaking Model 1 and 2 Fit Statistics

	CFI	RMSEA	SRMR	BIC	χ^2	p	Log likelihood	Satorra-Bentler scaling correction factor	Number of Parameters
Model 1^{A}	-	-	-	10748.515	-	-	-5251.392	1.0620	39
Model 2	0.961	0.210	0.021	918.659	24.964	$\leq .001$	-304.960	1.0087	49

Model 1^{A} is Model 2 with all path associated with Model 2 variables (*difficulty* and moderating variables) set to 0. Values not used to compare Model 1^{A} and 2 are not listed here.

Table 28. Reading Model 1^A and 2 Fit Statistics

	CFI	RMSEA	SRMR	BIC	χ^2	p	Log likelihood	Satorra-Bentler scaling correction factor	Number of Parameters
Model 1 ^{A}	-	-	-	11090.563	-	-	-5422.416	1.1081	39
Model 2	1.000	0.000	0.008	1393.963	0.671	0.4128	-541.816	1.0357	49

Model 1^A is Model 2 with all path associated with Model 2 variables (*difficulty* and moderating variables) set to 0. Values not used to compare Model 1 ^{A} and 2 are not listed here.

Utilizing the Satorra-Bentler scaling-corrected χ^2 value as calculated in Equation 6 and Equation 7, both the speaking ($\chi^2(10) = 13876.483, p =< .001$) and reading ($\chi^2(10) = 12957.230, p =< .001$) models with difficulty added as a moderating variable (Model 2) fit significantly better than the model with individual differences only (Model 1^A). Examination of the BIC fit statistics supported this finding, as the difference in BIC from Model 1 ^{A} to Model 2 was >10 in both modalities (reading: 9829.856, speaking: 9696.6), indicating that the evidence supporting Model 2 over Model 1 was "very strong" (Kass and Raffery, 1995, p. 777). These findings together provide strong evidence that the difficulty of the L2 learned influences the proficiency attainment throughout training above and beyond the influences of the individual differences.

Equation 6. Calculation of Satorra Bentler Test Statistic: Speaking Model 2 vs. 1

$$cd = \frac{(p_{Model\,1} * c_{Model\,1} - p_{Model\,2} * c_{Model\,2})}{(p_{Model\,1} - p_{Model\,2})} = \frac{(39*1.062-49*1.0087)}{(39-49)} = 0.801$$

$$TRd = -2(L_{Model\,1} - L_{Model\,2})/cd = -\frac{2(-5251.392 - 304.96)}{0.75334} = 13876.483$$

Equation 7. Calculation of Satorra Bentler Test Statistic: Reading Model 2 vs. 1

$$cd = \frac{(p_{Model\,1} * c_{Model\,1} - p_{Model\,2} * c_{Model\,2})}{(p_{Model\,1} - p_{Model\,2})} = \frac{(39*1.1081-49*1.0357)}{(39-49)} = 0.753$$

$$TRd = -\frac{2(L_{Model\,1} - L_{Model\,2})}{cd} = -\frac{2(-5422.416- -541.816)}{0.75334} = 12957.230$$

Moderating Effects of Second Language Difficulty

Speaking Model 2 Results

As previously described, speaking Model 2 fit the data acceptably well. The addition of difficulty and difficulty as a moderating variable on the influence of individual differences significantly improved the model fit from Model 1. However, the model explained 42.6% of the variance in EOT scores ($R^2 = .426$, Table 38. Model 2: Percent Assessment Score Variance Explained in Reading and Speaking), indicating that despite the models' stronger description of the data, it explained nearly the same amount of variance in EOT scores as Model 1 (42.7%), but higher amounts of variance in the formative progress assessment scores. Figure 15. Model 2: Speaking Results, below, displays the causal paths in the Model 2 speaking results using black arrows; gray arrows represent non-significant causal paths.

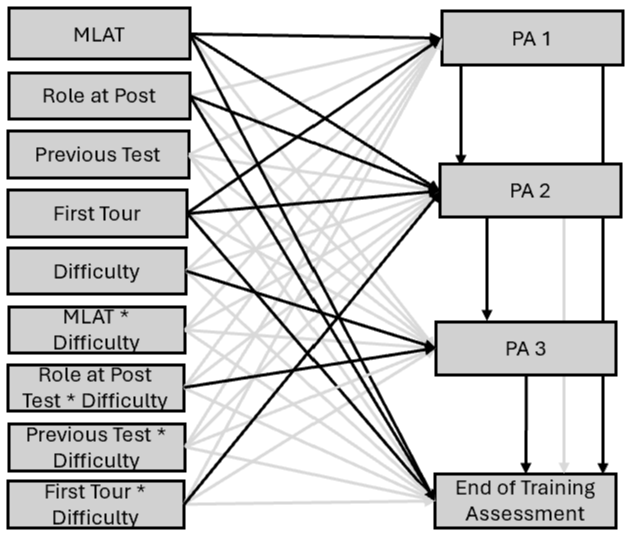

Figure 15. Model 2: Speaking Results

Relationships of Individual Differences and Difficulty Moderation to Speaking Proficiency Assessment Scores

As previously described, the addition of difficulty and difficulty as a moderating variable on the influence of individual differences significantly improved the model fit from Model 1. The ensuing results present a more complex explanation of the drivers of language proficiency scores throughout intensive training. Table 29-Table 32, below, display the non-standardized and standardized causal path coefficients, the standard errors of the path coefficients, and the significance level of the coefficients for the speaking Model 2 results. The following section discusses these results. A qualitative summary of these results will be presented as part of the response to research question 4 (Table 37).

As in Model 1, the individual-differences only model, MLAT was a key predictor of proficiency assessment scores. MLAT positively predicted PA 1, PA 2, and the EOT assessment scores. For every additional standard deviation (8.86 additional points) on the MLAT, a learner could be expected to score 0.186 higher on PA 1, 0.150 higher on PA 2, and 0.134 higher on the EOT assessment. In addition, the total effect of MLAT was significant, with a learner expected to score 0.273 higher on their EOT speaking assessment for every standard deviation higher they scored on MLAT. This means that, with all other variables held constant, a 23.82 point difference in MLAT scores would predict a proficiency difference of 0.5 on the ILR scale, or a half- ILR level increase. Difficulty was not found to moderate the relationship between MLAT and any assessment scores.

With the addition of difficulty and the mediating variables to the model, *Role at Post* was predictive of additional assessment scores across the training period compared to Model 1. Learners who had specialized job roles (*Role at Post*) were expected to score 0.105 points lower on their PA 2 and 0.096 points lower at EOT than learners with non-specialized jobs. The effects of job role summed to a significant total effect of learners with specialized jobs scoring 0.103 points for lower on their EOT speaking assessment. In addition, difficulty moderated the relationship between *Role at Post* and PA 3. This moderating variable of the difficulty of a language a student studied and their job role had a positive effect on PA 3, with an estimated parameter value of 0.756. This means that among students with specialized jobs, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks would be 0.756 at PA 3. This moderating effect occurred in the context of a non-significant path from *Role at Post* to PA 3, and a negative significant effect of difficulty on PA 3, described below. However, this difference in PA 3 scores between learners studying L2s with different proficiency expectations at 24 weeks did not exist between learners with non-specialized jobs. This indicates that the negative overall effect of difficulty at PA 3 (-.553) was essentially counteracted by the moderating effect (.756) among students whose curricula were misaligned with their job tasks: these learners scored lower if they studied harder language, while learners whose curricula were more closely aligned scored higher if they studied harder languages.

There was no observable relationship between *Previous Test* and any of the assessment scores, nor did difficulty moderate any causal paths between *Previous Test* and the assessment scores.

Learners who were on their *First Tour* of language training at the LTF scored 0.091 higher on PA 1 than learners who were returning to language training, but scored 0.080 lower on PA 2 and .110 lower on their EOT assessment, demonstrating a negative trend over training. This overall trend summed to a total effect of 0.103 lower scores for first time learners on their EOT assessments in speaking. The moderating variable of first tour and difficulty had a positive effect on PA 2, with an estimated parameter value of 0.241. This means that among students on their first tour of language study at the LTF, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks would be 0.241 at PA 3. This moderating effect occurred in the context of a non-significant path from *Difficulty* to PA 2, in addition to the significant effect of *First Tour* on PA 2. This indicates that the negative overall effect of *First Tour* at PA 2 (-.080) was counterbalanced by the positive moderating effect (.241) among students on their first tours of language study at the LFT: these learners demonstrated patterns of lower scores as language difficulty increase, while returning learners did not demonstrate this pattern.

L2 difficulty as a direct predictor was only significantly related to PA 3: learners who learned less difficult languages were expected to score lower on PA 3 than learners who learned harder languages. For every level higher on the ILR scale that a learner was expected to attain after 24 weeks of training, a learner was expected to score 0.553 lower on PA 3.

Several variables had significant total effects on EOT scores, especially those related to MLAT, *Role at Post*, and *First Tour.* For every standard deviation in MLAT score, or 8.86 points, a learner would be expected to score 0.273 higher on the EOT assessment. Learners who had specialized roles at post would be expected to score 0.149 points lower on the EOT assessment. In addition, the total effect of the interaction between *Role at Post* and *Difficulty* on EOT scores was significant. Among just learners who had specialized roles at post, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks was 0.599 at EOT. Similarly, the total effect of being on the first tour of duty in language training at LTF was an expected score that was 0.103 points lower than learners on subsequent tours. The total effect of the interaction of *First Tour* and *Difficulty* on EOT scores was significant. Among just learners who were on their first tours of duty as language students at LTF, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks was 0.587 at EOT.

Finally, the assessment scores over time were closely related: PA 1 performance predicted PA 2 performance, with a 0.480 higher score expected on PA 2 for every point on PA 1, and a 0.503 higher score expected on PA 3 for every point on PA 2. Finally, it is notable that PA 1 and PA 3 are both significant predictors of the EOT assessment score: for every point higher on the first progress assessment score, a learner can be expected to score 0.409 points higher at EOT and for every point higher on PA 3, a learner can be expected to score 0.523 points higher at EOT. However, there was no significant relationship between PA 2 and the EOT assessment score

Table 29. Speaking Model 2: Significant Paths of Exogenous Variables to Assessment Scores

	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT to PA 1	0.006	0.002	≤0.001	0.186	0.048	≤0.001
MLAT to PA 2	0.005	0.001	≤0.001	0.150	0.040	0.002
MLAT to EOT	0.006	0.002	≤0.001	0.134	0.037	≤0.001
Role at Post to PA 2	-0.105	0.035	0.002	-0.129	0.042	0.002
Role at Post to EOT	-0.096	0.048	0.048	-0.089	0.045	0.048
First Tour to PA 1	0.091	0.040	0.022	0.132	0.057	0.021
First Tour to PA 2	-0.080	0.035	0.021	-0.105	0.045	0.021
First Tour to EOT	-0.110	0.046	0.016	-0.109	0.045	0.016
Difficulty to PA 3	-0.553	0.255	0.030	-0.480	0.220	0.029

Table 30. Speaking Model 2: Significant Paths of Moderating Variables to Assessment Scores

	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
Role at Post * Difficulty to PA 3	0.756	0.283	0.008	0.196	0.075	0.009
First Tour * Difficulty to PA 2	0.241	0.098	0.014	0.069	0.029	0.019

Table 31. Speaking Model 2: Significant Paths between Progress Assessments

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
PA 1 to EOT	0.409	0.064	≤0.001	0.279	0.044	≤0.001
PA 3 to EOT	0.523	0.074	≤0.001	0.384	0.055	≤0.001
PA 1 to PA 2	0.480	0.043	≤0.001	0.433	0.038	≤0.001
PA 2 to PA 3	0.503	0.048	≤0.001	0.518	0.040	≤0.001

Table 32. Speaking Model 2 Total Effects from Exogenous Variables to End of Training Assessment Score

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT	0.013	0.002	≤0.001	0.273	0.044	≤0.001
Role at Post	-0.149	0.057	0.009	-0.138	0.053	0.009
First Tour	-0.103	0.050	0.040	-0.101	0.049	0.040
Role at Post * Difficulty	0.599	0.253	0.018	0.114	0.050	0.021
First Tour * Difficulty	0.587	0.194	0.002	0.127	0.045	0.004

Reading Model 2 Results

As previously described, reading Model 2 fit the data acceptably well. The addition of difficulty and difficulty as a moderating variable on the influence of individual differences significantly improved the model fit from Model 1. With the addition of difficulty, the model explained notably more variance (32.1%) of the variance in EOT scores ($R^2 = 32.1$) than Model 1 (23.1%, $R^2 = .231$). Figure 16. Model 2: Reading Results, below, displays the causal paths in the Model 2 reading results using black arrows; gray arrows represent non-significant causal paths. Dotted arrows represent causal paths which approach significance ($0.05 \leq p \leq 0.08$).

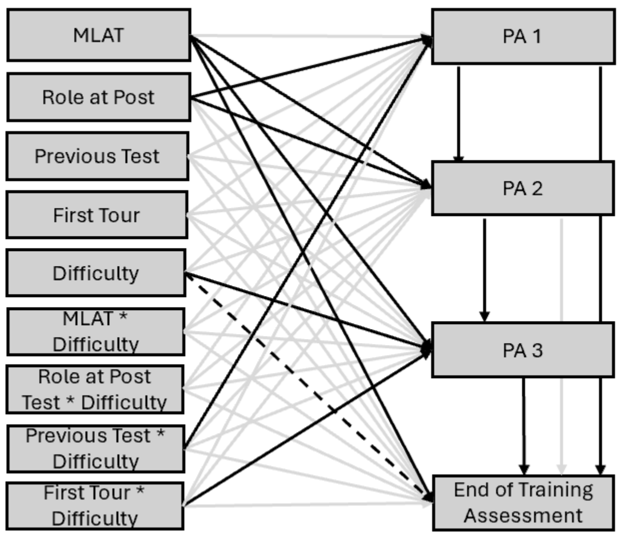

Figure 16. Model 2: Reading Results

Relationships of Individual Differences and Difficulty Moderation to Reading Proficiency Assessment Scores

As previously described, the addition of difficulty and difficulty as a moderating variable on the influence of individual differences significantly improved the model fit from Model 1. The ensuing results present a more complex explanation of the drivers of language proficiency scores throughout intensive training. Table 33-Table 36, below, display the non-standardized and standardized causal path coefficients, the standard errors of the path coefficients, and the significance level of the coefficients for the reading Model 2 results. The following section discusses these results. A qualitative summary of these results will be presented as part of the response to research question 4 (Table 37).

As in Model 1, the individual-differences only model, MLAT was a key predictor of proficiency assessment scores. MLAT positively predicted PA 2, PA 3, and the EOT assessment. For every additional standard deviation (8.92 points) on the MLAT, a learner could be expected to score 0.092 higher on PA 2, 0.166 higher on PA 3, and 0.259 higher on the EOT assessment. In addition, the total effect of MLAT was significant, with a learner expected to score 0.321 higher on their EOT reading assessment for every standard deviation higher they scored on MLAT. This means that, with all other variables held constant, a 48.48 point difference in MLAT scores would predict a proficiency difference of 0.5 on the ILR scale, or a half-ILR level increase. Difficulty was not found to moderate the relationship between MLAT and any assessment scores.

With the addition of difficulty and the moderating variables to the model, *Role at Post* was predictive of additional assessment scores across the training period. Learners who had specialized job roles were expected to score 0.133 points lower on PA 1 and 0.084 points lower on PA 2 than learners with non-specialized jobs. There was no significant total effect of *Role at Post.*

There was no observable direct relationship between *Previous Test* and any of the assessment scores. However, the moderating variable of difficulty and having previously taken a proficiency test at LTF had a positive effect on PA 1, with an estimated parameter value of 0.474. This means that among students who had previously taken an LTF test, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks would be 0.474 at PA 1. However, this difference in PA 1 scores between learners studying L2s with different proficiency expectations at 24 weeks did not exist among students who had not taken a previous test at the LTF. This moderating effect occurred in the context of a non-significant path from *Previous Test* to PA 1, and a non-significant path from *Difficulty* to PA 1, indicating that this effect of difficulty on PA 1 was only found in this group who had taken previous tests.

There was no observable direct relationship between *First Tour* and any of the assessment scores. However, difficulty moderated the relationship between *First Tour* on assessment scores. The moderating variable of difficulty and whether or not they had previously studied at the LTF had a positive effect on PA 3, with an estimated parameter value of 0.633. This means that among students who were studying at the LTF for the first time, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks would be 0.633 at PA 3. However, this difference in PA 3 scores between learners studying L2s with different proficiency expectations at 24 weeks did not exist among learners who were returning to LTF. This moderating effect occurred in the context of a non-significant path from *First Tour* to PA 3, and a significant path from *Difficulty* to PA 3, described below. This indicates that the negative overall effect of *Difficulty* at PA 3 (-.623) was largely counteracted by the moderating effect (.633) among students on their first tour of language study at the LTF: these learners scored lower if they studied harder languages. Given the close and opposite values of the path coefficients, this indicates the moderating effect largely served to counteract the effect of difficulty in first tour learners, while learners returning to the LTF scored higher if they studied harder languages.

L2 difficulty as a single predictor was only significantly related to PA 3 and EOT assessment scores: learners who learned easier languages were expected to score lower on PA 3 and higher on EOT than learners who learned harder languages. For every level higher on the ILR scale that a learner was expected to attain after 24 weeks of training, a learner was expected to score 0.623 lower on PA 3 and 0.426 higher at EOT.

Several variables had significant total effects on EOT reading scores, especially MLAT and the moderating interaction variables. For every standard deviation in MLAT score, or 8.92 points, a learner would be expected to score 0.321 higher on the EOT assessment. The total effect of the moderating variable of *Role at Post* and *Difficulty* on EOT scores was significant, with a parameter value of 0.705. This means that among learners who had specialized roles at post, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks was 0.705 at EOT. The total effect of the moderating variable of *Previous Test* and *Difficulty* on EOT scores was significant, with a parameter value of 0.381. This means that among learners who had taken an LTF test prior to the start of language training, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks was 0.381 at EOT. The total effect of the interaction of *First Tour* and *Difficulty* on EOT scores was significant. Among just learners who were on their first tours of duty as language students at LTF, the expected difference between a learner who studied a language where they were expected to achieve an ILR 0+ at 24 weeks and one who was expected to achieve an ILR 1+ at 24 weeks was 0.563 at EOT.

Finally, the assessment scores over time were closely related: PA 1 performance predicted PA 2 performance, with a 0.467 higher score expected on PA 2 for every point on PA 1, and a 0.179 higher score expected on PA 3 for every point on PA 2. Finally, it is notable that PA 1 and PA 3 are both significant predictors of the EOT assessment score: for every point higher on the first progress assessment score, a learner can be expected to score 0.232 points higher at EOT and for every point higher on PA 3 a learner can be expected to score 0.504 points higher at EOT. However, there was no significant relationship between PA 2 and the EOT assessment score.

Table 33. Reading Model 2: Significant Paths of Exogenous Variables to Assessment Scores

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT to PA 2	0.004	0.002	0.025	0.092	0.040	0.024
MLAT to PA 3	0.004	0.001	0.003	0.166	0.053	0.002
MLAT to EOT	0.014	0.002	≤0.001	0.259	0.042	≤0.001
Role at Post to PA 1	-0.133	0.058	0.022	-0.129	0.056	0.021
Role at Post to PA 2	-0.084	0.036	0.022	-0.086	0.038	0.023
Difficulty to PA 3	-0.623	0.241	0.010	-0.706	0.263	0.007
Difficulty to EOT	0.426	0.240	0.076†	0.233	0.131	0.076†

†Approaches significance

Table 34. Reading Model 2: Significant Paths of Moderating Variables to Assessment Scores

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
Previous Test * Difficulty to PA 1	0.474	0.204	0.020	0.308	0.133	0.021
First Tour * Difficulty to PA 3	0.633	0.271	0.020	0.267	0.113	0.018

Table 35. Reading Model 2: Significant Paths between Progress Assessments

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
PA 1 to EOT	0.232	0.062	≤0.001	0.186	0.050	≤0.001
PA 3 to EOT	0.504	0.106	≤0.001	0.244	0.049	≤0.001
PA 1 to PA 2	0.467	0.034	≤0.001	0.499	0.035	≤0.001
PA 2 to PA 3	0.179	0.039	≤0.001	0.277	0.054	≤0.001

Table 36. Reading Model 2 Total Effects from Exogenous Variables to End of Training Assessment Score

Path	Parameter Value	S.E.	Significance Level	Standardized Parameter Value	Standardized S.E.	Significance Level
MLAT	0.017	0.002	≤0.001	0.321	0.044	≤0.001
Role at Post * Difficulty	0.705	0.343	0.040	0.105	0.052	0.044
Previous Test * Difficulty	0.381	0.216	0.078†	0.198	0.111	0.075†
First Tour * Difficulty	0.563	0.182	0.002	0.115	0.040	0.004

†Approaches significance

<u>*Differences in Influence of Language Characteristics between Reading and Speaking*</u>

Research Question 4 (RQ4): Does the influence of individual differences and difficulty on L2 proficiency assessment scores throughout intensive language study differ between reading and speaking? If so, in what ways?

The impact of individual differences on progress and end of training assessment scores was similar in reading and speaking, with differences primarily in the times during training when the individual differences or moderating variables predicted proficiency scores Table 37 presents a summary of the hypothesized causal relationships between exogenous individual differences and difficulty variables and endogenous proficiency scores in speaking and reading alongside a summary of the observed causal relationships.

Table 37. Summary of Model 2 Results in Speaking and Reading

Individual Difference	Hypothesized Effect	Speaking Outcomes	Reading Outcomes
MLAT	Positive effect throughout training, most strongly in first and second PAs	• Positive effect at PA 1, 2 • Positive total effect on EOT	• Positive effect at PA 2, 3, EOT • Positive total effect on EOT
Role at Post	Negative effect in PAs 2 and 3 and at EOT	• Negative effect at PA 2 • Negative total effect on EOT	• Negative effect at PA 1, 2
Previous Test	Negatively effect in PA 1, non-significant in later training and at EOT	• No effect observed	• No effect observed
First Tour	Negatively effect in PA 1, non-significant in later training and at EOT	• Positive effect at PA 1; Negative effect at PA 2 • Negative effect at EOT	• No effect observed
Difficulty	Positive effect throughout (easier languages with higher scores)	• Negative effect of less difficult languages at PA 3	• Negative effect of less difficult languages at PA 3 • Positive effect of less difficult languages at EOT
Difficulty * MLAT	No effects predicted.	• No effect observed	• No effect observed

Table 37 (continued). Summary of Model 2 Results in Speaking and Reading

Individual Difference	Hypothesized Effect	Speaking Outcomes	Reading Outcomes
Difficulty * Role at Post	Learners with specialized roles more negatively influenced by difficulty throughout	• Learners with specialized roles scored higher on less difficult languages at PA 3 • Learners with specialized roles scored higher on less difficult languages in total effect on EOT	• Learners with specialized roles scored higher on less difficult languages at PA 3 • Learners with specialized roles scored higher on less difficult languages in total effect on EOT
Difficulty * Previous Test	Learners with previous test less influenced by difficulty	• No effect observed	• Learners who had taken a previous test scored higher on less difficult languages at PA 1 • Learners who had taken a previous test scored higher on less difficult languages in total effect on EOT
Difficulty * First Tour	Learners on first tour negatively influenced by difficulty especially at PA 2, 3, EOT	• First tour learners scored higher on less difficult languages at PA 2 • First tour learners scored higher on less difficult languages on EOT (total effect)	• First tour learners scored higher on less difficult languages at PA 3 • Learners with specialized roles scored higher on less difficult languages on EOT (total effect)

†Approaches significance

Aptitude as measured by the MLAT demonstrated a positive effect throughout training in both modalities, but was more associated with speaking proficiency scores in early training (PA 1 and 2) in speaking, compared to the later effect in reading (PA 2, 3, EOT). As in Model 1, MLAT scores had significant total effects on EOT assessment scores in both modalities. The hypothesized relationship between MLAT and proficiency was a significant effect in early language acquisition, so the reading outcome of a non-significant path connecting MLAT to PA 1 was unexpected. No significant interaction of MLAT score and difficulty was detected in either the reading or speaking model, indicating difficulty does not moderate the relationship between aptitude and proficiency scores at any point in training in either modality.

Role at Post had similar results between speaking and reading: a significant negative relationship to PA 1 and PA 2 scores. In the third PA in both reading and speaking, difficulty moderated the effect of role at post: learners with more specialized job roles were more affected by the difficulty of the language learned than learners without specialized roles: these learners with specialized jobs who studied less difficult languages scored higher on this assessment. A similar moderating effect was found between difficulty and *Role at Post* in total effect on EOT, again in both reading and speaking. Both of these findings align with the predicted effects.

Previous Test was hypothesized to have a significant negative relationship to performance on PA 1, but to be non-significant later in training across both modalities. However, in both reading and speaking there was no significant direct path from *Previous Test* to any of the proficiency assessments, and neither modality demonstrated a significant total effect of *Previous Test* on EOT scores. However, the difficulty of the L2 studied did have a moderating effect on the relationship of *Previous Test* to PA 3 and EOT in reading. In both PA 3 and the total effect on EOT: learners who had taken a LTF proficiency test prior to their current tour of duty studying at the LTF were more impacted by the difficulty of the language, and learners studying less difficult languages scored higher on the reading assessments.

Finally, *First Tour* also demonstrated some differences between reading and speaking. In speaking, learners on their first tour of language study at LTF scored higher on their PA 1 than returning learners. In speaking, the relationship of *First Tour* to PA 2 scores and the total effect of *First Tour* on EOT scores was moderated by difficulty: first tour learners studying less difficult languages scored higher on these assessments, while the scores of returning learners were not significantly impacted by the difficulty of the language. In reading, this positive moderating effect was found between *First Tour* and PA 3, and *First Tour* and the total effect on EOT scores: first tour learners studying less difficult languages scored higher on these assessments, while the scores of returning learners were not significantly impacted by the difficulty of the language.

Proportions of Explained Variance in Assessments

In addition to the contributions of each exogenous variable and moderating variable to the endogenous variables, the proportion of variance of each endogenous variable that is explained by the model can shed light on how well the individual differences, difficulty, and prior assessment scores explain progress at each assessment point. Table 38 shows the proportions and percents of explained variance for each assessment in both speaking and reading modalities. Overall, Model 2 explained higher proportions of speaking score variance than Model 1 at the start of training (PA 1 and 2) in speaking, but explained nearly the same percents of variance in PA 3 (33.8%, 1.1% more than Model 1) and EOT scores (42.6%, 0.1% less than Model 1). However, in reading, the model explained higher proportions of variance in every assessment score.

Table 38. Model 2: Percent Assessment Score Variance Explained in Reading and Speaking

	Speaking		Reading	
	R^2	% Explained Variance	R^2	% Explained Variance
PA 1	0.076	7.6%	0.067	6.7%
PA 2	0.342	34.2%	0.363	36.3%
PA 3	0.338	33.8%	0.228	22.8%
EOT	0.426	42.6%	0.321	32.1%

As in Model 1, the percent of variance explained in the first progress assessment scores is notably lower than the proportion of variance explained in subsequent assessments in both modalities. In speaking, the percent of variance explained rose with each assessment, from 7.6% in PA 1 to 42.6% at EOT. However, in reading, the second assessment had the highest percent of explained variance (36.3%), higher than the third (22.8%) or EOT (32.1%) assessments. A notably higher percent of variance in EOT speaking scores (42.6%) was explained than variance in EOT reading scores (32.1%).

Chapter 8: Summary and Discussion

Summary of Findings

Following the analyses describing the relationships between individual differences, difficulty, and proficiency assessment scores described in Chapters 6 and 7, this chapter summarizes the analyses findings and presents them within the context of instructed second language acquisition literature. Each research question is responded to below, followed by a discussion of the results and description of their implications.

Research Question 1: Individual Differences in Language Proficiency Attainment

The models examining the relationships between individual differences and proficiency scores explained the data to a significant degree. Across both modalities, aptitude as measured by MLAT was the most consistent predictor of proficiency scores.

Individual differences that predicted proficiency scores in the speaking model were MLAT, *Role at Post*, and *First Tour*: higher MLAT scores predicted higher proficiency scores in the first half of training, as well as the end of training assessment score. Misalignment between learners' *Role at Post* predicted poorer scores on the second progress assessment, and learners on their *First Tour* scored higher on the first progress assessment.

Individual differences that predicted proficiency scores in the reading model were MLAT and *Role at Post*: higher MLAT scores predicted higher proficiency scores in the second half of training. Misalignment between students' *Role at Post* predicted lower scores on the first and third progress assessment scores.

The effect of having taken a previous test, used in this study as a proxy for previous advanced language learning, did not significantly predict any of the assessment scores, nor did it have a significant total effect on the EOT score.

Research Question 2: Differences in Influence of Individual Differences between Reading and Speaking

In both speaking and reading, aptitude as measured by MLAT was the most consistent predictor of proficiency test scores. MLAT significantly and positively predicted performance on progress assessment scores and the EOT assessment scores. In speaking, MLAT predicted performance at earlier stages (PA 1 and 2), compared to reading (PA 2 and 3). In speaking, students whose curricula were less aligned with their job tasks (students with specialized jobs) scored lower on the second progress assessment, and scored lower at the end of training (though this end of training effect approached significance). In contrast, in reading, students misalignment between curricula and job tasks had significantly poorer scores on the first and third progress assessments in reading, and scored higher at the end of training (though this end of training effect approached significance). In speaking, learners in their first tour of language learning at the LTF scored higher on the first progress assessment than non-first time students, but there was no significant effect of being a first tour learner on any following assessments. In reading, no significant effect of being a first tour learner was observable on any assessments.

Research Question 3: Individual Differences and Difficulty Moderation in Language Proficiency Attainment

When difficulty was added as an independent predictor of proficiency scores and as a variable moderating the relationships of individual differences to proficiency scores, both the reading and speaking models significantly improved. This means that using difficulty in the prediction of proficiency in this way described variance in the data above and beyond using just individual differences.

When difficulty was added to the speaking model, the most notable predictors of proficiency scores were learners' MLAT scores, whether they were first tour learners, their roles at post, and difficulty. MLAT positively predicted assessment scores in the first half of training and at EOT. Learners whose curricula were misaligned with their job roles scored lower on their second progress assessment and at their end of training. This effect was moderated by difficulty at the third progress assessment: learners whose jobs were misaligned with the curricula they studied had especially lower scores on the third progress assessment. Learners who were on their *First Tour* of language training at the LTF scored higher than returning learners at the start of training, then lower on their second progress assessment and at EOT. Learners who studied more difficult languages scored lower on their third progress assessment. Finally, whether learners had taken a *Previous Test* was not related to their proficiency scores.

Reading

When difficulty was added to the speaking model, the most notable predictors of proficiency scores were learners' MLAT scores, their roles at post, and difficulty.

MLAT was again the most consistent predictor of proficiency assessment scores, positively predicting proficiency in the second half of training and at the end of training. Learners whose curricula were misaligned with their job roles scored lower on their third progress assessment and at their end of training. Among learners who had taken tests previously at LFT, learners who studied more difficult languages scored lower on their first progress assessment. Among learners who were studying a language at the LTF for first time, learners who studied more difficult languages scored lower on their third progress assessment.

Research Question 4: Differences in Individual Differences and Difficulty Moderation in Language Proficiency Attainment between Speaking and Reading

Aptitude as measured by the MLAT was the most consistent predictor across both modalities. MLAT scores were more associated with speaking proficiency scores in early training in speaking, compared to the later effect in reading. In both reading and speaking, learners whose curricula weren't aligned with their job tasks scored lower in the first half of training. Difficulty moderated this effect of misalignment later in training: learners with specialized jobs who studied less difficult languages scored higher on their third progress assessment and on the end of training test. In both reading and speaking there was relationship between whether or not a learner had taken a *Previous Test* at LTF with any of the proficiency scores. Finally, in speaking, learners on their first tour of language study at LTF scored higher on their PA 1 than returning learners. In speaking, first tour learners studying less difficult languages scored higher on their second progress assessment and EOT assessment, while the scores of returning learners were not significantly impacted by the difficulty of the language. In reading, effect of difficulty was found on the third progress assessment and the EOT assessment scores among first tour learners but not returning learners.

Discussion

Overall, the current investigation presents a robust analysis of individual differences' influence on intensive L2 acquisition, and demonstrates that any such analysis cannot be performed without examining the effects of language characteristics on proficiency outcomes. By using difficulty as a moderating variable, it was demonstrated that language features can confound results of crosslinguistic individual difference analyses, as relationships between individual differences and proficiency assessment scores throughout training shifted between models with and without the difficulty moderator. Furthermore, different individual differences that predicted proficiency at different times in training: while aptitude was consistently predictive of assessment scores throughout training, learner's previous interlanguage acquisition and the degree to which their training aligned to their professional work predicted proficiency at different points in training. Finally, different types of learners were affected by the difficulty of the language they were studying more than others: learners who were studying a language at the LFT for the first time and learners whose job roles were less related to the training content were especially challenged by more difficult languages. The findings support an understanding of adult L2 acquisition that more fully considers learners' goals and previous L2 experiences and consideration of the impact that difficulty can have on individual learners' abilities to achieve target proficiency goals.

Difficulty

The addition of a key L2 characteristic, difficulty, as an independent exogenous variable and as moderating variable on the individual differences, provided a more nuanced understanding of the relationships of individual differences to language proficiency scores throughout an intensive extended training period. This addition demonstrated that language difficulty explains the patterns of proficiency attainment throughout training beyond what is explained by individual student differences. This difference held out over both reading and speaking modalities: this indicates that any discussion of language proficiency gains cannot be held without consideration of the language characteristics and the difficulty of learning the specific languages involved. Furthermore, difficulty acts as a confounding variable: once the effects of L2 difficulty on the proficiency assessments are controlled for through its inclusion as an exogenous variable and moderating variable on the effects of individual differences, several nonsignificant causal paths from individual differences to assessment scores become significant. In this way, controlling for difficulty removes "noise" from the individual differences model, allowing for clearer understanding of the actual effects of individual differences.

Notably, difficulty was most often predictive of proficiency assessment scores when included as a moderating variable: across reading and speaking difficulty moderated the relationships of students' roles at post with several proficiency assessment scores across the duration of their training. The only case in which difficulty was directly related to a proficiency score through a significant causal path was in its relationship with progress assessment 3 in both speaking and reading. In both modalities, the negative sign of the path coefficient indicated that learners of less difficult languages scored lower on progress assessment 3. This unexpected finding contrasts with established crosslinguistic literature regarding the opposite effect of difficulty on long-term language study. A hypothesized explanation for this finding is found in literature regarding rater leniency: raters have been found to sympathetically score learners who attempt more difficult tasks (Morgan & Rotthoff, 2014, p. 1025), especially in language assessments in which the raters have L2 experience (Winke, Gass, & Myford, 2014, p. 38). Furthermore, raters may have been especially sympathetic at the third progress assessment due to their familiarity with the students after at least half a year of working with them and knowledge of their stress from the upcoming EOT assessment. These factors are all especially applicable in the case of difficult languages at LTF: more difficult languages tend to have smaller student populations with smaller numbers of staff, intensifying the raters' personal familiarity with the students. This factor is not often an issue with the EOT assessment, which is administered by an external testing unit.

Individual Differences

The most consistently predictive individual difference included in the models was aptitude, as measured by MLAT, in concordance with SLA literature. Notably, the relationship of MLAT scores to progress assessments remained consistent between Models 1 and 2. Furthermore, difficulty did not have a moderating effect on the relationship between MLAT and any proficiency assessment score. This finding of a lack of interaction corresponds to previous research by Mackey (2023, p. 204), and indicates that L2 learning aptitude is influential regardless of the difficulty of the

L2 studied. An unexpected finding was the lack of significant relationship between MLAT and the first progress assessment in reading: it is hypothesized that at this beginning stage, an excluded individual difference predicts L2 reading.

Learner's previous experience with L2 learning was related to their proficiency gains and achievement in training at different points in time and was highly interrelated with difficulty. In speaking, learners on their first tour of study at LTF demonstrated higher scores on their first progress assessments, but following this, scored lower on proficiency assessments than returning learners. This may demonstrate higher motivation in a novel language learning situation, and a lack of previous experience with which to associate negative language learning memories and diminish motivation. This initial strong performance followed by lower performance follows the pattern of "u-shaped behavior" in L2 acquisition, in which learning in a specific domain is first characterized by "targetlike" performance, then poorer "deviant" performance over time, before ultimately returning to "targetlike" performance (Kellerman, 1985, p. 346). This drop in performance may be attributed to more rote learning or "chunking" in production at the start, resulting in higher performance, followed by cognitive development of an interlanguage and variable application of these learned forms to novel utterances in the specific linguistic domain. This then leads to lower performance before the form is fully acquired and production returns to targetlike performance (ibid., p. 352), or development fossilizes and never becomes targetlike (Selinker, 1972; Ortega, 2011, p. 96). This "deviant", or irregular, performance may be due to the curriculum not matching the learner's "internal syllabus": this describes the learners' developmental ability to learn target sequences (Corder, 1967): if the syllabus or curriculum followed in a course is not aligned with the learner's internal syllabus and their ability to acquire a specified form, poorer performance can be expected (Ortega, 2011, p. 99).

Furthermore, returning LTF language students may be more familiar with the language training extension system which is utilized by a notable proportion of LTF students, and may understand and accept the effects of slower than recommended language proficiency growth, lessening their motivation or lowering their effort. The fact that this effect appears only in speaking may reflect the closer resemblance of the speaking assessments to work at post, because new students may be particularly encouraged by the heightened face validity of the speaking assessments. The other effect of previous language learning on PA 1 scores is the moderating effect of difficulty on reading scores. Among learners who had taken an LFT test previously, learners studying less difficult languages scored higher on reading assessments. This is explained by their having access to previous experiences interpreting L2 written language, and this experience being more accessible and applicable to L3s that were less difficult for English learners, and thus more likely to follow orthographic systems they were familiar with.

However, in assessments during the second half of training, the effects of previous interlanguages were moderated by difficulty, which is to say that the relationship between having previous interlanguages did not affect all students' assessment scores equally. First tour learners who studied less difficult languages in both speaking and reading, scored higher their mid-studies PA 2 and in total effect on their EOT assessments than first tour learners who studied more difficult languages.

Similarly, learners who had taken a previous LTF test scored higher reading PA 2 and EOT when they studied less difficult languages. This evidence of the influence of interlanguage demonstrates support for positive transfer of prior interlanguage and the positive effect of metalinguistic knowledge of curricula and assessment practices on language proficiency in subsequent L3 learning. This effect appears cumulative, as these effects related to prior L2 learning were found primarily on second and third progress assessments and on EOT. This indicates that although prior interlanguages may influence comprehension and production of individual language forms at earlier stages as widely attested in SLA literature, the cumulative effect of having prior interlanguage and language study experience is positive following initial stages of acquisition.

The alignment of the tasks students studied and were tested on to their future work at post was consistently negatively related to proficiency scores. Learners with specialized roles at post whose work tasks were less related to the tasks they studied scored lower at the first and second progress assessments in both speaking and reading. During later assessments in training, this effect was present among learners with specialized roles who were studying more difficult languages: the less difficult the language, the higher learners with specialized job roles scored on PA 3 and at EOT, again in both speaking and reading. Several conclusions may be drawn from these findings. First, the alignment of what students learn with their language use goals is associated with their proficiency gains in both reading and speaking. In the context of this study, this alignment is considered to drive student motivation, which influences their acquisition. Second, learners who are less motivated because the tasks they are learning are misaligned with their language use goals are especially impacted by language difficulty at later stages in language learning. The additional effort and stress of learning a more difficult language to an advanced level compounds the effect of lower motivation, lowering scores further. Finally, these findings suggest that the use of proxies for motivation such as alignment of curricula to language use outcomes in the place of traditional measures of self-reported motivation may be a reliable method of measuring L2 motivation in the context of assessing its impact on proficiency outcomes.

Practical Implications and Recommendations

The most consistent predictor of language proficiency gains during study is aptitude as measured by MLAT. While students at LTF are not selected for their MLAT scores, LTF should maintain records of the scores and to understand their implications regarding which students may require additional support or additional training time. Utilizing cohort population's mean MLAT scores may aid in predicting academic resource needs and on-time graduation rates. The lack of interaction with difficulty indicates that there is a lack of support for policies in other government training institutions (such as DLI) which select students who have already been selected for language training due to high aptitude for training in difficult languages based on MLAT scores.

While MLAT drives speaking and reading achievement as measured by the progress assessments and EOT scores, it is not the only student individual difference to do so. Considering the effect of misalignment between the tasks students study and the work of their roles at post, consideration should be given to the content of the training for these students to increase relevance of lessons and to align assessment tasks to their roles to increase motivation and engagement. This is especially key among students studying harder languages, who may be especially unmotivated due to misalignment with their ultimate language use goals. In extreme cases where learners will not consistently need to use the target language, consideration should be given to whether they should be trained on specific tasks in the L2, rather than to the high proficiency level of ILR 3. Furthermore, regular evaluation of the language needs of jobs should be continued to make sure that their target proficiency levels are appropriately identified: if the required proficiency level is misaligned with training, the specified target level should be reconsidered.

In addition, greater attention should be paid to the influence of learners' language learning backgrounds. In speaking, the first tour advantage at the start of training should be capitalized on, with attention given to how to extend it. As previously discussed, it is hypothesized that this stronger performance on the first speaking progress assessment is due to higher motivation or unclear understanding of the growth curve of L2 learning. Whichever case, learners would benefit from a clear understanding of the L2 learning process and the efforts they should make to enhance their learning. Understanding of the process and challenges involved may mitigate the negative effect of a lack of experience at LTF in the intermediate and EOT stages of learning. A similar effect of prior language learning is seen in reading: learners who have previously taken a test score higher on their first progress assessment when they study less difficult languages, indicating that higher language difficulty impacts returning learners. Whether this is due to interlanguage interference or potential overconfidence at the first assessment, knowing they have previously learned a language to an advanced level is unknown, but similar instruction in the process of language learning especially in the challenges of their high-difficulty target L2 would be practically possible and appropriate for these learners.

The current research examines difficulty as a moderating variable through a novel approach, examining its interactions with individual differences and their impacts on exogenous proficiency scores throughout federal language training. In the course of doing so, this research lays out an argument for the validity of this approach. Messick notes that validity research should test “theoretically relevant relationships” in ways that are *appropriate*, *meaningful*, and *useful* (1989, p. 5). The fact that difficulty has theoretical and practical relationships to language learning is well-established in the SLA field and in federal language training. However, the precise nature of these relationships is not established within the context of learner individual differences, making investigation into its influence on language learning highly *appropriate*. Furthermore, given the importance of L2 difficulty to federal training practices and the high costs, time, and resource use involved in training learners to high proficiency levels, the significant results in this research *meaningfully* establish the value of using difficulty as a moderating variable. As described above, these results provide *useful* guidance for the improvement language instruction for specific learner groups within U.S. government personnel as well as providing guidance for curricular and testing reform in the federal training context. Together, these factors establish the validity of examining language learners’ achievement through the lens of language difficulty, especially in a crosslinguistic adult learning context such as federal language training.

Finally, the results raise several concerns regarding the progress assessments. First, the reliability and validity of the second progress assessment should be examined, especially if it is specifically used to estimate progress toward their target proficiency level at EOT. Its lack of significant relationship to the EOT assessment is concerning, especially since first progress assessment scores had a consistently significant relationship to EOT scores across all models fit in the current investigation. Second, the finding that learners of less difficult languages scored lower on their third and final progress assessment in speaking and reading is concerning in the light of SLA literature. In the current study this is explained as sympathetic raters exhibiting difficulty bias, and consideration should be given to rater training and assessment specifications in order to account for these irregularities.

Chapter 9: Limitations and Future Research

Research Setting and Study Population

The LTF Setting

While the LTF presents a uniquely controlled setting in which to study adult second language acquisition, its uniqueness also limits the generalizability of this study. First, few adult language learners are granted the opportunity to study under such intensive conditions with such high levels of motivation. The long-term, full-time intensive study courses at the LTF are further unique due to the analytic syllabus and scenario-based curriculum designed to train language students to complete typical tasks of a U.S. federal employee working abroad. Further limitations are due to the restricted class sizes at the LTF: the target class size at the LTF is three students, with exceptions made due to instructional staff shortages, emergency online instruction periods, and interventions of class size reduction as a method of support for under-achieving students. This unusually small class size indicates that caution should be used when interpreting results for larger classes. In addition, the differences in the distribution of languages between the reading and speaking datasets indicate that comparisons between the two modalities should be limited. The specific languages and language programs students attended are not controlled for in this study. Because the reading data were less homogenous, generalizations regarding these findings may be less applicable outside of the current context, but the degree to which this is true is interdeterminate.

Finally, the LTF assessments are unique and findings cannot necessarily be compared to other proficiency assessments. Neither the LTF reading or speaking assessments have had an analysis of their concurrent nor predictive validity, limiting the generalizations that should be made regarding these findings to other settings. Both were developed at the LTF for the specific purpose of assessing ability to perform language tasks for specific job roles. While the methods of role play and oral assessments are well-attested in literature (Kasper & Youn, 2018; Kormos, 1999; Bachman & Savignon, 1986), a literature review revealed no other L2 reading assessment consisting of reporting the content of an L2 reading passage in the L1, indicating that the reliability and validity of this method of assessment has not been supported by previous research. This non-canonical assessment method, especially in light of the lack of concurrent and predictive validity analyses, limits generalizability of the current findings to wider L2 reading contexts.

LTF Population

A primary limitation of this research is the self-selected nature of the LTF population and the populations of federal employees. Eligibility for the majority of the LTF population is limited to U.S. citizens between 20 and 60 who: are able and willing to work abroad; able to pass a series of qualifying tests and interviews; and obtain a security clearance (U.S. Department of State, 2023g). Beyond these limitations, these employees are self-selected from the population of United States citizens, and go through multiple rounds of selection processes, including: written narrative applications, group and individual interviews, and a background and security clearance investigation. During the period of this study, this process eliminated the vast majority of candidates at either the first or second stage of the selection process (American Foreign Service Association, 2022), indicating that those who pass it are likely adults with time and inclination to study complex and detailed content, often college-educated adults who do not come from lower socio-economic status backgrounds. Language learning outcomes have been related to general intelligence (Silva and White, 1993), and if this selection process has created a study population with a higher average intelligence than the general population, it could potentially make the results less generalizable than ideal. The written application materials review and interview questions further narrow the field of candidates and likely introduce bias (American Foreign Service Association, 2022). Finally, security and suitability clearances require past history without significant drug use, crime, or association with foreign citizens from select countries that are not friendly to the United States.

Unobserved Exogenous Variables

A major limitation of any causal analysis is the influence of unobserved exogenous variables, also referred to as errors. These unmeasured or excluded variables that are unaccounted for in a model or study may significantly impact the outcomes modeled, and as such must be considered a significant limitation. While the models discussed in this study attempt to incorporate multiple influences on student proficiency outcomes including student background and cognitive measures, instructor quantities, and second language typologies, unmeasured variables likely strongly influence student learning. Among these sources of error are language learning orientation, general cognitive ability, gender, and language preference.

Methodological Limitations

The outcomes in the current investigation should not be interpreted as part of a proficiency growth curve, but interpreted as the differences on four temporally ordered assessment scores. Furthermore, the comparisons between speaking and reading outcomes in the current investigation were performed through comparison of the similarity of the significance and direction of the causal paths between the models fit using data from the two modalities. Typically, such an analysis would be performed utilizing a multi-group structural equation model which would allow for direct comparison of the causal paths through formal inferential tests. However, an assumption of multi-group SEM is the independence of the groups: in this case, the speaking and reading samples overlapped due to being drawn from the same populations: 61 cases were included in the reading analysis but not the speaking, and 31 cases were included in the speaking but not reading, indicating that the majority of the cases were in both models and violating this assumption. As such, direct comparison of the parameter values of the causal paths cannot be performed, limiting comparability of causal effects between modalities. Finally, while the number of cases in the analyses met several standards for minimum sample size, the relatively small sample sizes for these analyses leave the possibility that larger numbers of cases would have more power to detect smaller effects, thus detecting relationships that were non-significant in the current findings.

Future Research

Future research in this domain should duplicate the analyses performed in this study with a more generalizable data set. A larger sample size and stronger balance of languages studied would grant more power to control for differences such as specific language typologies, rather than the more general difficulty variable. Furthermore, although the ordinal ILR scale scores were able to be used as continuous endogenous variables in this data set, a finer scale with more measurement points for assessment scores would enable detection of smaller effects. For example, using the individual items or rubrics utilized to determine the ILR scale score would enable power analysis and more specific outcomes. In addition, models that incorporate multiple language characteristics would be able to control for their effects simultaneously to establish which typological characteristics are most associated with language proficiency throughout training.

Future research should also utilize more fine-grained operationalizations of the variables used in the present investigation to increase power and interpretability. For example, utilizing continuous values rather than dichotomous variables to represent previous L2 acquisition would both align with previous research and allow for understanding of what degrees of previous language learning experience are influential on subsequent learning. Utilizing dummy variables representing categories of students trained at LTF would allow for clearer understanding of how much this variable predicts language achievement, and which particular categories of students perform more poorly, thus providing information to guide curricula and different instruction between students with different goals. Finally, given the componential nature of language aptitude, MLAT sub scores as well as aptitude measures complementary to MLAT such as Hi-LAB subtests should be utilized to determine which of these cognitive measures are most influential at which times in training and how they interact with non-cognitive individual differences.

Several additional exogenous variables should be incorporated into future research. For example the individual difference mentioned above in limitations, such as language learning orientation, general cognitive ability, gender, language preference, and language-learning grit. Furthermore, language learning environment factors should be included. Given the rising prevalence of computer-assisted language learning technology and its capabilities, the degrees and types of computer-based or virtual language training such as virtual interaction with instructors or classmates and computer-based language learning exercises should be explored to determine its influence on language proficiency growth. In addition, variables such as instructor quality and class size are of constant interest in L2 literature. The LTF environment presents an ideal controlled environment in which to examine these variables to ascertain a clearer understanding of their influence on L2 acquisition. Finally, future research should incorporate additional variables known to specifically influence L2 speaking or reading to explain the differences between the reading and speaking findings in the present investigation. For example, as described above, L1 word decoding and text comprehension (Sparks et al., 2023, p. 492-494), or L1 and L2 reading attitude (Yamashita, 2004, p. 11), or L1 print exposure (Sparks et al., 2012, p. 497), varied levels of classroom participation (Ely, 1986), willingness to communicate (Hernandez, 2010), and integrative motivation (Hernandez, 2006).

Appendix A. Language Characteristics

Table A1. Language Training Weeks and Difficulty Categories

Language	Training Weeks to ILR-3	FSI Difficulty Category	Difficulty[a]
Albanian	44	III	2†
Arabic (Formal Std)	88	IV	1.5
Armenian (Eastern)	44	III	2†
Azerbaijani	44	III	2†‡
Bulgarian	44	III	2
Burmese	44	III	1.75
Chinese (Mandarin)	88	IV	1.5
Czech	44	III	2
Danish	24	I	2.25
Dutch/Flemish	24	I	2.75
Estonian	44	III	2†
French	30	I	2.5
Georgian	44	III	2†
German	36	II	2.25
Greek	44	III	1.75
Hebrew	44	III	2
Hindi	44	III	1.75
Hungarian	44	III	2
Indonesian	36	II	2
Italian	24	I	2.5
Kazakh	44	III	2†‡
Latvian	44	III	2†
Macedonian	44	III	2†
Malay	36	II	2.75
Nepali/Nepalese	44	III	1.75
Norwegian	24	I	3
Persian/Farsi/Irania	44	III	2
Persian/Tajiki	44	III	2†
Polish	44	III	2
Portuguese	24	I	2.5
Romanian	24	I	3
Russian	44	III	2.25
Slovak	44	III	2†
Slovenian	44	III	2†
Spanish	24	I	2.25
Swahili/Kiswahili	36	II	2.75
Tamil	44	III	1.5†‡
Thai	44	III	2
Turkish	44	III	2
Turkmen	44	III	2†
Ukrainian	44	III	2†
Urdu	44	III	1.6†‡
Vietnamese	44	III	1.5

[a] Difficulty values marked by † are from LTF data request. All other difficulty values are from Chiswick and Miller (2005). Distance values marked by ‡ indicate a 21-week benchmark.

Table A2. Language Typological Characteristics

Language	Family (Institute Data)	Indo-European	Ergative[a]	Use of Postpositions[b]	Nominal Plural[c]	Orthography (Institute Data)
Albanian	Other Indo-European	Indo-European	Absolutive	Prepositions	Always mandatory (Joseph, Costanzo, & Slocum, n.d.)	Roman
Arabic (Formal Std)	Afroasiatic	Non-Indo-European	Absolutive	Prepositions	Always mandatory	Non-Roman
Armenian (Eastern)	Other Indo-European	Indo-European	Absolutive	Postpositions or No Adpositions	Always mandatory	Non-Roman
Azerbaijani	Turkic	Non-Indo-European	Absolutive (Siegel, n.d., pp. 39-42)	Postpositions or No Adpositions (Siegel, n.d., p. 97)	Always mandatory (Siegel, n.d., pp. 15-16)	Roman
Bulgarian	Balto-Slavic	Indo-European	Absolutive (Leafgren, 2012, pp. 72-74)	Prepositions	Always mandatory (Leafgren, 2012, pp. 26-32)	Non-Roman
Burmese	Sino-Tibetan	Non-Indo-European	Ergative	Postpositions or No Adpositions	Always mandatory (Schmidt, 2017, pp. 3-5)	Non-Roman
Chinese (Mandarin)	Sino-Tibetan	Non-Indo-European	Ergative	Postpositions or No Adpositions	Not always mandatory	Non-Roman
Czech	Balto-Slavic	Indo-European	Absolutive (Naughton, 2006, p. 216)	Prepositions	Always mandatory (Naughton, 2006, pp. 18-19)	Roman
Danish	Germanic	Indo-European	Absolutive (Allan et al., 2005, pp. 151-155)	Prepositions	Always mandatory (Allan et al., 2005, p. 22)	Roman
Dutch/Flemish	Germanic	Indo-European	Absolutive	Prepositions	Always mandatory	Roman
Estonian	Uralic	Non-Indo-European	Absolutive (Erelt, 2007, pp. 93-95	Postpositions or No Adpositions	Always mandatory (Viitso, 2007, p. 35)	Roman
French	Romance	Indo-European	Absolutive	Prepositions	Always mandatory	Roman
Georgian	Kartvelian	Non-Indo-European	Absolutive	Postpositions or No Adpositions	Always mandatory	Non-Roman
German	Germanic	Indo-European	Absolutive	Prepositions	Always mandatory	Roman

Language	Family (Institute Data)	Indo-European	Ergative[a]	Use of Postpositions[b]	Nominal Plural[c]	Orthography (Institute Data)
Greek	Other Indo-European	Indo-European	Absolutive	Prepositions	Always mandatory	Non-Roman
Hebrew	Afroasiatic	Non-Indo-European	Absolutive	Prepositions	Always mandatory	Non-Roman
Hindi	Indo-Aryan	Indo-European	Absolutive	Postpositions or No Adpositions	Always mandatory	Non-Roman
Hungarian	Uralic	Non-Indo-European	Absolutive	Postpositions or No Adpositions	Always mandatory	Roman
Indonesian	Austronesian	Non-Indo-European	Absolutive	Prepositions	Not always mandatory	Roman
Italian	Romance	Indo-European	Absolutive	Prepositions	Always mandatory	Roman
Kazakh	Turkic	Non-Indo-European	Absolutive (Dotton & Wagner, n.d., p. 53)	Postpositions or No Adpositions (Dotton & Wagner, n.d., pp. 65-66)	Always mandatory (Dotton & Wagner, n.d., pp. 17-18)	Non-Roman
Latvian	Balto-Slavic	Indo-European	Absolutive	Prepositions	Always mandatory	Roman
Macedonian	Balto-Slavic	Indo-European	Absolutive (Friedman, 2001, pp. 49-50)	Prepositions	Always mandatory (Friedman, 2001, p. 18)	Non-Roman
Malay	Austronesian	Non-Indo-European	Absolutive (Shehadeh Ali, n.d., p. 3)	Prepositions (Shehadeh Ali, n.d., p. 3)	Always mandatory (Malay, 2023)	Roman
Nepali/Nepalese	Indo-Aryan	Indo-European	Absolutive (Lindemann, 2019, pp. 300-301)	Postpositions or No Adpositions	Not always mandatory	Non-Roman
Norwegian	Germanic	Indo-European	Absolutive (Askedal, J. O., 1986, pp. 25–45)	Prepositions	Always mandatory (Strandskogen & Stranskogen, 1995, p. 66)	Roman
Persian/Farsi/ Iranian	Iranian	Indo-European	Absolutive	Prepositions	Not always mandatory (Roberts, 2004, p. 13)	Non-Roman
Persian/Tajiki	Iranian	Indo-European	Absolutive (Khojayori & Thompson, 2009, p. 55)	Prepositions	Not always mandatory (Khojayori & Thompson, 2009, pp. 21-22)	Non-Roman

Language	Family (Institute Data)	Indo-European	Ergative[a]	Use of Postpositions[b]	Nominal Plural[c]	Orthography (Institute Data)
Polish	Balto-Slavic	Indo-European	Absolutive	Prepositions	Always mandatory (Feldstein, 2001, pp. 38-56)	Roman
Portuguese	Romance	Indo-European	Absolutive (Galves & Avelar, 2021, pp. 98-99)	Prepositions	Always mandatory	Roman
Romanian	Romance	Indo-European	Absolutive (Tomescu, n.d., 40 – 41)	Prepositions	Always mandatory	Roman
Russian	Balto-Slavic	Indo-European	Absolutive	Prepositions	Always mandatory	Non-Roman
Slovak	Balto-Slavic	Indo-European	Absolutive (Short, 1980, pp. 565-566)	Prepositions (Slovak, 2010)	Always mandatory (Short, 1980, p. 543)	Roman
Slovenian	Balto-Slavic	Indo-European	Absolutive (Greenberg, 2006, p. 124)	Prepositions	Always mandatory (Greenberg, 2006, pp. 29-36)	Roman
Spanish	Romance	Indo-European	Absolutive	Prepositions	Always mandatory	Roman
Swahili/Kiswahili	Other	Non-Indo-European	Absolutive	Prepositions	Always mandatory	Roman
Thai	Austroasiatic	Non-Indo-European	Ergative	Prepositions	Always mandatory (Jenks, 2011, pp. 101-102)	Non-Roman
Turkish	Turkic	Non-Indo-European	Absolutive	Postpositions or No Adpositions	Always mandatory	Roman
Turkmen	Turkic	Non-Indo-European	Absolutive (Gray, 2011, p. 27)	Postpositions or No Adpositions	Always mandatory (Peace Corps, n.d., p. 20)	Non-Roman
Ukrainian	Balto-Slavic	Indo-European	Absolutive (Matthews, 1949, p. 117)	Prepositions	Always mandatory (Pelishenko, 2006, p. 25-27)	Non-Roman
Urdu	Indo-Aryan	Indo-European	Ergative (Uddar, 2014, pp. 2-3)	Postpositions or No Adpositions	Always mandatory (Ahmed, 2014, p. 2846)	Non-Roman
Vietnamese	Austroasiatic	Non-Indo-European	Ergative	Prepositions	Not always mandatory	Roman

Table A3. Distribution of Language Categories: Speaking

	I		II		III		IV	
	n	%	n	%	n	%	n	%
Dataset with Complete Cases Only	287	81.07%	18	5.08%	49	13.84%	-	-
Dataset with Incomplete Cases	184	69.70%	11	4.17%	56	21.21%	13	4.92%
Total	471	76.21%	29	4.69%	105	16.99%	13	2.10%

Table A4. Distribution of Language Categories: Reading

	I		II		III		IV	
	n	%	n	%	n	%	n	%
Dataset with Complete Cases Only	282	80.57%	18	5.14%	50	14.29%	-	-
Dataset with Incomplete Cases	226	75.84%	10	3.36%	49	16.44%	13	4.36%
Total	508	78.40%	28	4.32%	99	15.28%	13	2.01%

Appendix B. Model Covariance Tables

Table B1. Speaking Covariance Table

	PA 1	PA 2	PA 3	EOT	MLAT	Role at Post	Previous Test	First Tour	Difficulty	MLAT * Difficulty	Role at Post * Difficulty	Previous Test * Difficulty	First Tour * Difficulty
PA 1	0.084												
PA 2	0.046	0.103											
PA 3	0.033	0.049	0.098										
EOT	0.058	0.063	0.075	0.185									
MLAT	0.418	0.673	0.675	1.115	78.498								
Role at Post	-0.008	-0.02	-0.018	-0.03	-1.246	0.157							
Previous Test	0	0.014	0.025	0.022	1.068	-0.049	0.187						
First Tour	0.016	0.002	-0.012	-0.007	-0.186	-0.032	-0.088	0.177					
Difficulty	0.013	0.024	-0.006	0.012	-0.258	0.012	-0.012	0.013	0.074				
MLAT * Difficulty	0.047	0.053	-0.008	0.091	0.729	-0.017	-0.038	0.01	0.182	5.161			
Role at Post * Difficulty	0	0.002	0.003	0.002	-0.067	0.01	-0.001	-0.002	0.007	-0.032	0.007		
Previous Test * Difficulty	0.012	0.023	-0.005	0.011	-0.231	0.01	-0.003	0.01	0.067	0.199	0.005	0.067	
First Tour * Difficulty	0.003	0.005	0	0.004	-0.049	-0.002	-0.002	0.01	0.009	-0.009	0.001	0.005	0.009

Table B2. Reading Covariance Table

	PA 1	PA 2	PA 3	EOT	MLAT	Role at Post	Previous Test	First Tour	Difficulty	MLAT * Difficulty	Role at Post * Difficulty	Previous Test * Difficulty	First Tour * Difficulty
PA 1	0.149												
PA 2	0.077	0.131											
PA 3	0.007	0.021	0.054										
EOT	0.056	0.057	0.032	0.233									
MLAT	0.143	0.454	0.513	1.26	80.473								
Role at Post	-0.013	-0.022	-0.019	-0.014	-1.069	0.139							
Previous Test	-0.009	0.01	0.012	0.012	1.075	-0.05	0.195						
First Tour	0.012	0.009	0.001	-0.002	-0.123	-0.028	-0.087	0.183					
Difficulty	0.02	0.026	-0.011	0.038	-0.204	0.009	-0.011	0.012	0.07				
MLAT * Difficulty	0.072	0.106	0.009	0.163	1.162	-0.028	-0.027	-0.003	0.172	4.578			
Role at Post * Difficulty	0	0	-0.001	0.004	-0.062	0.007	-0.002	-0.002	0.005	-0.027	0.005		
Previous Test * Difficulty	0.02	0.026	-0.009	0.035	-0.177	0.006	-0.003	0.009	0.063	0.189	0.003	0.063	
First Tour * Difficulty	0.001	0.003	0.002	0.008	-0.053	-0.001	-0.002	0.009	0.01	-0.004	0.001	0.007	0.01

Bibliography

Ahmed, T. (2014, May). Automatic acquisition of Urdu nouns (along with gender and irregular plurals). In LREC (pp. 2846-2850).

Alhija, F. A. N. (2010). Factor analysis: An overview and some contemporary advances. *International encyclopedia of education, 3*, 162-170.

Allan, R., Holmes, P., & Lundskær-Nielsen, T. (2005). *Danish: An Essential Grammar*. Routledge. https://ia601303.us.archive.org/17/items/DanishGrammar/Danish%20Grammar.pdf

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology, 112*(4), 545.

American Foreign Service Association. (2022, October). *FSO Selection: Changing the Path to the Oral Assessment*. The Foreign Service Journal. https://afsa.org/fso-selection-changing-path-oral-assessment

Askedal, J.O. (1986). On Ergativity in Modern Norwegian. Nordic Journal of Linguistics, 9(1), pp. 25 – 45. https://doi.org/10.1017/S0332586500001396

Bachman, L. F., & Savignon, S. J. (1986). The Evaluation of Communicative Language Proficiency: A Critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380–390. https://doi.org/10.2307/326817

Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: the case of Germanic syntax. *Second Language Research, 23*(4), 459-484. https://doi.org/10.1177/0267658307080557

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186–203. doi: 10.1207/s15328007sem1302_2

Bentler, P.M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 88,* pp. 588-606.

Bentler, P.M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bermudez-Mendez, J. (2020). *Student success factors at Defense Language Institute Foreign Language Center*. [Masters Thesis, United States Naval Postgraduate School]. Defense Technical Information Center. https://apps.dtic.mil/sti/citations/AD1114144

Boomsma A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50.* pp. 229–242.

Brown, T. (2023). Confirmatory factor analysis. In R.H. Hoyle, *Handbook of Structural equation modeling* (pp. 261–276). The Guilford Press.

Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research : LTR, 23*(5), 603–632. https://doi.org/10.1177/1362168817744389

Busse, V., & Walter, C. (2013). Foreign language learning motivation in higher education: A longitudinal study of motivational changes and their causes. *The modern language journal, 97*(2), 435-456.

Carroll, J. B. (1973). Implications of aptitude test research and psycholinguistic theory for foreign language teaching. *Linguistics, 11*(112), 5–14. https://doi.org/10. 1515/ling.1973.11.112.5

Carroll, J. B., Sapon, S. M., Reed, D. J., & Stansfield, C. W. (2010). *Modern Language Aptitude Test: Manual*. Rockville, MD: Second Language Testing Foundation.

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.

Chua, H. W., & Lin, C. Y. (2020). The Effect of Task-Based Language Teaching in Learning Motivation. *International Journal on Social and Education Sciences, 2*(1), 41-48.

Clark, K. N., & Malecki, C. K. (2019). Academic Grit Scale: Psychometric properties and associations with achievement and life satisfaction. *Journal of School Psychology, 72*, 49–66. https://doi.org/10.1016/j.jsp.2018.12.001

Clark, J. L. D., & Clifford, R. T. (1988). THE FSI/ILR/ACTFL proficiency scales and testing techniques: Development, current Status, and needed research. *Studies in Second Language Acquisition, 10*(2), 129–147. http://www.jstor.org/stable/44488169

Clifford, R. T., & Fischer, D. C. (1990). Foreign language needs in the U. S. government. *The Annals of the American Academy of Political and Social Science, 511*, 109–121. http://www.jstor.org/stable/1047374

Cohen, J. (1988). A power primer. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 279–284). American Psychological Association. https://doi.org/10.1037/14805-018

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330.

Corder, S. P. (1967). The significance of learner's errors. *IRAL: International Review of Applied Linguistics in Language Teaching, 5*(4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Cysouw, M. (2013). Predicting language-learning difficulty. In L. Borin & A. Saxena (Eds.), *Approaches to Measuring Linguistic Differences* (pp. 57–82). De Gruyter. https://doi.org/10.1515/9783110305258.57

Defense Language Institute Foreign Language Center. (2023). Mission, vision, and values. Defense Language Institute Foreign Language Center. https://www.dliflc.edu/about/mission-vision/

Defense Language Institute Foreign Language Center. (2023a). DLI-Washington. Defense Language Institute Foreign Language Center. https://www.dliflc.edu/about/dli-washington/

Defense Language Institute Foreign Language Center. (2023c). Continuing Education. https://www.dliflc.edu/academics/continuing-education-2/

DeKeyser, R. (2016). Of moving targets and chameleons: Why the concept of difficulty is so hard to pin down. *Studies in Second Language Acquisition, 38*(2), 353–363. https://www.jstor.org/stable/26330923

Depaoli, S., Kaplan, D., & Winter, S.D. (2023). Foundations and extension of Bayesian structural equation modeling. In R.H. Hoyle, *Handbook of Structural equation modeling* (pp. 261–276). The Guilford Press.

Dörnyei, Z. (2019). Task motivation. In Ahmadian, M.J., & Wen, Z. (E.), *Researching L2 task performance and pedagogy: In honour of Peter Skehan Volume 13*, (pp. 53-66). Philadelphia: John Benjamins Publishing Company.

Dotton, Z., & Wagner, J. D. (n.d.). A Grammar of Kazakh. Duke University. https://slaviccenters.duke.edu/sites/slaviccenters.duke.edu/files/file-attachments/kazakh-grammar.pdf

Doughty, C. J. (2013). Optimizing post-critical-period language learning. In G. Granena & M. H. Long (Eds.)*, Sensitive periods, language aptitude, and ultimate L2 Attainment* (pp. 153–174). Oxford, England: Wiley-Blackwell.

Doughty, C. J. (2019). Cognitive language aptitude. *Language Learning, 69*, 101–126.

Dryer, M. S. (2013). Order of adposition and noun phrase. In: Dryer, Matthew S. & Haspelmath, Martin (eds.). *WALS Online* (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

(Available online at http://wals.info/chapter/85, Accessed on 2024-02-09.)

Edwards, J. D. (2004). The role of languages in a post-9/11 United States. *The Modern Language Journal*, 88(2), 268–271. http://www.jstor.org/stable/3588756

Ehrman, M. (1994). A Study of the Modern Language Aptitude Test for predicting learning success and advising students. In *Language Aptitude Invitational Symposium Program Proceedings* (pp. 74-90).

Ehrman, M., Leaver, B, & Oxford, R. (2003). A brief overview of individual differences in second language learning. *System*. 31. 313-330. 10.1016/S0346-251X(03)00045-9.

Ehrman, M., & Oxford, R. (1990). Adult language learning styles and strategies in an intensive training setting. *The modern language journal, 74*(3), 311-327.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition Plus: Correlates of Language Learning Success. *The Modern Language Journal*, *79*(1), 67–89. https://doi.org/10.2307/329394

Ely, C. M. (1986). An analysis of discomfort, risktaking, sociability, and motivation in the L2 classroom. *Language learning, 36*(1), 1-25.

Enders, C. (2013). Analyzing structural equation models with missing data. In *Structural equation modeling: A second course* (2nd ed., pp. 493–519). essay, Information Age Publishing.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling, 8*(3), 430-457.

Erelt, M. (2007). Structure of the Estonian Language. Syntax. In M. Erelt (Ed), *Estonian Language*. Estonian Academy Publishers. https://dspace.ut.ee/server/api/core/bitstreams/adbce35b-935e-4ef8-afd3-e3bf44f79040/content

Falk, Y., & Bardel, C. (2011). Object pronouns in German L3 syntax: Evidence for the L2 status factor. *Second Language Research, 27*(1), 59-82. https://doi.org/10.1177/0267658310386647

Feldstein, R. (2001). *A Concise Polish Grammar*. SEELRC. http://seelrc.org:8080/grammar/pdf/compgrammar_polish.pdf

Finney, S. J., DiStefano, C., & Kopp, J. P. (2016). Overview of estimation methods and preconditions for their application with structural equation modeling. In *Principles and methods of test construction: Standards and recent advances*, Schewizer, K. & DiStefano, C., eds. pp. 135-165.

Flynn, S., Foley, C., & Vinnitskaya, I. (2004). The Cumulative-Enhancement Model for Language Acquisition: Comparing Adults' and Children's Patterns of Development in First, Second and Third Language Acquisition of Relative Clauses. *International Journal of Multilingualism, 1*(1), 3–16.

Foreign Affairs Manual, 3 FAM 2218.2-1, Certification of need (2023a).

Foreign Affairs Manual, 5 FAM 470, Accessing and using Department of State information (2023b).

Friedman, V. A. (2001). *Macedonian*. SEELRC. http://www.seelrc.org:8080/grammar/pdf/compgrammar_macedonian.pdf

Galves, C. & Avelar, J. (2021). Case and agreement in Brazilian Portuguese: Between Bantu and Romance. In Bárány, A., Biberauer, T., Dougla, J. & Vikner, S. (eds.), *Syntactic architecture and its consequences III: Inside syntax*, 95–124. Berlin: Language Science Press. DOI: 10.5281/zenodo.4680304

Gardner, R. C. (1985). Social psychology and second language learning: the role of attitude and motivation. Baltimore, MD: Edward Arnold.

Gardner, R. C. (2001). Language learning motivation: the student, the teacher, and the researcher. *Texas Papers in Foreign Language Education*, *6*(1), 1-18.

Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second-language acquisition. *Canadian Journal of Psychology / Revue canadienne de psychologie, 13(*4), 266–272. https://doi.org/10.1037/h0083787

Gardner, R. C., & Lambert, W. E. (1972). Attitudes and Motivation in Second Language Learning. Rowley, MA: Newbury House Publishers.

Goshko, K (1986, November 12.) Diplomats Weak in Vital Languages. *The Washington Post.* https://www.washingtonpost.com/archive/politics/1986/11/12/diplomats-weak-in-vital-languages/c838de32-a9f6-4923-96a5-b81ab814380e/

Government Accountability Office, (2017). Foreign Language Proficiency Has Improved, but Efforts to Reduce Gaps Need Evaluation 9–9. Washington, DC; Government Accountability Office. https://www.gao.gov/assets/gao-17-318.pdf

Government Accountability Office, (2002). Foreign languages: Human capital approach needed to correct staffing and proficiency shortfalls. Washington, DC; Government Accountability Office. https://www.gao.gov/assets/gao-02-375.pdf

Government Accountability Office, (2003). U.S. public diplomacy: State Department expands efforts but faces significant challenges. Washington, DC;

Government Accountability Office. https://www.gao.gov/assets/gao-03-951.pdf

Government Accountability Office, (2023). Enhanced training, analysis, and monitoring could improve foreign language proficiency. Washington, DC; Government Accountability Office. https://www.gao.gov/assets/gao-24-105849.pdf

Granena, G. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied psycholinguistics, 37*(3), 577-600.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second language research, 29*(3), 311-343.

Gray, D. (2011). A Short Descriptive Grammar of the Turkmen Language.

Greenberg, M. L. (2006). *A short reference grammar of Standard Slovene*. SEELRC Reference Grammar Network.

Grigorenko, E. L., Sternberg, R., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal, 84*, 390–405.

Hart-Gonzalez, Lucinda and Stephanie Lindemann, (1993). “Expected Achievement in Speaking Proficiency, 1993,” School of Language Studies, Foreign Services Institute, Department of State, mimeo.

Haspelmath, M. (2013). *Occurrence of nominal plurality*. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: a Regression-Based Approach*. The Guilford Press.

Hernández, T. (2006). Integrative motivation as a predictor of success in the intermediate foreign language classroom. Foreign Language Annals, 39, 605–617.

Hernández, T. A. (2010). The relationship among motivation, interaction, and the development of second language oral proficiency in a study-abroad context. *The Modern Language Journal, 94*(4), 600-617.

Herzog, M. (n.d.). *An overview of the history of the ILR Language proficiency skill level descriptions and scale*. History of the ILR. https://govtilr.org/Skills/IRL%20Scale%20History.htm

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign Language Classroom Anxiety. *The Modern Language Journal, 70*(2), 125–132. https://doi.org/10.2307/327317

Housen, A., & Simoens, H. (2016). INTRODUCTION: COGNITIVE PERSPECTIVES ON DIFFICULTY AND COMPLEXITY IN L2 ACQUISITION. *Studies in Second Language Acquisition, 38*(2), 163–175. https://doi.org/10.1017/S0272263116000176

Hoyle, R. H. (2023). Structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 3–16), The Guilford Press.

Hoyle, R. H., & Pearl, J. (2023). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 49–75), The Guilford Press.

IBM Corp. (2022). IBM SPSS Statistics for Windows (Version 28.0) [Computer software]. IBM Corp.

Interagency Language Roundtable. (n.d.). *Interagency Language Roundtable Language Skill Level Descriptions - Listening*. ILR Listening Scale Stage. https://www.govtilr.org/Skills/ILRscale3.htm

Jenks, P. S. E., (2011). The hidden structure of Thai noun phrases [Doctoral dissertation, Harvard University]. http://www.linguistics.berkeley.edu/~jenks/Research_files/Jenks_2011.pdf

Joseph, B., Costanzo, A., & Slocum, J. (n.d.). *The Albanian Noun*. Albanian Online. https://lrc.la.utexas.edu/eieol/albol/10#grammar_104

Kasper, G., & Youn, S. J. (2018). Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review, 9*(4), 589-616.

Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773-795).

Kellerman, E. (1985). If at first you do succeed. *Input in second language acquisition. Rowley: Newbury House,* 345-353.

Kim, Y., Kong, D.-K., Lee, J.-H., & Lee, Y.-G. (2000). Implementation and evaluation of an approach to task-based Korean language teaching. Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i at Mānoa. [An abridged version of this paper appeared with the same title in

Korean Language in America (Vol. 6, pp. 45– 51, 2001), the Proceedings of the 6th Annual American Association of Teachers of Korean.].

Klein, E. C. (1995). Evidence for a ‘wild’ L2 grammar: when PPs rear their empty heads. *Applied linguistics, 16*(1), 87-117.

Khojayori, N., & Thompson, M. (2009). *Tajiki Reference Grammar for Beginners*. Georgetown University Press.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language testing, 16*(2), 163-188.

Language Learning and Testing Foundation. (2021, February 1). Modern language aptitude test . Language Learning and Testing Foundation . https://lltf.net/aptitude-tests/language-aptitude-tests/modern-language-aptitude-test-2/#:~:text=The%20Modern%20Language%20Aptitude%20Test,but%20particularly%20speaking%20and%20listening.

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability. *Journal of Research in Reading, 15*(2), 95-103.

Leafgren, J. (2012). *A Concise Bulgarian Grammar*. Duke University. http://www.seelrc.org:8080/grammar/pdf/stand_alone_bulgarian.pdf

Lee, J. W., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *Tesol Quarterly, 31*(4), 713-739.

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(3)*, pages 385–408. https://doi.org/10.1093/applin/amu054

Linck, J., Hughes, M., Campbell, S., Silbert, N., Tare, M., Jackson, S., et al. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63*, 530–566.

Lindemann, L. (2019). *A Jewel Inlaid: Ergativity and Markedness in Nepali* [Graduate Dissertation, Yale University]. Yale University Department of Linguistics https://elischolar.library.yale.edu/cgi/viewcontent.cgi?article=1003&context=ling_graduate

Little, R. J. A., & Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, *18*(2-3), 292-326. https://doi.org/10.1177/0049124189018002004

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley Blackwell.

Long, M. H., & Crookes, G. (1992). Three Approaches to Task-Based Syllabus Design. *TESOL Quarterly, 26*(1), 27–56. https://doi.org/10.2307/3587368

MacIntyre, P. D. (2017). An overview of language anxiety research and trends in its development. In C. Gkonou, M. Daubney, & J.-M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 11–30). Bristol, UK: Multilingual Matters.

Mackey, B. (2023). *A multilevel analysis of language proficiency growth* [Doctoral dissertation, University of Maryland]. Digital Repository at the University of Maryland. http://hdl.handle.net/1903/30718

Malay (individual language) Grammar - Plurals. (2023, March). In Polyglot Club Wiki. https://polyglotclub.com/wiki/Language/Malay-individual-language/Grammar/Plurals

Masters, M. (2018). *Pathways to proficiency: Examining the coherence of initial second language acquisition patterns within the language difficulty categorization framework* [Doctoral dissertation, University of Maryland]. Digital Repository at the University of Maryland. http://hdl.handle.net/1903/20928

Matthews, W. K. (1949). The Ukrainian System of Declension. *Language, 25*(2), 117–129. https://doi.org/10.2307/409939

Messick, S. (1984). The psychology of educational measurement. *Journal of ,Educational Measurement, 21*, 215-237.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Meyer, J., Fleckenstein, J., & Köller, O. (2019). Expectancy value interactions and academic achievement: Differential relationships with achievement measures. *Contemporary Educational Psychology*, *58*, 58-74.

Microsoft Corporation. (2018). Microsoft Excel. Retrieved from https://office.microsoft.com/excel.

Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry, 52*(3), 1014-1026.

Muthén, L. [Comparing Non-Nested Models]. (2010, June 16). You need to have the same set of observed variables to compare models. Use the full set in both analyses and fix the paths you don't want to zero. [Online forum post]. Mplus Discussion > Structural Equation Modeling. https://www.statmodel.com/discussion/messages/11/3521.html?1480430295

Muthén, L. [Correlation of exogenous variables in...]. (2011, July 25). If you include them in the analysis, they are treated as dependent variables and distributional assumptions are made about them. [Online forum post]. Mplus Discussion > Structural Equation Modeling. https://www.statmodel.com/discussion/messages/11/7833.html?1602181191#:~:text=By%20default%2C%20Mplus%20correlates%20the,the%20one%20endogenous%20latent%20factor.

Muthén, L. [Using BIC for SEM]. (2015, October 25). To use BIC to compare non-nested models, the models must have the same set of dependent variables. [Online forum post]. Mplus Discussion > Structural Equation Modeling. https://www.statmodel.com/discussion/messages/11/5682.html

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

NamazianDost, I., Bohloulzadeh, G., & Pazhakh, A. (2017). The Effect of Task-Based Language Teaching on Motivation and Grammatical Achievement of

EFL Junior High School Students. Advances in Language and Literary Studies, 8(2), 243-259.

National Security Agency. (n.d.). *Overview*. Education and Professional Development. https://www.nsa.gov/academics/national-cryptologic-university/education-and-professional-development/

National Virtual Translation Center. (2007). Language difficulty. Language Learning Difficulty for English Speakers. https://web.archive.org/web/20080430141810/http://www.nvtc.gov/lotw/months/november/learningExpectations.html

Naughton, J. (2006). *Czech An Essential Grammar*. Routledge. https://jakobson.korpus.cz/~rosen/public/GGG/Czech_essent_grammar.pdf

Newson, J. T. (2023). *Missing data and missing data estimation in SEM*. 623 Structural Equation Modeling. https://web.pdx.edu/~newsomj/semclass/ho_missing.pdf

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230-244.

Nunnally, J.C. (1967). *Psychometric theory.* New York, NY: McGraw-Hill.

O'Grady, S. (2023). Halo effects in rating data: Assessing speech fluency. *Research Methods in Applied Linguistics, 2*(2), 100048.

Office of Resource Management and Organization Analysis, Five-Year Workforce and Leadership Succession Plan Fiscal Years 2018 – 2022 49–53 (2019). Washington, DC; U. S. Department of State Bureau of Human Resources.

Ornstein, J. (1956). Structurally Oriented Texts and Teaching Methods since World War II: A Survey and Appraisal. *The Modern Language Journal, 40*(5), 213–222. https://doi.org/10.2307/320234

Ortega, L. (2009). Sequences and processes in language learning. *The handbook of language teaching,* 81-105.

Ozili, P. K. (2023). The acceptable R-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native English speakers* (pp. 134-143). IGI global.

Peace Corps. (n.d.). Turkmen Language Grammar Guide. U.S. Peace Corps Turkmenistan.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2023). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed.) (pp.184-205). The Guilford Press.

Pelishenko, O., (2006). A contrastive analysis of Ukrainian and German nominal morphology [Masters thesis, The University of Georgia]. https://getd.libs.uga.edu/pdfs/pelishenko_olena_200908_ma.pdf

Preacher, K. J., & Yaremych, H.E. (2023). Model selection in structural equation modeling. In R.H. Hoyle, *Handbook of Structural equation modeling* (pp. 261–276). The Guilford Press.

President's Commission on Foreign Language and International Studies. (1979). (rep.). *President's Commission on Foreign Language and International*

Studies: Background Papers and Studies. (pp. 197–202). Washington, DC: Office of Education (DHEW). https://files.eric.ed.gov/fulltext/ED179117.pdf

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Rhoades, E. (2023). *Investigating individual differences' prediction of language proficiency outcomes: a latent growth curve modeling approach.* [Doctoral dissertation, University of Maryland]. Digital Repository at the University of Maryland.

Robinson, P. (2002). Learning conditions, aptitude complexes, and SLA. *Individual differences and instructed language learning*, *2*, 113-133.

Robitzsch, A. (2020, October). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in education, 5*. https://doi.org/10.3389/feduc.2020.589965

Rothman, J. (2011). L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research, 27*(1), 107–127. https://doi.org/10.1177/0267658310386439

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Saito, K., Dewaele, J.-M., Abe, M. and In'nami, Y. (2018), Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning, 68*(pp. 709-743). https://doi.org/10.1111/lang.12297

Satorra, A., & Bentler, P. M. (2010). Ensuring Positiveness of the Scaled Difference Chi-square Test Statistic. *Psychometrika, 75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language learning*, *43*(3), 313-344.

Sasaki, M. (2012). The modern language aptitude test (paper-and-pencil version). *Language Testing*, *29*(2), 315-321.

Sawyer, D. (2022, October). *Language testing reforms: What you need to know*. The Foreign Service Journal. https://afsa.org/language-testing-reforms-what-you-need-know

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching,(10)* pp. 209-231.

Shehadeh Ali, A. (n.d.). Contrastive analysis of sentences of Arabic Language and Malay. International Islamic University Malaysia. https://www.academia.edu/34894405/The_Syntactical_structure_of_Malay_and_Arabic_doc

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the Model Size Effect on SEM Fit Indices. *Educational and psychological measurement, 79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Shieh, G. (2011). Clarifying the role of mean centering in multicollinearity of interaction effects. *The British Journal of Mathematical and Statistical Psychology, 64*(3), 462–477. https://doi.org/10.1111/j.2044-8317.2010.02002.x

Short, D. (1980). Slovak. In *West Slavonic Languages* (pp. 533–592). essay, Slavica Publishers.

Siegel, A. H. (n.d.). Essentials of Azerbaijani: An Introductory Course. https://minio.la.utexas.edu/webeditor-files/slavic/pdf/azeri.pdf

Siewierska, A. (2013.) *Alignment of verbal person marking.* In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533
(Available online at http://wals.info/chapter/100, Accessed on 2024-04-27.)

Silva, J. M., & White, L. A. (1993). Relation of cognitive aptitudes to success in foreign language training. *Military Psychology, 5*(2), 79–93. https://doi.org/10.1207/s15327876mp0502_1

Slovak adposition. Slovak Adposition. (2010, December 5). https://nl.ijs.si/ME/Vault/V4/msd/html/msd.S-sk.html

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-384.

Sparks, L., Dale, S., & Patton, M. J. (2023). Individual differences in L1 attainment and language aptitude predict L2 achievement in instructed language learners. *The Modern Language Journal, 107*(2), 479-508.

Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2012). Do L1 Reading Achievement and L1 Print Exposure Contribute to the Prediction of L2

Proficiency? *Language Learning, 62*(2), 473–505. https://doi.org/10.1111/j.1467-9922.2012.00694.x

Steiger, J.H., & Lind, J.C., (1980, May). *Statistically-based tests for the number of common factors.* Paper presented at annual meeting of Psychometric Society, Iowa City, IA.

Strandskogen, Åse-Berit, & Strandskogen, R. (1995). *Norwegian: An Essential Grammar* (2nd ed.) Routledge. https://tanuljunknorvegul.files.wordpress.com/2014/02/learn-norwegian-language-routledge-norwegian-an-essential-grammar.pdf

Tomescu, V. (n.d.). The acquisition of subjects in 2L1 Romanian. *Bucharest Working Papers in Linguistics XX*, *1*, 39-63. https://bwpl.unibuc.ro/wp-content/uploads/2019/03/3_BWPL_2018_special_issue_Tomescu.pdf

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology, 4*(1), 25-29.

Thomas, J. (1988). The role played by metalinguistic awareness in second and third language learning. *Journal of Multilingual & Multicultural Development, 9*(3), 235-246.

Uddar, U., Kaur, G., & Chandra, P. (2014). Revisiting morphological ergativity in Hindi-Urdu. *Acta Linguistica, 8*(1), 3–15.

University of Maryland Institutional Review Board. (6 September, 2023). [DETERMINATION OF EXEMPT STATUS]. Retrieved from https://www.irbnet.org/ (2100612-1).

U.S. Air Force Academy. (2023, October 27). *Department of Languages and Cultures Archives*. https://www.usafa.edu/department/languages-and-cultures/

U.S. Department of Defense. (2024). *About* We Are Your Defense. https://www.defense.gov/about/

U.S. Department of State. (2023a, May 3). *Foreign language training - United States Department of State.* Foreign Language Training. https://www.state.gov/foreign-language-training/

U.S. Department of State. (2023b, May 3). *About Foreign Service Assignments* Careers Representing America. https://careers.state.gov/career-paths/foreign-service/officer/about-foreign-service-assignments/

U.S. Department of State. (2023c). *About* Our Mission. https://www.state.gov/about/

U.S. Department of State. (2023d, September 29). *The Foreign Service Institute's History - United States Department of State*. U.S. Department of State. https://www.state.gov/the-foreign-service-institutes-history/

U.S. Department of State. (2023f). *13 FAM 201.2 Language designated positions (LDPS).* U.S. Department of State. https://fam.state.gov/FAM/13FAM/13FAM020102.html

U.S. Department of State. (2023g, September 9). *Is the Foreign Service right for you?*. Foreign Service. https://careers.state.gov/career-paths/foreign-service/

U.S. Department of State. (2023h). *3 FAM 3900 Incentive pay (LDPS).* U.S. Department of State. https://fam.state.gov/fam/03fam/03fam3910.html

U.S. Department of State. (2023i). *Which languages qualify for bonus points?*. Foreign Service. https://careers.state.gov/faq-items/language-bonus-points/

U.S. Department of State Foreign Service Institute. (2020). *FSI: Task-Based Language Teaching* [YouTube channel]. YouTube. Retrieved July 8, 2024, from https://www.youtube.com/playlist?list=PLozqW4cd_hzWeaEveUOV8nqJClRKQvExM

Viitso, T. (2007). Structure of the Estonian Language. Phonology, morphology, and word formation. In M. Erelt (Ed), *Estonian Language* (p. 35). Estonian Academy Publishers. https://dspace.ut.ee/server/api/core/bitstreams/adbce35b-935e-4ef8-afd3-e3bf44f79040/content

Wagener, T. R. (2015). *The Influences of Aptitude, Learning Context, and Language Difficulty Categorization on Foreign Language Proficiency* [Doctoral Dissertation, University of Maryland]. Digital Repository at the University of Maryland.

West, S. G., Wu, W., McNeish, D., & Savord, A. (2023). Model fit in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed.) (pp.184-205). The Guilford Press.

Widaman, K. F. (2006). Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development.*

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12*(3), 265–310. https://doi.org/10.1016/0273-2297(92)90011-P

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary educational psychology, 25*(1), 68-81.

Winke, P., Gass, S. and Myford, C. (2011), THE RELATIONSHIP BETWEEN RATERS' PRIOR LANGUAGE STUDY AND THE EVALUATION OF FOREIGN LANGUAGE SPEECH SAMPLES. ETS Research Report Series, 2011: i-67. https://doi.org/10.1002/j.2333-8504.2011.tb02266.x

Winke, P., Zhang, X., Rubio, F., Gass, S., Sonenson, D., & Hacking, J. (2020). The proficiency profiles of language students: Implications for programs. *Second Language Research & Practice, 1*(1), 25–64. http://hdl.handle.net/10125/69840

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and psychological measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wu, J., Liu, M., & Liao, C. (20l3). Analytic scoring in an interpretation test: Construct validation and the halo effect. In H-H. Liao, T-E. Kao & Y. Lin (Eds.), *The making of a translator: Multiple perspectives* (pp.277-292). Taipei: Bookman.

Yamashita, J. (2004). Reading attitudes in L1 and L2, and their influence on L2 extensive reading. *Reading in a foreign language, 16*(1), 1-19.