

ABSTRACT

Title of Dissertation: PATHWAYS TO PROFICIENCY:
EXAMINING THE COHERENCE OF INITIAL
SECOND LANGUAGE ACQUISITION
PATTERNS WITHIN THE LANGUAGE
DIFFICULTY CATEGORIZATION
FRAMEWORK

Megan Christina Masters, Doctor of Philosophy,
2018

Dissertation directed by: Dr. Steven J. Ross, Professor, Second Language
Acquisition

It has perhaps never been clearer that in order to effectively communicate with global governments and develop reasoned foreign policy, the United States Intelligence Community requires the support of trained linguists. The development of foreign language proficiency is a complex process requiring a significant investment of time and resources. For learners involved in intensive foreign language training within the United States Government (USG), the Department of Defense (DoD) has developed various Language Difficulty Categorization (LDC) frameworks aimed at standardizing the amount of time learners are given to meet established proficiency criteria. Despite the widespread adoption of LDC frameworks over the past 60 years, few empirical studies have examined the systematicity in proficiency patterns for languages grouped within the

same difficulty category. By situating the analysis within the framework of a logic model, data-mining techniques were used to statistically model, via path analysis, the relationships between program inputs, activities, and outcomes.

Two main studies comprised the investigation. Study 1 employed a contrastive-analytic approach to examine the coherence with which both cognitive (e.g., general aptitude, language-specific aptitude, and average coursework outcomes) and non-cognitive (e.g., language preference self-assessment scores) variables contributed to the development of foreign language achievement and proficiency outcomes for three languages grouped within the same category. For Study 1, only learners who completed the entire foreign language-training program were included in the analysis. Results of Study 1 found a great deal of coherence in the role that language-specific aptitude and 300-level average coursework grades play in predicting end-of-program proficiency outcomes. To examine the potential hidden effects of non-random attrition, Study 2 followed the same methodological procedures as Study 1, but it imputed missing coursework and proficiency test score data for learners who attrited (that is, “dropped out”) during the intensive foreign language-training program. Results of the imputation procedure confirmed that language-specific aptitude plays a robust role in predicting average coursework outcomes across languages. Study 2 also revealed substantial differences in the role that cognitive and non-cognitive variables play in predicting end-of-program proficiency outcomes between the observed and imputed datasets as well as across languages and skills.

PATHWAYS TO PROFICIENCY: EXAMINING THE COHERENCE OF INITIAL
SECOND LANGUAGE ACQUISITION PATTERNS WITHIN THE LANGUAGE
DIFFICULTY CATEGORIZATION FRAMEWORK

by

Megan Christina Masters

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
2018

Advisory Committee:
Professor Steven J. Ross, Chair
Professor Michael Long
Professor Robert DeKeyser
Dr. Jared Linck
Professor Donald J. Bolger, Dean's Representative

© Copyright by
Megan Christina Masters
2018

Dedication

To my husband, Dave, I could not be more grateful for your love, support, encouragement, and commitment to seeing me through accomplishing this goal. It meant the world to me to have you by my side and I feel so blessed to spend my life with you. To my son, Ryan James, and daughter, Cara Marion, may this manuscript be a reminder to always finish what you started, and that nothing worth earning comes easily.

The honor of being the first person in our family to earn a Doctor of Philosophy is not lost on me. To the women in my family whose intelligence and perseverance set the standard against which I could benchmark success, namely, the late Marion Kanachis, and Jeanette Krol, I also dedicate this to you. To the late Dr. James Michael, thank you for encouraging me to apply, and for imparting in me the importance of the constant pursuit of knowledge. To my parents, Mary and Edward Krol, thank you for instilling, from a young age, the importance of challenging yourself and working hard as well as for always asking, “how’s that degree coming along?” Lastly, to my dear sister Kelly Bezak and brother-in-law, Shaun Bezak, I could not have accomplished this goal without the both of you. Thank you for being by my side, every step of the way, and for teaching me, by example, the way of the ant.

Acknowledgments

First and foremost, I would like to offer a heartfelt acknowledgement to my advisor, Dr. Steven J. Ross, who has the unparalleled ability of making the most complex statistical procedures approachable and understandable. I am so grateful to have been the recipient of your training. It has been an honor to work with and learn from you. To my committee members, Drs. Long and DeKeyser, it has truly been a pleasure to be your student. The insights you have offered into the field of SLA have shaped the way I think as an academic and approach argumentation. To my committee member and colleague, Dr. Jared Linck, your ability to make significant contributions to cross-disciplinary research is truly remarkable. I hope to apply a similar approach to my own research. To Dr. Catherine J. Doughty, I have learned so much from you during our years of working together. Your commitment to the field and to maintaining standards of excellence in your research is inspiring.

To my dear friends and colleagues who have been there to lend an ear, offer important input, provide critical perspectives, share a bottle of wine, or help take care of the kids while I squeezed in a few hours of writing, I could not have earned this degree without you - namely, Dr. Marcio Oliveira, Dr. Martyn Clark, Dr. Amber Bloomfield, Dr. Susan Benson, Dr. Katharine Nielson, Dr. Ana Palla-Kane, Dr. Emma Gregory, Eric Pelzl, Alia Lancaster, Stephen O'Connell, Cari Killian, Liz Crawford, Kenny Jarosz, Jess Larche, Caitlin Cahalan, Kelly Arnold, Allison Volinsky, Rebecca Titus, and Erin Hoffman. You've seen me through thick and thin, and I don't know how I became so lucky to have people like you in my life. From the bottom of my heart, *thank you*.

Table of Contents

Dedication	ii
Acknowledgments	iii
Table of Contents	iv
List of Tables	vi
List of Figures	viii
List of Abbreviations	xii
Chapter 1: Early Approaches to Investigating the L2 Acquisition Process and Their Influence on the Development of Language Categorization Frameworks	1
Chapter 2: Logic Modeling as a Tool for Examining L2 Instructional Program Coherence	15
Program Theory and Logic Modeling	18
Applying Program Theory and Logic Modeling Within an L2 Program Evaluation Context	22
Current Dataset	27
Chapter 3: Examining the Coherence of Initial Acquisition Proficiency Development	30
Study 1: Purpose of the Study	30
Research Questions	31
Dataset	31
Description of Variables in Model	32
Research Design: Panel Study	51
Path Analysis	53
Study 1 predicted results: Wave 1 to Wave 2	63
Study 1 predicted results: Wave 2 to Wave 3	64
Study 1 predicted results: Wave 1 to Wave 3	65
Study 1 Results	67
Study 1 Results: Wave 1 to Wave 2 Path Analyses: Reading, Listening, and Speaking	67
Study 1 Results: Wave 2 to Wave 3 Path Analyses: Reading	71
Study 1 Results: Wave 2 to Wave 3 Path Analyses: Listening	75
Study 1 Results: Wave 2 to Wave 3 Path Analyses: Speaking	78
Study 1 Results: Wave 1 to Wave 3 Full Path Analyses: Reading	80
Study 1 Results: Wave 1 to Wave 3 Full Path Analyses: Listening	85
Study 1 Results: Wave 1 to Wave 3 Full Path Analyses: Speaking	87
Discussion: Study 1	89
Study 2: Validity of the Imputation Procedure for Path-Analytic Analyses Within a Large-Scale, L2 Instructional Context	94
Missing Data: Overview	94
Results: Study 2	106
Study 2 Results: Wave 1 to Wave 2 Path Analyses: Reading, Listening, and Speaking	106
Study 2 Results: Wave 2 to Wave 3 Path Analyses: Reading	112
Study 2 Results: Wave 2 to Wave 3 Path Analyses: Listening	116
Study 2 Results: Wave 2 to Wave 3 Path Analyses: Speaking	120

Study 2 Results: Wave 1 to Wave 3 Full Path Analyses: Reading.....	123
Study 2 Results: Wave 1 to Wave 3 Full Path Analyses: Listening.....	128
Study 2 Results: Wave 1 to Wave 3 Full Path Analyses: Speaking.....	131
Discussion: Study 2	134
Conclusions and Directions for Future Research	140
Limitations	147
Appendices	150
Appendix A: Overview of the Contrastive Analysis Hypothesis.....	150
Appendix B: List of DLAB 2 Predictor and Outcome Variables.....	157
Appendix C: Correlation Matrices for Average Coursework Outcomes.....	159
Appendix D: Correlation Matrices for Path-Analytic Model.....	162
Appendix E: Descriptive Statistics for Imputed Data.....	165
Appendix F: Wave 2 and Wave 3 Observed and Imputed Mean-Level Comparisons.....	169
Appendix G: Minimum, Maximum, and Median RMSEA Values of Imputed Models.....	170
Appendix F: Sample DLPT Listening and Reading Specimen.....	171
References.....	173

List of Tables

Table 1. Current LDC framework at the DLIFLC	6
Table 2. Outlier languages by difficulty category (taken from Clark et al. (2016 b; c)...	10
Table 3. Difficulty categories and associated length of instruction at DLIFLC (adapted from Mackey, 2014).....	23
Table 4. Descriptive Statistics (Arabic)	33
Table 5. Descriptive Statistics (Chinese)	33
Table 6. Descriptive Statistics (Korean)	33
Table 7. Descriptive Statistics (AFQT Transformed Values-Correcting for Negative Skew)	35
Table 8. Descriptive Statistics (DLAB Transformed Values).....	36
Table 9. Descriptive Statistics (Language Preference Self-Assessment Transformed Values).....	37
Table 10. Descriptive Statistics (Arabic Coursework).....	40
Table 11. Descriptive Statistics (Chinese Coursework).....	40
Table 12. Descriptive Statistics (Korean Coursework).....	40
Table 13. Descriptive Statistics: Transformed Values (Arabic Coursework).....	42
Table 14. Descriptive Statistics: Transformed Values (Chinese Coursework).....	42
Table 15. Descriptive Statistics: Transformed Values (Korean Coursework).....	42
Table 16. Descriptive Statistics (Arabic End-of-Program Outcomes; Scale = 0 - 30)	47
Table 17. Descriptive Statistics (Chinese End-of-Program Outcomes; Scale = 0 - 30) ..	47
Table 18. Descriptive Statistics (Korean End-of-Program Outcomes; Scale = 0 - 30)....	47
Table 19. Percentage of learners to meet established DLPT and OPI criterion scores....	50
Table 20. Structural equations for each endogenous variable in model.....	63
Table 21. Arabic Wave 1 to Wave 2 path-analytic outcomes (n =241).....	69
Table 22. Chinese Wave 1 to Wave 2 path-analytic outcomes (n = 98).....	69
Table 23. Korean Wave 1 to Wave 2 path-analytic outcomes (n = 75).....	69
Table 24. z-transformation of DLAB to 100-, 200-, and 300-level course outcomes	70
Table 25. Arabic Reading: Wave 2 to Wave 3 path-analytic outcomes (n = 241).....	72
Table 26. Chinese Reading: Wave 2 to Wave 3 path-analytic outcomes (n = 98).....	73
Table 27. Korean Reading: Wave 2 to Wave 3 path-analytic outcomes (n = 75).....	73
Table 28. Multi-Group Invariance Testing: Wave 2 to Wave 3: Reading	74
Table 29. Arabic Listening: Wave 2 to Wave 3 path-analytic outcomes (n = 241).....	76
Table 30. Chinese Listening: Wave 2 to Wave 3 path-analytic outcomes (n = 98).....	76
Table 31. Korean Listening: Wave 2 to Wave 3 path-analytic outcomes (n = 75).....	76
Table 32. Multi-Group Invariance Testing: Wave 2 to Wave 3: Reading	77
Table 33. Arabic Speaking: Wave 2 to Wave 3 path-analytic outcomes (n = 241).....	79
Table 34. Chinese Speaking: Wave 2 to Wave 3 path-analytic outcomes (n = 98).....	79
Table 35. Chinese Reading: Wave 1 to Wave 3 path-analytic outcomes (n = 98).....	82
Table 36. Korean Reading: Wave 1 to Wave 3 path-analytic outcomes (n = 75).....	82
Table 37. Multi-Group Invariance Testing: Wave 1 to Wave 3 Reading	83
Table 38. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean reading models	84
Table 39. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean listening models	86

Table 40. Korean Speaking: Wave 1 to Wave 3 path-analytic outcomes (n = 75).....	88
Table 41. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean speaking models.....	88
Table 42. Arabic Wave 1 to Wave 2 path-analytic outcomes (Observed (n= 241) and Imputed (n = 411)).....	110
Table 43. Chinese Wave 1 to Wave 2 path analytic outcomes (Observed (n = 98) and Imputed (n = 161)).....	111
Table 44. Korean Wave 1 to Wave 2 path analytic outcomes (Observed (n = 75) and Imputed (n = 118)).....	111
Table 45. Differences in Observed versus imputed correlations	112
Table 46. Arabic Reading: Wave 2 to Wave 3 path-analytic outcomes (Observed (n = 241) and Imputed (n = 411)).....	114
Table 47. Chinese Reading: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161)).....	115
Table 48. Korean Reading: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118)).....	115
Table 49. Differences in observed versus imputed correlations	116
Table 50. Arabic Listening: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 241) and Imputed (n = 411))	118
Table 51. Chinese Listening: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161)).....	118
Table 52. Korean (Observed n = 75; Imputed n = 118).....	118
Table 53. Differences in observed versus imputed correlations	119
Table 54. Arabic Speaking: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 241) and Imputed (n = 411)).....	122
Table 55. Chinese Speaking: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161)).....	122
Table 56. Differences in observed versus imputed correlations	123
Table 57. Chinese Reading: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161)).....	125
Table 58. Korean Reading: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118)).....	125
Table 59. Differences in observed versus imputed correlations	126
Table 60. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean reading models	127
Table 61. Arabic Listening: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 241) and Imputed (n = 411))	129
Table 62. Korean Listening: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118)).....	129
Table 63. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean listening models	130
Table 64. Korean Speaking: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118)).....	132
Table 65. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean speaking models.....	133

List of Figures

Figure 1. Distributions of empirical language difficulty by category (taken from Clark et al. (2016 b).....	11
Figure 2. Main components of a logic model (adapted from Fretchling, 2007, p. 22)....	19
Figure 3. The four main components of a logic model applied within a FL program evaluation context.	26
Figure 4. Arabic AFQT Weighted	34
Figure 5. Arabic DLAB	34
Figure 6. Arabic Lang.Pref.	34
Figure 7. Chinese AFQT Weighted	34
Figure 8. Chinese DLAB	34
Figure 9. Chinese Lang.Pref.	34
Figure 10. Korean AFQT Weighted	34
Figure 11. Korean DLAB.....	34
Figure 12. Korean Lang.Pref.....	34
Figure 13. Arabic AFQT (T).....	35
Figure 14. Chinese AFQT (T).....	35
Figure 15. Korean AFQT (T).....	35
Figure 16. Arabic DLAB (T)	37
Figure 17. Chinese DLAB (T)	37
Figure 18. Korean DLAB (T)	37
Figure 19. Arabic Lang.Pref. (T)	38
Figure 20. Chinese Lang.Pref. (T)	38
Figure 21. Korean Lang.Pref.(T)	38
Figure 22. Non-Transformed Arabic, Chinese, and Korean AFQT, DLAB, and Language Preference Self-Assessment scores.....	39
Figure 23. Arabic 100-level	41
Figure 24. Arabic 200-level	41
Figure 25. Arabic 300-level	41
Figure 26. Chinese 100-level	41
Figure 27. Chinese 200-level	41
Figure 28. Chinese 300-level	41
Figure 29. Korean 100-level	41
Figure 30. Korean 200-level	41
Figure 31. Korean 300-level	41
Figure 32. Arabic 100-level (T)	44
Figure 33. Chinese 100-level (T).....	44
Figure 34. Korean 100-level (T)	44
Figure 35. Arabic 200-level (T)	44
Figure 36. Chinese 200-level (T).....	44
Figure 37. Korean 200-level (T)	44
Figure 38. Arabic 300-level (T).....	44
Figure 39. Chinese 300-level (T).....	44
Figure 40. Korean 230-level (T)	44

Figure 41. Non-Transformed Arabic, Chinese, and Korean 100-, 200-, and 300- average coursework outcomes.....	46
Figure 42. Arabic DLPT Listening	48
Figure 43. Arabic DLPT Reading	48
Figure 44. Arabic OPI Speaking.....	48
Figure 45. Chinese DLPT Listening	48
Figure 46. Chinese DLPT Reading	48
Figure 47. Chinese OPI Speaking.....	48
Figure 48. Korean DLPT Listening	48
Figure 49. Korean DLPT Reading.....	48
Figure 50. Korean OPI Speaking.....	48
Figure 51 Transformed Arabic, Chinese, and Korean DLPT and OPI outcomes.....	50
Figure 52. Visual depiction of the variables associated with each wave of the panel study	52
Figure 53. Example Wave 1 to Wave 2 Partial Path Model	55
Figure 54. Example Wave 2 to Wave 3 Partial Path Model	57
Figure 55. Complete path-analytic diagram.....	58
Figure 56. Wave 1 to Wave 2 Model Arabic	68
Figure 57. Wave 1 to Wave 2 Model Chinese	68
Figure 58. Wave 1 to Wave 2 Model Korean	68
Figure 59. Wave 2 to 3 Arabic Reading.....	72
Figure 60. Wave 2 to 3 Chinese Reading.....	72
Figure 61. Wave 2 to 3 Korean Reading.....	72
Figure 62. Wave 2 to 3 Arabic Listening.....	76
Figure 63. Wave 2 to 3 Chinese Listening.....	76
Figure 64. Wave 2 to 3 Korean Listening.....	76
Figure 65. Wave 2 to 3 Arabic Speaking.....	78
Figure 66. Wave 2 to 3 Chinese Speaking.....	78
Figure 67. Wave 2 to 3 Korean Speaking.....	78
Figure 68. Full Arabic Path Model Reading.....	81
Figure 69. Full Chinese Path Model Reading.....	81
Figure 70. Full Korean Path Model Reading.....	81
Figure 71. Full Arabic Path Model Listening.....	85
Figure 72. Full Chinese Path Model Listening.....	85
Figure 73. Full Korean Path Model Listening.....	85
Figure 74. Full Arabic Path Model Speaking.....	87
Figure 75. Full Chinese Path Model Speaking.....	87
Figure 76. Full Korean Path Model Speaking.....	87
Figure 77. Arabic Pattern of Missing Data (Coursework).....	96
Figure 78. Arabic Missing Data: Course averages	96
Figure 79. Chinese Pattern of Missing Data (Coursework).....	97
Figure 80. Chinese Missing Data: Course averages	97
Figure 81. Korean Pattern of Missing Data (Coursework).....	98
Figure 82. Korean Missing Data: Course averages.....	98
Figure 83. Arabic Wave 3 Missingness	99
Figure 84. Chinese Wave 3 Missingness	99

Figure 85. Korean Wave 3 Missingness	99
Figure 86. Patterns of Missingness across language groups	100
Figure 87. Visual depiction of the Multiple Imputation Process	104
Figure 88. Wave 1 to Wave 2 Model Arabic (Observed)	109
Figure 89. Wave 1 to Wave 2 Model Arabic (Imputed)	109
Figure 90. Wave 1 to Wave 2 Model Chinese (Observed)	109
Figure 91. Wave 1 to Wave 2 Model Chinese (Imputed)	109
Figure 92. Wave 1 to Wave 2 Model Korean (Observed)	109
Figure 93. Wave 1 to Wave 2 Model Korean (Imputed)	109
Figure 94. Wave 2 to 3 Arabic Reading (O)	113
Figure 95. Wave 2 to 3 Arabic Reading (I)	113
Figure 96. Wave 2 to 3 Chinese Reading (O)	113
Figure 97. Wave 2 to 3 Chinese Reading (I)	113
Figure 98. Wave 2 to 3 Korean Reading (O)	114
Figure 99. Wave 2 to 3 Korean Reading (I)	114
Figure 100. Wave 2 to 3 Arabic Listening (O)	117
Figure 101. Wave 2 to 3 Arabic Listening (I)	117
Figure 102. Wave 2 to 3 Chinese Listening (O)	117
Figure 103. Wave 2 to 3 Chinese Listening (I)	117
Figure 104. Wave 2 to 3 Korean Listening (O)	117
Figure 105. Wave 2 to 3 Korean Listening (I)	117
Figure 106. Wave 2 to 3 Arabic Speaking (O)	121
Figure 107. Wave 2 to 3 Arabic Speaking (I)	121
Figure 108. Wave 2 to 3 Chinese Speaking (O)	121
Figure 109. Wave 2 to 3 Chinese Speaking (I)	121
Figure 110. Wave 2 to 3 Korean Speaking (O)	121
Figure 111. Wave 2 to 3 Korean Speaking (I)	121
Figure 112. Full Arabic Path Model Reading (O) (n = 241, CFI = 0.999, RMSEA = 0.034)	124
Figure 113. Full Arabic Path Model Reading (I) (n = 411, CFI = 1.00, RMSEA = 0.000 (all models))	124
Figure 114. Full Chinese Path Model Reading (O) (n = 98, CFI = 1.000, RMSEA = 0.000)	124
Figure 115. Full Chinese Path Model Reading (I) (n = 161, CFI = 1.000, RMSEA = 0.000)	124
Figure 116. Full Korean Path Model Reading (O)	124
Figure 117. Full Korean Path Model Reading (I) (n = 118, CFI = 1.000, RMSEA = 0.000)	124
Figure 118. Full Arabic Path Model Listening (O)	128
Figure 119. Full Arabic Path Model Listening (I)	128
Figure 120. Full Chinese Path Model Listening (O)	128
Figure 121. Full Chinese Path Model Listening (I)	128
Figure 122. Full Korean Path Model Listening (O)	129
Figure 123. Full Korean Path Model Listening (I)	129
Figure 124. Full Arabic Path Model Speaking (O)	131
Figure 125. Full Arabic Path Model Speaking (I)	131

Figure 126. Full Chinese Path Model Speaking (O) (n = 98, CFI = 1.000, RMSEA = 0.000)	131
Figure 127. Full Chinese Path Model Speaking (I)	131
Figure 128. Full Korean Path Model Speaking (O)	132
Figure 129. Full Korean Path Model Speaking (I).....	132

List of Abbreviations

AFQT	Armed Forces Qualifying Test
ASVAB	Armed Services Vocational Aptitude Battery
CAH	Contrastive Analysis Hypothesis
Df	Degrees of Freedom
DLAB	Defense Language Aptitude Battery
DLPT	Defense Language Proficiency Test
DLIFLC	Defense Language Institute Foreign Language Center
DoD	Department of Defense
FL	Foreign Language
ILR	Interagency Language Roundtable
L1	First Language
L2	Second Language
LDC	Language Difficulty Categorization
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
OPI	Oral Proficiency Interview
SLA	Second Language Acquisition
USG	United States Government

Chapter 1: Early Approaches to Investigating the L2 Acquisition Process and Their Influence on the Development of Language Categorization Frameworks

Vital to the success of international relationships and the development of reasoned foreign policy is the need for adequately trained linguists within the United States Intelligence Community. Under Department of Defense (DoD) directive, the Department of the Army is responsible for providing foreign language training to four military services (the United States Army, the United States Navy, the United States Air Force, and the United States Marine Corps). The Defense Language Institute Foreign Language Center (DLIFLC), located in Monterey, California, is home to a residential foreign language training and testing program designed to train military linguists to proficiency levels approximating those of traditional four-year undergraduate foreign language majors (J. Lett, personal communication, May 2007). Managing and coordinating the training of military linguists for over 20 languages is a complex process requiring unparalleled investments of time and taxpayer resources. Recently costs for training a single military linguist have been estimated at upwards of \$250,000 (Brecht, personal communication, March 2018). Therefore, the United States Government (USG) has developed various Language Difficulty Categorization (LDC) frameworks in order to predict the length of initial second language (L2) acquisition instructional time and to plan for resource allocation.¹ The main goal of these frameworks is to establish a comprehensive, policy-based system through which large numbers of DoD language analysts can matriculate as efficiently as possible (Lett, 1990; Clark, Jackson, Kim, O'Rourke, Aghajanian-Stewart, & Ross, 2016a). In addition to being used to group

¹ The frameworks are designed from the perspective of adult monolinguals with English as a first language.

² Minimum DoD proficiency test score criterion outcomes will be discussed in detail in Chapter 2.

languages for initial acquisition training, they are also used for a wide variety of purposes once learners meet minimum proficiency criteria, including instructor-to-student classroom training ratios, academic credit earnings, and incentive pay (see Clark et al., 2016a, p. 15).

Despite the widespread adoption of LDC frameworks across all USG agencies, virtually few research studies have examined empirically the coherence of initial acquisition patterns of languages grouped within the same category. The purpose of the current chapter is to review findings from the most recent empirical research examining the coherence of L2 categorization frameworks working with learners who have already met minimum DoD proficiency criteria (Clark et al., 2016a,b,c).² Foreign language groupings within the various frameworks tend to be based largely upon the subjective expertise of L2 instructors or upon typological differences between a given L2 and English. The categorization system assumes that languages within the same category require the same amount of instructional time to meet minimum proficiency standards required to perform a variety of mission-related L2 tasks. Systematic examination of the of the LDC framework can either provide corroborating evidence with respect to how well languages grouped within the system appear to be functioning, or reveal differential patterns in proficiency development for languages otherwise assumed to function similarly. If the categorization system is wrong, it could be the case that military personnel graduate from the DLIFLC without the language skills necessary to perform their assignments. The chapter will conclude with the argument that replication analyses examining patterns of proficiency development using data from learners who are just

² Minimum DoD proficiency test score criterion outcomes will be discussed in detail in Chapter 2.

beginning their L2 language studies, rather than post L2 acquisition data, is warranted. As the stakes for training, and subsequently deploying, military linguists are quite high, the need for robust empirical evidence validating the categorization and use of the LDC framework is critical.

Over the past 60 years, L2 researchers have employed various data-analytic methodologies to investigate the complex cognitive processes associated with the acquisition of a second language for both children and adults.³ As is the case with all emerging fields, available analytical techniques determined how researchers initially examined Second Language Acquisition (SLA) processes, which, in turn, informed early theories of how learners could most efficiently acquire a second language. The notion of developing a systematic framework through which large numbers of military linguists could be efficiently trained received attention throughout the 1950s through 1970s as the need for government linguists with advanced levels of foreign language proficiency vastly increased.⁴ During this time, government policy makers grew keenly interested in developing training programs that efficiently produced linguists who could serve as expert transcriptionists, translators, and/or interpreters. Perhaps borrowing from Lado's (1957) and Stockwell et al.'s (1965) early research (which applied the well-known Contrastive Analysis Hypothesis (CAH) to the language-learning context), a number of USG entities extended the conceptualization of grouping language features both *typologically*, in terms of English L1 and L2 linguistic and structural differences, and

³ For a detailed overview of the early research pertaining to the evolution of empirical issues in early SLA research, see Hakuta and Cancino, 1977 and Larsen-Freeman and Long, 1991.

⁴ The unexpected launch of Sputnik in 1957 by the former Union of Soviet Socialist Republics (USSR) prompted an increased USG investment in science, engineering, and foreign language studies, with the ultimate goal of ensuring a robust, well-trained army of scientists, engineers, and linguists that could readily anticipate and/or respond to international crises as well as advance the nation's scientific and academic contributions in these main areas.

hierarchically, in terms of anticipated difficulty, from isolated linguistic features to entire languages themselves.⁵ As described in detail by Clark et al., (2016a), this approach resulted in a number of different language difficulty hierarchies, referred to across the DoD as LDC frameworks. In the first academic research aimed at examining the LDC system, the authors note:

While there are many sources that cite the LDC, only a few detail the initial, conceptual development of the framework. Furthermore, the LDC is not a uniform system within the United States Government (USG). For example, the difficulty category system at the Foreign Service Institute (FSI) is similar, but distinct from, the framework used at the Defense Language Institute Foreign Language Center (DLIFLC) and in the Department of Defense (DOD) at large. In addition, there have been some fluctuations over time in the number of categories as well as the category assignment of languages (p. 2).

The authors also point out that USG-based approaches to grouping languages within a hierarchy of difficulty have typically been informed by subjective observations made by expert instructors or training specialists but have not examined the extent to which empirical evidence supports or refutes hypothesized groupings.

Given the scant amount of published research detailing the grouping of languages with the LDC framework, Clark et al. conducted semi-structured interviews with subject matter experts. The interviewees relayed their recollections about how the initial frameworks were designed. After aggregating the results collected across each interview, the authors concluded that the initial LDC frameworks were first designed by the Department of State's Foreign Service Institute and later adopted by the DLIFLC in the

⁵ See Appendix A for a description of how this early work likely informed the hierarchical nature of the current categorization system.

1950s.⁶ These frameworks were informed by post hoc, expert observations from USG course instructors about how long it had been taking students to meet established graduation criteria in their target language.⁷ The authors note that frameworks were designed using a bottom-up rather than top-down approach. That is, rather than conjecturing where a given language would fall within a hierarchy a priori, the creators of the framework worked from their observations of student learning and placed the language into a respective category within the hierarchy a posteriori. Jackson and Kaplan (2001) argue that this approach is thus based on empirical, albeit merely observational, evidence.

Clark et al. (2016 a) note that the first USG LDC frameworks comprised three categories, arranged from hypothesized least (Category I) to most (Category III) difficult, similar to the manner in which Stockwell et al. (1965) organized their framework based on typological language features. They state:

Category I consisted of languages that are perceived to be easier for native English speakers to learn (French, Spanish, Portuguese, German, etc.), Category II comprised languages that were somewhat more difficult (Russian, Greek, Turkish, Thai, etc.), and Category III consisted of the languages believed to be the hardest (Arabic, Chinese, Japanese, etc.) (p. 2).

The authors note that in the late 1960s, additional categories were added to make room for the incorporation of more languages, such as Indonesian, Malay, and Romanian. They state, “Category II as it is used today was created to accommodate [languages] that were similar in terms of time and demand, and the previous Categories II and III became III

⁶ Detailed information about the mission, purpose, selection, and training process associated with DLIFLC and its students will be provided in Chapter 2. To this day, the DLIFLC continues to serve as the primary USG institution at which service women and men are trained as military linguists.

⁷ What the subject matter experts define as “learned” is not described.

and IV at DLIFLC, respectively. Table 1 below, taken from Clark et al. (2016 a; b; c), denotes the current categorization system as of September 2016.

Table 1. Current LDC framework at the DLIFLC* 8

Category I (26 weeks)	Category II (35 weeks)	Category III (48 weeks)	Category IV (64 weeks)
French Portuguese Spanish	German Indonesian	Hebrew Hindi Persian Farsi Russian Serbian/Croatian Tagalog Turkish Urdu	Modern Standard Arabic Arabic-Egyptian Arabic-Iraqi Arabic-Levantine Arabic-Sudanese Chinese Mandarin Japanese Korean Pashto

* According to the fiscal year 2016 catalog; the DLIFLC website lists Category I and Category II languages at 36 weeks.

Upon review of Table 1, of particular note is the way in which the LDC appears to be based largely on the same principles as Lado’s CAH, which also took shape during the 1950s and 1960s.⁹ Similar to the CAH, the LDC framework not only informed instructors about which target languages an L1 English learner would likely have trouble acquiring, but also provided instructors with guidelines associated with the amount of time learners would need to acquire a given language. Consistent with the CAH, the higher the difficulty category of a given language, the greater the predicted interference between the two languages and the longer the amount of time that instructors and students would need to teach and learn the language. It is important to note that the number of weeks of instruction associated with the categorization system does not necessarily correspond to

⁸ Clark et al. (2016) also note, “While the composition of the categories in the DLIFLC’s system has been quite stable over the years, two languages have changed categories. In 2009, Pashto was moved from Category III to Category IV as graduation rates failed to reach the levels of other Category III languages. Similarly, Hindi was moved from Category II to Category III at some point in the 1980s.”

⁹ See Appendix A for a detailed overview of inaugural theoretical perspectives within the SLA field related to L2 learning.

the amount of instructional time recommended by DoD training and testing working groups. As Clark et al. (2016 a) state,

After interviewing language teachers and reviewing the outcomes of various government and academic programs, the [working group] committee determined that roughly 104 weeks of training were needed for the hardest languages, roughly 75 weeks of training for the medium languages, and roughly 50 weeks were needed for the easiest languages. Courses of that length were determined unrealistic, however, due to the time and cost (p. 7).

That is, military linguists are expected to meet minimum proficiency requirements (and, if successful, to perform a range of L2-related job tasks) after engaging in instruction that is equivalent to roughly *half* of that recommended by working group members. The incongruity between recommended instructional time and that allowed by DoD policy analysts makes even more obvious the need to examine empirically the LDC framework.

In a white paper written by researchers at the Human Resources Research Organization, Koch and McCloy (2015), after synthesizing available USG policy and research documents, concluded that the order in which languages were assigned to categories within the LDC were based on two main criteria: (1) difficulty and (2) distance from English. Clark et al. (2016 a; b; c) define difficulty as the use of “different indices that reflect the relationship between time spent learning and [an established] proficiency [level].” Distance refers to “the degree of difference in vocabularies, phonetic inventories, grammars, etc. between...two languages” (p. 3). In their review, the authors found that difficulty was typically defined by two criteria: (1) the number of hours it takes a learner to reach Interagency Language Roundtable 3 (ILR-3) or (2) the level of speaking proficiency that students reach after 24 weeks of instruction (p. 3).¹⁰ The

¹⁰ The Interagency Language Roundtable (ILR) is an unfunded federal interagency organization that sets common standards about language-related activities at the federal level. It was originally founded in 1955

researchers calculated a Spearman's rank order correlation coefficient between data from research previously conducted by Cysouw (2013), which reported the results of the observed relationship between hours required to reach ILR-3, and data from Hart-Gonzalez and Lindemann (1993), which investigated the level of speaking proficiency attained after 24 weeks of instruction. The authors note that the results revealed a "strong relationship between DLIFLC's categories and the two indices of difficulty: time to ILR-3 ($r = 0.95, p < 0.001$) and proficiency level after 24 weeks [of instruction] ($r = -0.88, p < 0.001$), providing a modicum of evidence for DLIFLC's current categorization system.

Turning to the examination of how distance from English was measured, Clark et al.'s (2016 a) research did not uncover a robust empirical foundation for grouping languages by linguistic features, stating "there is no evidence that it was based on a rigorous analysis of linguistic features" (p. 3). Although it can be noted that all of the languages grouped in Categories III and IV differ substantially from English in that they have different scripts than English, as noted by Lowe (1998), this does not necessarily indicate homogeneity in language features within a given category grouping. In other words, the fact that languages have been grouped within the same category does not indicate that they contain typological features that function in a similar manner.¹¹ To explore empirically the homogeneity of languages grouped within the same difficulty categories, Clark et al. (2016b,c) systematically mined more than 234,000 unique test records representing 108 different languages for nearly 20,000 language analysts within a

by members of the Air Force, Foreign Service Institute, and Central Intelligence Agency. The ILR sets expected proficiency standards for novice- through advanced-level learners, ranging from 0 (no functional proficiency) to 5 (functional proficiency equivalent to that of a highly articulate, well-educated user of the language). For more information, see <http://www.govtilr.org/IRL%20History.htm> and <http://www.govtilr.org/skills/ILRscale1.htm>

¹¹ The closest approximation to a systematic categorization of language difficulty based on typological language features was research completed by Child (1998).

large government organization who had already reached L2 proficiency criteria of ILR scores of 2/2/1+ in the listening, reading, and speaking skills, respectively. The goal of their research was to investigate statistically the extent to which similar patterns in Defense Language Proficiency Test (DLPT) outcomes would be observed for languages grouped within the same category.¹² The authors state, “if the language difficulty categories are a primary driver in explaining the success with which individuals can master a foreign language, we would expect to see a great deal of homogeneity [in observed DLPT outcome patterns] among languages in the same difficulty category” (p. 5). In other words, if the contrastive analytic approach of grouping languages based on difficulty or distance from English is valid, it would follow that languages grouped within the same category would display largely invariant patterns. The authors subjected the data to three separate analyses in order to explore observed statistical patterns of languages grouped within the same category.

In their first analysis, the authors completed an event history analysis in order to predict the likelihood of a language analyst reaching an ILR-3 proficiency level, while controlling for covariates such as hours of training and test version change.¹³ In their comparison of languages within Category IV with fewer than 100 records (used as the reference category) against nine other languages grouped within the same category, results indicated significant variation in the amount of time it took analysts to reach the ILR-3 proficiency level. Although the attainment patterns show an increased likelihood

¹² The DLPT is the foreign language proficiency test of record for the majority of DoD language analysts. The DLPT is used to test reading and listening proficiency. The Oral Proficiency Interview (OPI) is used to test speaking proficiency skills.

¹³ Previous research completed by Bloomfield et al. (2012) found that analysts who took a different version of the DLPT (e.g., moved from older to new versions of the proficiency test) are more likely to show a decrease in subsequent test scores.

for all nine languages to reach an ILR-3 level eventually, this result suggests that it takes less time for some languages to reach this benchmark than others.¹⁴

To explore further potential within-category variation, the authors next subjected the data to logistic mixed effects modeling, with the goal of predicting the likelihood of an analyst scoring an ILR-3 on his or her subsequent test, controlling for previous DLPT scores, the amount of time between tests, the amount of training received between tests, and whether or not the examinee took a different version of the test. Working from this model, the authors completed two separate analyses. The first analysis, within this study, focused on determining how difficult it was for an analyst testing within a Category IV, for example, to reach an ILR-3 on a subsequent test, in comparison to a subset of “Small N” languages grouped within the same category.¹⁵ The same analysis was completed separately for the reading and listening skills for Categories I, III, and IV.¹⁶ Table 2 below, taken from Clark et al. (2016 b, p. 7), details the results.

Table 2. Outlier languages by difficulty category (taken from Clark et al. (2016 b; c)

	Category I		Category III		Category IV	
	<i>Reading</i>	<i>Listening</i>	<i>Reading</i>	<i>Listening</i>	<i>Reading</i>	<i>Listening</i>
<i>Harder</i>	Euro. Portuguese	--	Tagalog Russian Urdu	Tagalog Russian Urdu	Chinese Mandarin	Chinese Mandarin
<i>Easier</i>		Spanish	Hebrew	Hebrew	Modern Standard Arabic Korean	

As shown above in Table 2, results of this analysis revealed a number of within-category outliers. Portuguese (Category I, Listening), Tagalog, Russian, and Urdu (Category III), and Chinese Mandarin (Category IV) were identified as being more difficult for language

¹⁴ Since Clark et al. (2016 a) were working from de-identified testing and training data, specific information about which languages exhibited which patterns was not provided.

¹⁵ “Small N Languages” were defined as languages with fewer than 100 records within the dataset.

¹⁶ Category II languages were not included in the analysis due to sparsity of languages within the category.

analysts to attain an ILR-3 on subsequent proficiency tests than other languages grouped within the same category. Spanish (Category I, Listening), Hebrew (Category II), and Modern Standard Arabic and Korean (Category IV, Reading) were identified as being less difficult.

The research team then ran a second mixed-effects model, but this time collapsed all languages into a single analysis in order to estimate any observed differences between language difficulty categories. Results of this analysis indicated that, while some languages varied randomly in terms of difficulty, many of the languages exhibited complete overlap in terms of estimated difficulty.¹⁷ Figure 1 below, taken from Clark et al. (2016b), visually depicts these results.

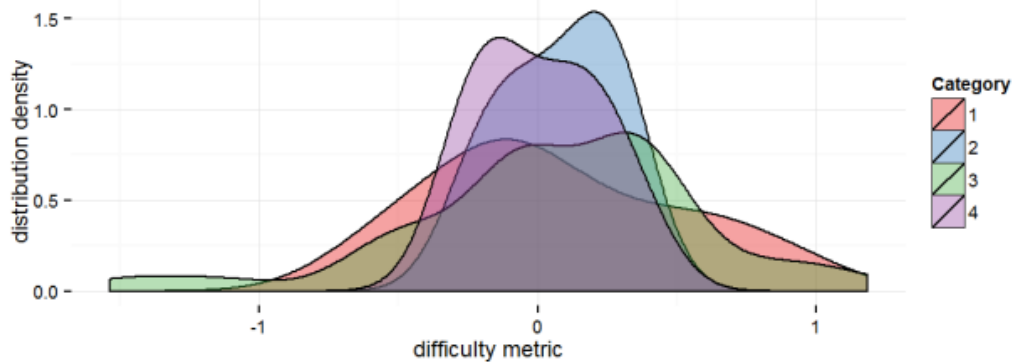


Figure 1. Distributions of empirical language difficulty by category (taken from Clark et al. (2016b)

As shown above in Figure 1, the overlapping distributions observed for each category suggests a lack of “significant differences between categories as a whole” (p. 7). That is, although the LDC framework rests on a variety of assumptions concerning languages’ distance from English and predicted difficulty, the results depicted above show that, for the majority of languages, no meaningful distinctions between languages are observed.

¹⁷ Difficulty was defined by Clark et al. (2016 b) as the likelihood of scoring an ILR-3 on one’s next testing event (p. 6).

The authors acknowledge, however, that learners' aptitude, a variable not modeled in the results discussed above, likely plays an important role in the observed patterns, particularly given the lack of a right-hand "tail" for the more difficult language groupings. Also influencing the patterns displayed in Figure 1 is the restricted range associated with the data. That is, proficiency test scores used as input for the analyses only include test/retest scores for those who have undergone intensive foreign language training and met established DLIFLC proficiency test score criteria. High-aptitude learners who participated in intensive language study but did not meet DLIFLC proficiency test score criteria, do not go on to work in the target language as military linguists and are therefore not included in the above model.

To address this concern, working from a subset of data for military language analysts, which contained aptitude-related data (rather than both military and civilian language analysts), Clark et al. (2016b,c) performed separate regression analyses for languages grouped within the same category. The analyses were designed to predict analysts' final DLPT scores, controlling for number of training hours, the Armed Services Vocational Aptitude Battery (ASVAB) score, the Defense Language Aptitude Battery (DLAB) score, branch of military service, and number of testing attempts.¹⁸ The authors then modeled the standardized residuals (differences between predicted and observed scores) observed for languages grouped within the same category.¹⁹ Results of this analysis indicated that several Category IV languages appeared to be densely clustered below the -2.0 value, suggesting that test scores for these languages were significantly below their predicted scores and thus potentially more difficult to maintain

¹⁸ The DLAB and ASVAB serve as the aptitude-related tests of record for military personnel. Each will be explained in detail in Chapter 2.

¹⁹ When standardized, it is generally expected that residual values will fall between -2.0 and +2.0.

than their counterparts.²⁰ All three analyses conducted by Clark et al. (2016b,c) appear to provide converging evidence concerning heterogeneity in testing patterns observed for languages grouped within the same category. This finding indicates that the categorization system, designed to group languages during the process of initial L2 acquisition, may not adequately discriminate between language groups once minimum proficiency criteria have been attained. In other words, even though it may take learners longer to initially acquire a Category IV language, it is not necessarily the case that it takes more effort to retain their language skills thereafter.

Although the data-analytic techniques within the SLA field have advanced significantly over the past 60 years, many of the category-based frameworks within which second language teaching and training regimens are situated have not. It is hoped that this first chapter has provided sufficient detail concerning the background, historical development, and most recent empirical explorations concerning the development, design, and initial validation research investigating the use of language difficulty categorizations or hierarchies. Research by Clark et al. (2016b,c) explored within-category variation of test/retest data involving L2 learners who had already reached established ILR proficiency levels. However, to the author's knowledge, few published research studies investigate statistically the development of within-category *initial* acquisition patterns. Although languages grouped within the same category are typologically distinct, situating languages within the same category rests on the assumption that each language will display comparable proficiency development patterns.

²⁰ The authors only discuss the results of this analysis for Category IV. The results of these analyses for Categories I and II were not included in the technical report.

An empirical investigation of these frameworks, modeling patterns of initial acquisition using more sophisticated, non-observational data-analytic techniques, is thus warranted.

Chapter 2: Logic Modeling as a Tool for Examining L2 Instructional Program Coherence

The purpose of the current chapter is to introduce the application of logic modeling as a useful tool for examining the observed degree of coherence of initial L2 acquisition patterns within the LDC framework. As noted in Chapter 1, and described in detail in Appendix A, conventional approaches to grouping languages for pedagogical or training purposes have been based on typological differences between languages and impressionistic judgments concerning the L2 acquisition process. These approaches have typically been informed by observations made by expert instructors or training specialists, but there has been no systematic examination of the extent to which empirical evidence supports or refutes hypothesized typological groupings or the time required to meet end-of-program criteria. Situated within an L2 program review framework, this chapter will first review existing literature within the SLA field aimed at examining the effectiveness of large-scale programs of instruction. It will then introduce the concept and application of logic modeling as a novel method of examining hypothesized acquisition patterns as learners progress through large-scale programs of study.

Increased evaluation, assessment, and accreditation demands across all levels of educational institutions have led to the completion of various forms of program reviews, particularly during times of fiscal austerity. Despite the widespread need for universities and program administrators to continually demonstrate the impact of a given program, there is considerable variation in the strategies and methods employed when documenting program processes and outcomes (Ross, 2003; Watanabe et al., 2009). The format of such initiatives differs substantially, ranging from experimental or quasi-experimental research

designs to those focused on gathering more subjective evidence through the use of survey instruments, focus groups, and one-on-one interviews.

Early research investigating learner outcomes within L2 instructional programs was influenced heavily by behavioral psychology; most studies involved large-scale, quasi-experimental designs focusing almost exclusively on the summative evaluation process (Watanabe et al., 2009, p. 8-9). The main purpose of these early research agendas was to determine the superiority of a given pedagogical or instructional approach (Genesee, 1985; Keating, 1963; Scherer & Wertheimer, 1964; Smith, 1970; Swain & Lapkin, 1982). Criticizing the narrow focus on outcomes from summative assessment measures and the impossibility and impracticality of conducting randomized comparative experiments within the foreign language program context, later research within L2 program reviews called for expanding the repertoire of quasi-experimental evaluation methodology to include the incorporation of more diverse types of qualitative program evidence, such as interviews, surveys, and questionnaires (Alderson, 1992; Beretta, 1986, 1992, a, b; Kiley & Rea-Dickens, 2005; Long, 1984; Lynch, 1996; Norris 2006, 2009; Rea-Dickens & Germaine, 1998) (for a more detailed discussion see also Watanabe, et al. 2009; Ross, 2003). Frechtling (2007) writes:

Evaluation is not a process or an approach; it is a family of activities...evaluation includes description, but by its very nature evaluation cannot stop there. It looks at what is happening in various ways and through various lenses, and assesses the value of what is found. "Objectivity" is frequently a criterion for sound, unbiased evaluation, but assessing, valuing, interpreting, and engaging the stakeholder community—all activities that involve judgment—are the essential components of the evaluation enterprise. Evaluation is more than a judgment of ultimate success or failure...what has been called 'summative evaluation.' Instead we believe that evaluation should be designed to inform improvement or modification both in the future and as a project is unfolding. This *formative evaluation* [emphasis original], which acknowledges the importance of looking at both implementation and progress, has become increasingly important over time (pp. 2-3).

The addition of subjective evidence and judgments to the program review repertoire of available techniques made way for a shift in focus of L2 program evaluation efforts, from summative, objective evidence-only approaches to more formative evaluation approaches, underscoring the importance of providing feedback to both teachers and students on the learning process itself or on program improvement.

Most recently, as an alternative to both quasi-experimental, summative research designs, and more formative approaches to program reviews, researchers have applied data-mining strategies to existing datasets in order to conduct longitudinal analyses of programs of interest (Bloomfield et al., 2012; 2013; Mackey, 2014; Ross, 2011; Wagner, 2014). Data mining allows researchers to work with systematically maintained datasets in order to statistically model observed relationships and patterns. For example, working with existing employee testing records and official training records, Bloomfield et al. (2012, 2013) examined rates of skill retention at varying lag-times between test occasions, patterns of change in language skills over time, and the impact of language training and changes in test versions on overall patterns of change. Results indicated that reading and listening L2 proficiency test scores improved over time, while speaking L2 proficiency test scores declined. This finding had important implications for testing policy in that the researchers recommended that the annual proficiency-testing mandate be changed to every other year for the reading and listening skills. Using data-mining strategies, although labor-intensive and time-consuming for researchers in terms of data organization and restructuring, allows researchers to conduct sophisticated analyses without requiring one to directly intervene with a given population of interest.

Irrespective of their design, the goal of most L2 program reviews is to systematically document how well a given intervention, product, or system is working (Brousselle & Champagne, 2011; Frechtling, 2007). While there is an abundance of research describing how best to engage with clients and stakeholders during the design and implementation of a program of interest, there is an absence of literature detailing recommended best practices for using empirical models to document learner progress as students proceed through a given program (see Norris 2006, 2008, 2009a,b, 2013, 2014). To address this issue, logic models are often used to specify the intended causal flow between each phase of a program of study. The following section will introduce the role of evaluation theory and logic modeling as a means of establishing an empirical model within an L2 instructional framework.

Program Theory and Logic Modeling

Before undertaking any form of an empirically-based program investigation, it is first advised to make explicit the assumptions associated with the instructional program or intervention (Brousselle & Champagne, 2011; Connell & Kubish, 1998; Frechtling, 2007). This process is referred to as establishing a Program Theory. Program Theory is defined as “the set of assumptions about the manner in which the program relates to the social benefits it is expected to produce and the *strategy* [original emphasis] and tactics the program has adopted to achieve its goals and objectives” (Callow-Heusser et al., 2005, p. 38). Therefore, a Program Theory defines, a priori, each of the causal links that are expected to occur from project start to goal attainment, allowing for the process of program review to play an integral role in each stage of program planning.

One method of making explicit each of the assumptions associated with a program of interest is the use of a logic model. A logic model is a visual representation of the underlying flow, or logic, of a program. Oftentimes, program managers and stakeholders have an implicit understanding of the underlying purpose of a program but have not made their implicit assumptions explicit (Fretchling, 2007; McLaughlin & Jordan, 1999). The purpose of a logic model is to make clear, via visual representation, each of the causal factors, or elements, related to a program of interest. The individual elements of a logic model include resources, activities, objectives, indicators, impacts (short-term actions), and long-term outcomes (Renger & Titcomb, 2002, p. 494). After the underlying rationale and key components of a program of interest have been defined, the resources required to address a given program's overall goals become clearer. Renger and Titcomb (2002) state, "knowing what causal factors are being targeted first is essential to assessing whether an activity is appropriately targeted, identifying appropriate indicators of change, and writing sound objectives" (p. 494). Figure 2 visually depicts the four major components of a logic model.



Figure 2. Main components of a logic model (adapted from Fretchling, 2007, p. 22)

The first component, "inputs," is defined as the resources brought to a project, typically in terms of monetary funding. The next component, "activities," is defined as the actual actions that are completed to achieve a program's overall goal. The next step, "outputs," is composed of the direct results of implementing an activity. Outputs are represented by the tangible services or products that result from a completed activity, and they are typically expressed numerically. Finally, "outcomes" represent progress toward desired

program goals. Program outcomes are usually associated with an established timeframe outlining expected short-term, medium-term, and long-term goals.

The purpose of employing a logic model is to define the components of each phase of a program and to make the connections between them explicit. How is each piece of the program defined? How does each component work? Do expected connections occur where predicted? When well defined, Program Theory can substitute for classical experimental study using random assignment. Weiss writes, “if predicted steps between an activity and an outcome can be confirmed in implementation, this matching of ‘theory’ to outcomes will lend a strong argument for ‘causality;’ if the evaluation can show a series of micro-steps that lead from input to outcomes, then causal attribution *for all practical purposes* [original italics] seems to be within reach” (as cited in Frechtling, 2007, p. 4). Although the logic model depicted in Figure 2 suggests movement from left to right, when building a model, it is recommended to first start with specifying program outcomes rather than inputs (Frechtling, 2007). It is only by first establishing what a program is trying to accomplish that one can begin building a model that adequately describes the underlying theory of change.

Situating logic modeling within an L2 instructional framework, large-scale foreign language instructional programs are typically composed of both localized indicators of student achievement (e.g., class GPA scores) and standardized summative assessments (e.g., proficiency tests) documenting end-of-program outcomes. In most cases, foreign language programs are viewed as successful if the graduates a given program has produced are adequately prepared to engage in a variety of foreign-language-related activities. Evidence of short-, mid-, or long-term success will vary

widely depending on program goals, but could consist of such information as the number of foreign language majors who are accepted into graduate programs, hired as foreign language instructors, or take positions involving the use of their foreign language on the job. This type of information would represent evidence of program outcomes.

Following Frechtling's (2007) advice and working backwards from established program outcomes, output is typically measured by summative, end-of-program proficiency tests. This type of evidence allows high-level program administrators to ascertain the extent to which students are meeting or exceeding established proficiency requirements necessary for graduation. Observed changes in the number of students meeting or exceeding required program output measures may prompt administrators to review the program activities, typically in the form of coursework or study-abroad activities, through which learners engage in order to meet the established proficiency goals. Finally, program input within the foreign language program, as with nearly all programs, represents the required investment of both time and money to facilitate learners' progression through all required activities, allowing for evidence of program output and outcomes to be documented. Program inputs can also consist of the requirements necessary for entering a given program, such as SAT scores or required thresholds on prerequisite coursework. As stated by Norris (2016), "logic models of this sort are virtually non-existent in language education and applied linguistics, and they certainly have not been utilized as a means for fully testing a program's theory of change about language teaching and learning" (p. 177). The following section will address the dearth of research in this area by discussing how logic modeling can be applied to a large-scale foreign language program.

Applying Program Theory and Logic Modeling Within an L2 Program Evaluation Context

As mentioned briefly in Chapter 1, the DLIFLC provides intensive foreign language instruction for active and reserve members of the military, foreign military students, and civilian personnel working in the federal government. Implementing the strategy of working backwards from the logic model depicted in Figure 2, the expected **output** of the DLIFLC program of study is to produce foreign-language-enabled enlisted personnel capable of supporting a wide variety of foreign language mission-related work across the DoD (D.K. Chapman, Commandant of DLIFLC, personal communication, May 1, 2015). In terms of expected program **outcomes**, upon completion of the DLIFLC program of study, students are expected to achieve ILR proficiency ratings of 2 on the DLPT for both the reading and listening skills and a 1+ on the Oral Proficiency Interview (OPI) for the speaking skill. The length of time students have to achieve these program outcomes varies relative to the difficulty category to which the target language is assigned (for ease of reference, see Table 3 below, adapted from Table 1 in Chapter 1). Higher difficulty categories are associated with both higher minimum DLAB scores and longer target language instructional time (additional information on the DLAB is provided below).

Table 3. Difficulty categories and associated length of instruction at DLIFLC (adapted from Mackey, 2014)

Language Category	Languages	Minimum DLAB Score	No. Weeks of Instruction
I	Spanish, French, Portuguese	95	26
II	German, Indonesian	100	35
III	Dari, Persian Farsi, Russian, Hindi, Urdu, Hebrew, Serbian/Croatian, Tagalog, Turkish	105	48
IV	Arabic (MSA), Arabic (Levantine), Arabic (Iraqi), Chinese Mandarin, Korean, Japanese, Pashto	110	64

In terms of language program **activities** in which learners engage, students enrolled in courses at the DLIFLC attend classes five days per week for seven hours per day.

Learners progress through a series of courses composed of four to five classes at the 100, 200, and 300 levels. Students typically have about two to three hours of homework per night, in addition to their regular military duties as enlisted personnel.

Continuing to move backwards from program activities to program **inputs**, for enlisted personnel to be accepted at the DLIFLC, they must meet minimum performance thresholds on subsections of the ASVAB. The ASVAB is a multiple choice aptitude battery that measures developed abilities and is designed to predict future academic and occupational success in the military. The ASVAB is administered annually to more than one million military applicants, high school students, and post-secondary students. The ASVAB is used for three primary purposes: (1) to determine enlistment eligibility, (2) to assign applicants to military jobs, and (3) to aid students in career exploration. The test itself is designed to measure aptitude in four main ability domains: (1) verbal, (2) mathematical, (3) science, and (4) technical/spatial (Lett et al., 2003). Scores on the

ASVAB are reported by an Armed Forces Qualifying Test (AFQT) score, which is a composite score created from four of nine ASVAB subtests: Paragraph Comprehension, Word Knowledge, Mathematics Knowledge, and Arithmetic Reasoning.²¹ Those who achieve the minimum required percentile scores (calculated relative to a given population of test takers) then qualify to take the DLAB.

The DLAB is an aptitude battery used by the DoD to test an individual's potential for learning a foreign language within a formal training program. It is a selection tool designed to differentially predict foreign language program outcome measures within the DLIFLC language-learning context (Peterson & Al-Haik, 1976; Lett & O'Mara, 1990). The DLAB includes four main sections: (1) biographical data, (2) spoken stress/tone, (3) deductive rule application, and (4) inductive pattern application (Lett, n.d.). Those who demonstrate high scores on the DLAB are encouraged to accept assignments requiring foreign language proficiency within the armed services. To attend the DLIFLC, enlisted personnel must achieve a minimum DLAB score of 95 (scores on the DLAB can range from 12-164); to enroll in Category IV language study, learners must earn a minimum DLAB score of 105, but this score can be waived down to a 95 if necessary (Army Pocket Recruiter Guide, 2013; Lett, n.d.). For an in-depth discussion of both the ASVAB and DLAB tests, see Mackey (2014, pp. 3-9).

In addition to the ASVAB and DLAB screening tools, prior to beginning foreign language training at the DLIFLC, enlisted personnel are required to take a five-day Introduction to Language Studies within the Student Learning Center. This course is designed to "help increase students' levels of preparation for language studies, increase

²¹ Paragraph Comprehension and Word Knowledge are then combined to create a Verbal Expression score.

student motivation, and ultimately help reduce academic attrition rates” (DLIFLC website: <http://www.dliflc.edu/introductiontola.html>). After receiving their language assignments, and prior to engaging beginning language training coursework, students are asked to indicate an answer relative to their preference for studying their assigned language:

I am here to study this language, which is...

1. Not my choice. I would prefer to do something else rather than study a foreign language.
2. Not my choice. I am not motivated to study the assigned language.
3. Not my choice, but I am still motivated to study the assigned language.
4. Based on my second or third choice.
5. Based on my first choice (Lett, n.d.)

The purpose of the activity above is to gauge learners’ self-assessed preference for their assigned language prior to beginning intensive foreign language study at the DLIFLC. Previous research studies have found most correlations between self-ratings on the motivation questionnaire and DLPT/OPI outcomes to be non-significant and to vary substantially across target languages and subskills, as well as from year to year among enlisted personnel (Lett, n.d.; Lett & O’Mara, 1990). Research by Lett (n.d.) referred to as the Language Skill Change Project, which included 1,903 students across four languages, found that learner motivation and preference for studying their assigned target language at the beginning of training was less predictive of attaining DLPT and OPI outcomes than it was *during* language study. It should be noted that within the multiple regression model, Lett (n.d.) used nine blocks of predictors using forced order-of-entry. The Language Preference Self-Assessment variable was entered later in the model (in Block 5). This decision may have attenuated the potential predictive power of the “language preference” variable, since most of the model’s variance was likely to have

already been explained by more robust predictor variables entered earlier in the model, such as general cognitive ability (Block 1) and aptitude (Block 2). Although previous research by Lett (n.d.) has established that a learner’s self-assessed language preference plays a variable role in attaining DLPT/OPI outcomes, it is maintained that language preference likely plays an important role in initial language acquisition, particularly when learners are situated in an intensive training environment such as that of the DLIFLC. Figure 3 below chronologically depicts each component of the DLIFLC program of study relative to its chronological position within a logic model.

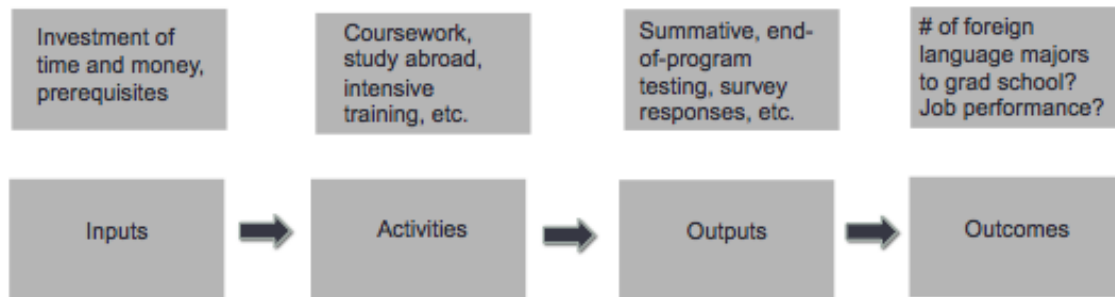


Figure 3. The four main components of a logic model applied within a FL program evaluation context (adapted from Frechtling, 2007, p. 22).

Situating the main components of the DLIFLC L2 instructional program, as shown in **Figure 3** above, addresses the virtual absence of documenting the “variety of factors that conspire to bring out language-learning in a certain way and with specific resulting outcomes” (Norris, 2016, p. 177). The following section details how each of the components within the input, activities, and output categories can be empirically modeled to examine individual variation in learner initial acquisition patterns as students progress through the DLIFLC program of study.

Current Dataset

As mentioned above, previous empirical studies aimed at examining the process through which learners develop proficiency have largely focused on the role of cognitive (e.g., aptitude-related) and non-cognitive (e.g., motivation-related) factors in predicting ultimate program outcomes (e.g., end-of-course DLPT proficiency outcomes) and/or growth, maintenance, or loss in foreign language proficiency test scores over one's career (Bloomfield et al., 2012, 2013; Lett & O'Mara, 1990; Mackey, 2015; Ross, 2011; Wagner, 2014). Lacking in this body of research is consideration of the role that achievement-related variables, such as course grades, may have on influencing end-of-program outcomes while engaging in intensive language training. That is, while previous research has found that general cognitive abilities and language aptitude positively predict end-of-course outcomes (Lett, n.d.; Lett & O'Mara, 1990; Mackey, 2014; Wagner, 2015), few research studies have examined individual patterns in foreign language achievement or how both cognitive and non-cognitive individual difference variables may jointly influence the development of initial acquisition proficiency patterns as learners progress through a program of interest.

To the author's knowledge, just one empirical study has been completed that compares observed initial acquisition achievement and proficiency patterns between Category I and Category IV languages. In research employing the use of path analysis to examine L2 acquisition patterns for Spanish (Category I) and Arabic (Category IV) learners, Masters (2016) found that both AFQT percentile scores and DLAB scores were significant predictors of Arabic coursework success, but did not predict Spanish

coursework success at the DLIFLC.²² This finding provided initial validity evidence for requiring higher aptitude scores for learners studying languages in higher difficulty categories, although the strength of their causal influence waned over time as Arabic learners progressed through coursework. Interestingly, for both languages, and consistent with Lett's (n.d.) findings, learners' Language Preference Self-Assessment scores were found to significantly predict course achievement outcomes but not end-of-program proficiency outcomes.

Masters (2016) also modeled the predictive influence of course-achievement-related variables on end-of-program DLPT outcomes. Results indicated that Arabic 200- and 300-level average outcomes significantly predicted DLPT listening, reading, and speaking results. Arabic 200-level course averages predicted DLPT listening outcomes most strongly, followed by speaking, and reading. Arabic 300-level course averages predicted DLPT listening outcomes most strongly, followed by reading and then speaking. For Spanish, 100-level course averages predicted DLPT speaking outcomes and 200-level course averages predicted DLPT reading outcomes. Lastly, when examining the causal influence of AFQT, DLAB, and Language-Preference Self-Assessment scores on DLPT and OPI outcomes, Masters found that AFQT and DLAB outcomes were significant predictors of reading and speaking outcomes (although a negative causal pathway was found between AFQT and OPI outcomes). Overall the results of Masters' research revealed striking differences in the development of proficiency between learners of languages grouped in different difficulty categories, providing corroborating evidence of the LDC system. That is, with respect to the

²² The path-analytic framework will be described in detail in Chapter 3

aptitude-related predictor variables, since Arabic and Spanish are grouped within different difficulty categories, they were not expected to show similar patterns of proficiency development. The difference in observed proficiency acquisition patterns between Arabic and Spanish corroborates the DLIFLC policy requiring higher aptitude scores for Category IV languages than Category I languages. It is important to note that Masters' research only examined one Category IV language (Arabic) and one Category I language (Spanish). Additional research is necessary to examine the amount of variation that may be present for languages grouped within a single category. Findings from this type of empirical analysis would have important implications for the language difficulty categorization system, particularly since the category to which a language is assigned has a direct impact on the number of weeks of instruction with which learners are provided within the DLIFLC program of study.

Chapter 3: Examining the Coherence of Initial Acquisition Proficiency Development

The purpose of the current chapter is to examine empirically the coherence of initial acquisition patterns for languages grouped within the same category. Building on Masters' (2016) research employing the use of path analysis to investigate L2 language acquisition patterns between languages grouped in different categories, the current chapter will examine patterns in the development of achievement and proficiency outcomes for learners enrolled in languages grouped *within the same category*. Within this chapter, the research questions of interest, associated datasets, and data-analytic procedures employed to address each question will be described in detail. The research questions of interest and associated analyses are organized relative to two main studies, each described in detail below.

Study 1: Purpose of the Study

The purpose of Study 1 is to examine the coherence of initial L2 acquisition patterns within the LDC framework. A given grouping of languages within an LDC framework assumes that the patterns of initial acquisition proficiency development associated with the acquisition of an L2 are invariant for languages situated within the same difficulty category, since languages within the same category are constrained to the same amount of instructional time to meet end-of-program proficiency-testing criteria. That is, unlike the findings by Masters (2016), which found striking differences in acquisitional patterns between languages in different difficulty categories, it logically follows that languages grouped *within the same category* should show similar patterns of development, thus providing validity evidence for the categorization schema. By

grouping languages within the same category, DLIFLC policy makers assume that L2 learners will acquire their assigned target languages in a similar manner, along the same required timeframe. If this assumption is correct, all languages grouped within the same category should display similar acquisitional patterns in the development of the listening, reading, and speaking proficiency skills. The path-analytic approach allows for hypothesized causal pathways between each wave of the DLIFLC instructional context to be modeled statistically, allowing for implicit assumptions concerning L2 acquisition patterns to be modeled explicitly. Each of the research questions of interest associated with Study 1, along with the methodological procedures that were employed to test each research question, are specified below.

Research Questions

Three main research questions are associated with Study 1:

Research Question 1 (RQ1): For languages grouped within the same category, are similar patterns of individual differences in general aptitude, language-specific aptitude, and language preference self-assessment observed in *the prediction of learners' success as they progress through coursework?*

Research Question 2 (RQ2): For languages grouped within the same category, are similar patterns of individual differences in general aptitude, language-specific aptitude, and language preference self-assessment observed in *the prediction of learners' end-of-program outcomes?*

Research Question 3 (RQ3): For languages grouped within the same category, *are homogenous patterns of initial language acquisition observed across languages as learners progress through the DLIFLC program of study?*

Dataset

To investigate the above research questions, extant data from previously conducted research at the University of Maryland Center for Advanced Study of Language was used to conduct Study 1. In order to create longitudinal records for

learners progressing through the DLIFLC, data from four different systematically maintained databases were restructured, merged, and coded, resulting in the aggregation of 244 separate learner variables. Variables of interest associated with these datasets fall into six main categories: (1) Existing test scores (ASVAB and DLAB), (2) Demographic and Biographical Variables, (3) Cognitive and Perceptual Measures, (4) Personality Measures, (5) Motivational Measures, and (6) DLIFLC achievement and proficiency measures.²³ To determine the extent to which observed variation in L2 acquisition patterns found by Masters (2016) is attributable to language categorization differences, analyses were constrained to Category IV languages only. The languages of interest included Arabic, Chinese, and Korean, which were the only Category IV languages with adequate sample sizes for the planned inferential analyses. For Study 1, learners with incomplete records or those who had been “re-cycled” or “re-languaged” were not included in the analysis due to the increased variability these types of learners would likely add to the model.²⁴

Description of Variables in Model

As detailed in Chapter 2, and observed in **Figure 3** on page 26, the logic model associated with the current investigation is composed of seven main variables: (1) ASVAB scores, (2) DLAB Scores, (3) Language Preference Self-Assessment scores, (4) 100-, 200-, and 300-level coursework variables, (5) DLPT reading test score outcomes, (6) DLPT listening test score outcomes, and (7) OPI speaking test score outcomes. The current

²³ For a complete list of predictor and outcome variables, please see Appendix B.

²⁴ “Re-cycled” students are those who are asked to retake previously completed coursework within a given language due to subpar achievement grades. “Re-languaged” students are those who begin their course of study in one language but are asked to switch to another (usually to a language that is lower on the language difficulty categorization scale) due to subpar achievement grades.

section will review the descriptive statistics and mean comparisons for each of the main model variables. Starting with the left-most section of the logic models, three main variables constitute program inputs into the model: (1) AFQT Weighted scores, DLAB scores, and Language Preference Self-Assessment scores.²⁵ For the Arabic, Chinese, and Korean languages, the final datasets were composed of learner data collected between February 12, 2009 and March 22, 2010; February 19, 2009 and February 23, 2010; and February 12, 2009 and February 23, 2010; respectively.

Table 4, Table 5, and Table 6 display the descriptive statistics associated with each dataset.

Table 4. Descriptive Statistics (Arabic)

	N	Minimum	Maximum	Mean	Std. Deviation
AFQT Weighted	241	207	286	252.81	15.23
DLAB	241	95	150	120.00	10.81
Language Preference	241	1	5	4.03	0.98

Table 5. Descriptive Statistics (Chinese)

	N	Minimum	Maximum	Mean	Std. Deviation
AFQT Weighted	98	215	288	254.65	13.79
DLAB	98	102	156	123.86	10.46
Language Preference	98	2	5	4.40	0.81

Table 6. Descriptive Statistics (Korean)

	N	Minimum	Maximum	Mean	Std. Deviation
AFQT Weighted	75	207	278	251.53	13.37
DLAB	75	104	145	119.72	9.03
Language Preference	75	1	5	3.60	1.31

As shown in

²⁵ The DLIFLC transforms ASVAB scores to AFQT weighted scores from subsections of the ASVAB battery. The equation is ((Mathematics Knowledge + Arithmetic Reasoning) + 2(Verbal Expression))
<https://www.thebalance.com/how-the-asvab-afqt-score-is-computed-3354094>.

Table 4 through Table 6 above, the highest AFQT, DLAB, and Language Preference Self-Assessment scores are associated with the Chinese language. Figure 4 through Figure 12 show the associated distributions of the AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores for the Arabic, Chinese, and Korean languages.

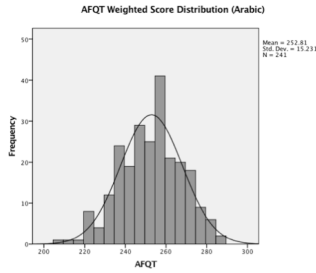


Figure 4. Arabic AFQT Weighted

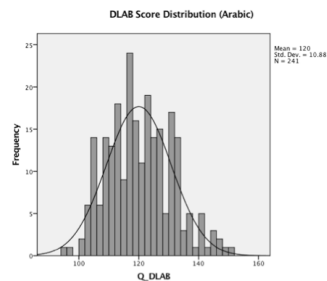


Figure 5. Arabic DLAB

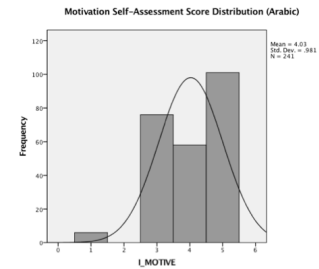


Figure 6. Arabic Lang.Pref.

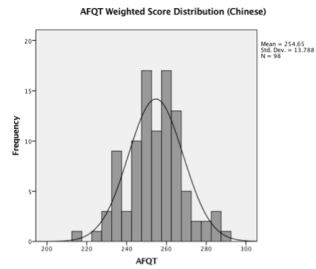


Figure 7. Chinese AFQT Weighted

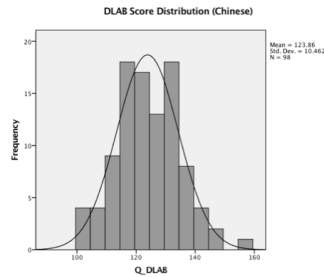


Figure 8. Chinese DLAB

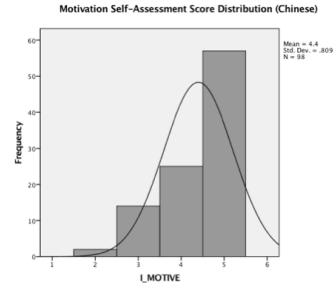


Figure 9. Chinese Lang.Pref.

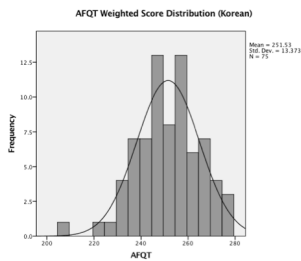


Figure 10. Korean AFQT Weighted

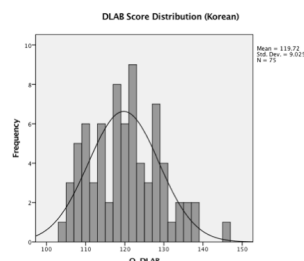


Figure 11. Korean DLAB

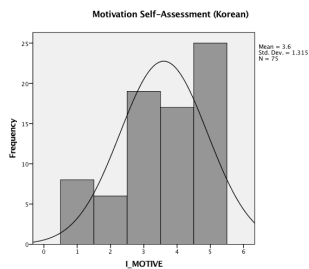


Figure 12. Korean Lang.Pref.

As can be gleaned by visually inspecting Figure 4, Figure 7, and Figure 10, the AFQT weighted scores for all three languages exhibit moderately negative skews. Statistical tests for skewness confirmed this finding (Arabic skewness = -0.185, Chinese skewness =

-0.052, Korean skewness = -0.394). Although the skewness statistic for all three languages was found to be within acceptable ranges (+/- 1.5 to 2.0), it is often found to be unreliable when generated from non-random samples, such as the datasets associated with the current investigation (Lomax, 2011, p. 182). To correct for potential inferential bias when performing the planned path analyses, the AFQT variables were transformed following the recommendation for correcting for moderate negative skew outlined by Tabachnick and Fidell (2007, p. 89).²⁶ **Table 7** outlines the descriptive statistics associated with the AFQT transformation.²⁷

Table 7. Descriptive Statistics (AFQT Transformed Values-Correcting for Negative Skew)

	N	Minimum	Maximum	Mean	Std. Deviation
Arabic	241	1.00	8.94	5.67	1.42
Chinese	98	1.00	8.60	5.71	1.31
Korean	75	1.001	8.49	5.05	1.39

As shown in **Table 7**, the highest AFQT scores are associated with the Chinese language, consistent with what was found for the non-transformed values. Figure 13 through Figure 15 below show the transformed AFQT distribution for the Arabic, Chinese, and Korean languages, as well as their associated skewness and kurtosis statistics.

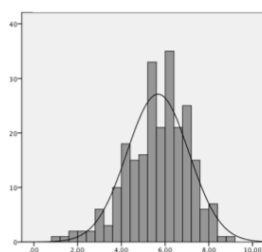


Figure 13. Arabic AFQT

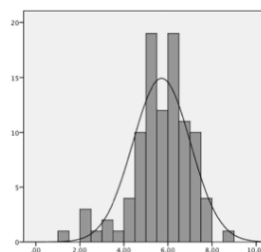


Figure 14. Chinese AFQT

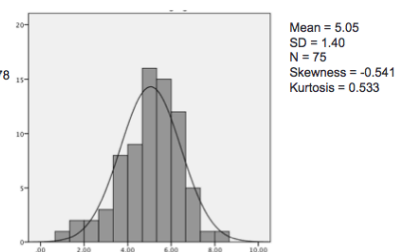


Figure 15. Korean AFQT

²⁶ The correction formula is: $AFQT_NS = \sqrt{K - AFQT \text{ score}}$. K is equivalent to a constant from which each score is subtracted, which is equal to the largest score for the variable of interest + 1. For the Arabic, Chinese, and Korean datasets, K was equivalent to 287, 289, and 279, respectively.

²⁷ Of note, however, is that “the interpretation of a reflected variable is just the opposite of what it was: if big numbers meant good things prior to the reflecting the variable, big numbers mean bad things afterwards” (Tabachnick & Fidell, 2007, p. 88, footnote 16). The nature of the transformations (e.g., positive or negative) will be taken into account when interpreting the output in the planned path-analytic procedure.

(Transformed)

(Transformed)

(Transformed)

As shown in Figure 13 through Figure 15, even after the transformations, all three languages continue to exhibit slight skewness with reversed distributions given the performed reflections on the previously negative distributions. Although the transformations did not completely correct for the observed skewness across languages, the transformed values will be used as input for the planned inferential analyses since they generally improve the stability of inferences made from statistical analyses by reducing the impact that outliers have on observed outcomes (Tabachnick & Fidell, 2007, p.86).

With respect to DLAB-related distributions, all three languages exhibit moderately positive skews (Arabic skewness = 0.293, Chinese skewness = 0.182, Korean skewness = 0.322), as shown in Figure 5, Figure 8, and Figure 11. Also following guidance outlined by Tabachnick & Fidell (2007), transformations were performed for the Arabic, Chinese, and Korean languages to correct for positive skew. **Table 8** below reports the descriptive statistics associated with the transformation.²⁸

Table 8. Descriptive Statistics (DLAB Transformed Values)

	N	Minimum	Maximum	Mean	Std. Deviation
Arabic	241	9.75	12.25	10.94	0.50
Chinese	98	10.10	12.50	11.12	0.47
Korean	75	10.12	12.04	10.93	0.41

As shown in Table 8, the highest DLAB outcomes are associated with the Chinese language, consistent with what was found with the non-transformed values. **Figure 16** through **Figure 18** show the transformed DLAB distribution for the Arabic, Chinese, and Korean languages, as well as their associated skewness and kurtosis statistics.

²⁸ The correction for moderate positive skew is: $DLAB_PS = \sqrt{DLAB \text{ score}}$.

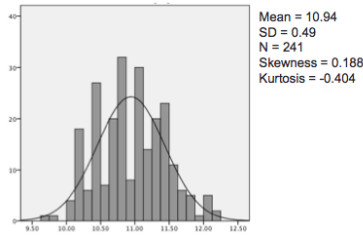


Figure 16. Arabic DLAB (Transformed)

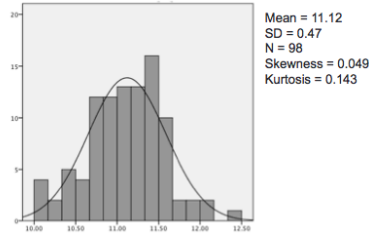


Figure 17. Chinese DLAB (Transformed)

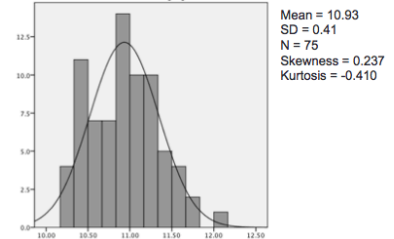


Figure 18. Korean DLAB (Transformed)

As shown above, the correction for positive skew improved the skewness statistic for all three languages. These transformed values will be used as inputs into the planned inferential analyses.

Lastly, as shown in Figure 6, Figure 9, and Figure 12, the Language Preference Self-Assessment distributions for all three languages exhibited moderately negatively skewed distributions. Following the same procedure as the AFQT weighted scores, the Language Preference Self-Assessment variables were also corrected for moderate negative skew following the guidance outlined by Tabachnick & Fidell (2007, p. 89).²⁹

Table 9 reports the descriptive statistics associated with the transformation.

Table 9. Descriptive Statistics (Language Preference Self-Assessment Transformed Values)

	N	Minimum	Maximum	Mean	Std. Deviation
Arabic	241	1.00	2.24	1.36	0.34
Chinese	98	1.00	2.00	1.23	0.30
Korean	75	1.00	2.24	1.49	0.42

Figure 19 through Figure 21 show the transformed Language Preference Self-Assessment distributions for the Arabic, Chinese, and Korean languages, as well as their associated skewness and kurtosis statistics.

²⁹ The correction formula is: $Lang_Pref_NS = \sqrt{K - Lang_Pref\ self\text{-}assessment\ score}$. K is equivalent to a constant from which each score is subtracted, which is equal to the largest score for the variable of interest + 1. For the Arabic, Chinese, and Korean datasets, K was equivalent to 6.

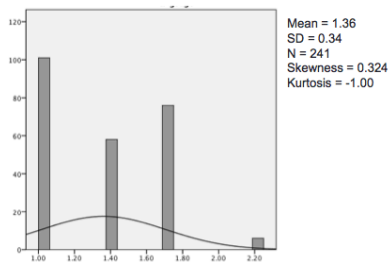


Figure 19. Arabic Lang.Pref.
(Transformed)

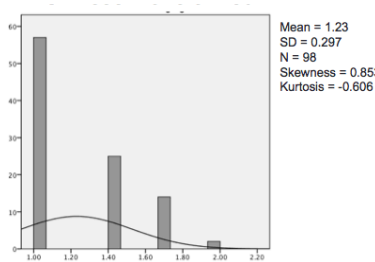


Figure 20. Chinese Lang.Pref.
(Transformed)

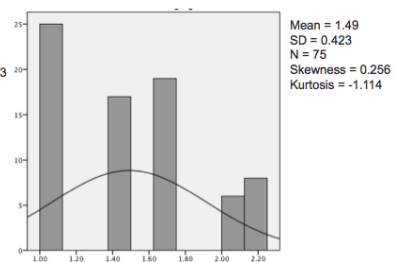


Figure 21. Korean Lang.Pref.
(Transformed)

As shown in Figure 19 through Figure 21, the transformation yielded modest improvements to the overall skewness statistics for the Arabic, Chinese, and Korean languages. These values will be used as input for the planned statistical analyses. Overall, although each initial screening variable required transformations, similar distributional patterns were found for the Arabic, Chinese, and Korean languages, which are expected since they are grouped within the same language category.

To test for statistically significant differences in mean AFQT, DLAB, and Language Preference Self-Assessment scores between the Arabic, Chinese, and Korean languages, one-way ANOVAs were calculated between each language group for each of the three screening variables using the non-transformed values, for ease of interpretability. For the AFQT variable, no significant differences were found between languages, $F(2, 411) = 1.03, p = 0.358$. For the DLAB variable, significant differences were found between languages, $F(2, 411) = 5.27, p = 0.006$. Tukey post hoc tests revealed that DLAB scores were higher for the Chinese language ($M = 123.86, SD = 10.462$) than for the Arabic language ($M = 120.00, SD = 10.881$) and for the Chinese language ($M = 123.86, SD = 10.462$) than for the Korean language ($M = 119.72, SD = 9.025$). No significant differences in DLAB scores were found between the Arabic and

Korean languages. Lastly, significant differences were also found for Language Preference Self-Assessment variable, $F(2, 411) = 13.191, p = 0.000$. Tukey post hoc tests revealed that Language Preference Self-Assessment scores were higher for the Chinese language ($M = 4.40, SD = 0.81$) than for the Arabic language ($M = 4.03, SD = 0.98$) or Korean language ($M = 3.60, SD = 1.32$) and also higher for the Arabic language ($M = 4.03, SD = 0.98$) than the Korean language ($M = 3.60, SD = 1.32$). **Figure 22** below visually confirms the findings discussed above for each Wave 1 variable for each language.³⁰

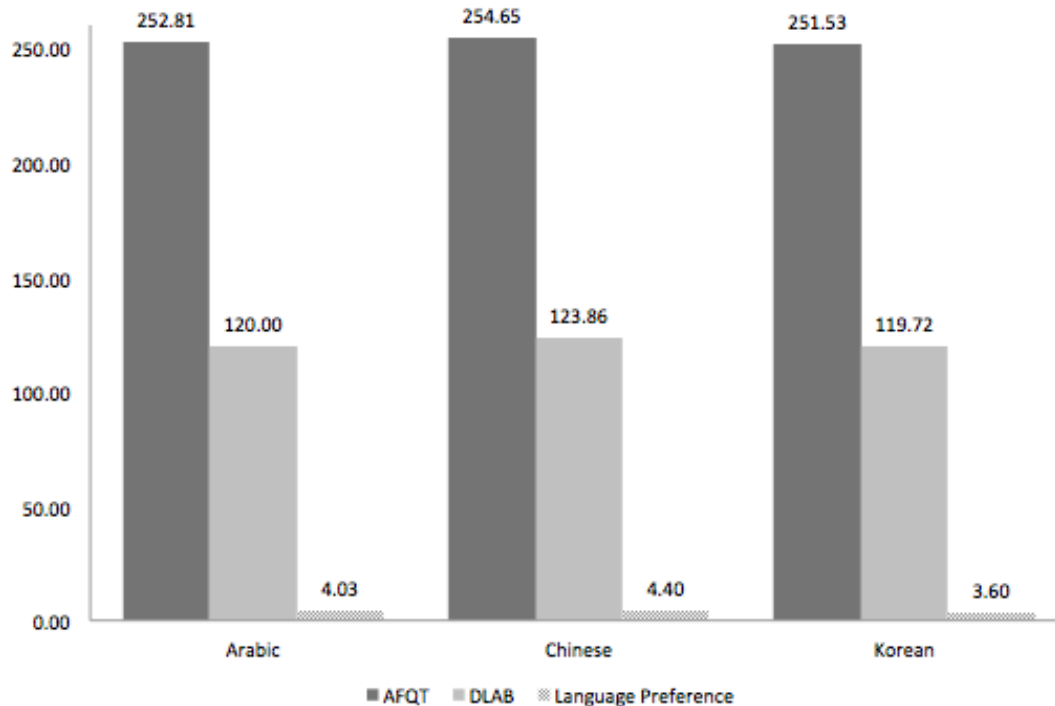


Figure 22. Non-Transformed Arabic, Chinese, and Korean AFQT, DLAB, and Language Preference Self-Assessment scores; AFQT scale: 0-300, DLAB scale: 0-160, Language Preference Self-Assessment scale: 0-5)

As shown in Figure 22, the highest AFQT, DLAB, and Language Preference Self-Assessment scores are associated with the Chinese language.

³⁰ A drawback of performing transformations to raw data is the loss of a comprehensible scale of reference. For ease of interpretability for the reader, although the transformed variables will be used as input into the planned analyses, the non-transformed data is visualized in Figures 22 and 41.

The next section of the logic model is composed of “Program Activities,” which are represented as learners’ average 100-, 200-, and 300-level coursework outcomes.

Table 10 through Table 12 display the descriptive statistics associated with the Arabic, Chinese, and Korean languages.

Table 10. Descriptive Statistics (Arabic Coursework)

	N	Minimum	Maximum	Mean	Std. Deviation
Average 100-level coursework outcomes	241	2.34	4.00	3.27	0.39
Average 200-level coursework outcomes	241	2.12	4.00	3.33	0.39
Average 300-level coursework outcomes	241	2.20	4.00	3.34	0.41

Table 11. Descriptive Statistics (Chinese Coursework)

	N	Minimum	Maximum	Mean	Std. Deviation
Average 100-level coursework outcomes	98	2.68	4.00	3.49	0.35
Average 200-level coursework outcomes	98	2.38	4.00	3.49	0.32
Average 300-level coursework outcomes	98	2.86	4.00	3.53	0.27

Table 12. Descriptive Statistics (Korean Coursework)

	N	Minimum	Maximum	Mean	Std. Deviation
Average 100-level coursework outcomes	75	2.60	4.00	3.56	0.30
Average 200-level coursework outcomes	75	2.60	4.00	3.49	0.29
Average 300-level coursework outcomes	75	2.82	4.00	3.58	0.27

As can be gleaned from

Table 10 through Table 12, the highest 100-level coursework averages are associated with the Korean language, the highest-200-level coursework averages are associated with the Chinese and Korean languages, and the highest 300-level coursework averages are associated with the Korean language. Figure 23 through Figure 31 show the associated distributions for the 100-, 200-, and 300-level coursework outcomes.

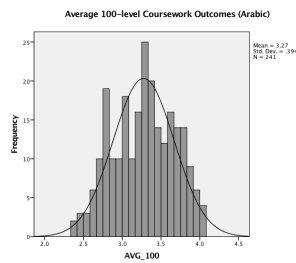


Figure 23. Arabic 100-level

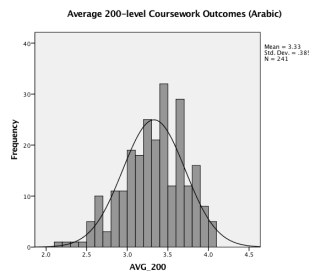


Figure 24. Arabic 200-level

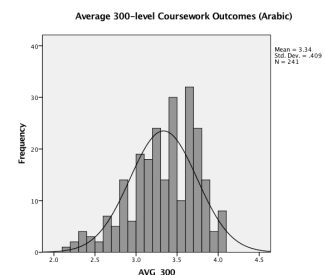


Figure 25. Arabic 300-level

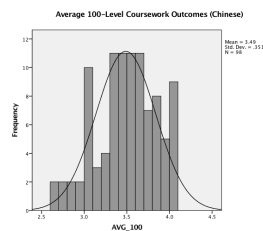


Figure 26. Chinese 100-level

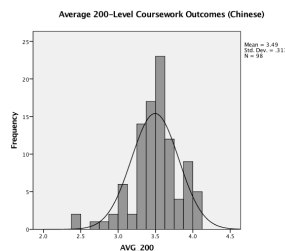


Figure 27. Chinese 200-level

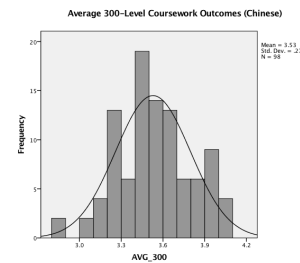


Figure 28. Chinese 300-level

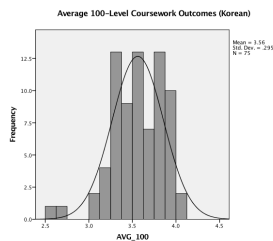


Figure 29. Korean 100-level

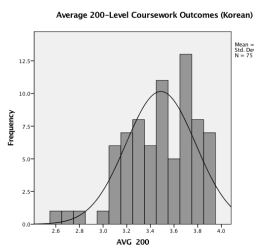


Figure 30. Korean 200-level

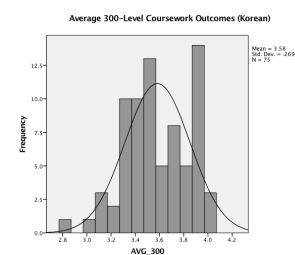


Figure 31. Korean 300-level

As visually depicted in Figure 23 through Figure 31 above, the average 100, 200-, and 300-level coursework outcomes show moderately negatively skewed distributions.

Statistical tests for skewness confirm this finding (Arabic 100-, 200-, and 300-level

skewness = -0.128, -0.439, -0.570, respectively; Chinese 100-, 200-, and 300-level skewness = -0.384, -0.847, -0.340, respectively; Korean 100-, 200-, and 300-level skewness = -0.750, -0.588, -0.271, respectively). Following the same procedure as the AFQT and Language Preference Self-Assessment variables, the 100-, 200-, and 300-level coursework outcomes were corrected for moderate negative skew, following the guidance outlined by Tabachnick & Fidell (2007, p. 89).³¹ **Table 13** through **Table 15** display the descriptive statistics associated with the transformed variables.³²

Table 13. Descriptive Statistics: Transformed Values (Arabic Coursework)

	N	Minimum	Maximum	Mean	Std. Deviation
Average 100-level coursework outcomes	241	1.00	1.63	1.30	0.15
Average 200-level coursework outcomes	241	1.00	1.70	1.28	0.15
Average 300-level coursework outcomes	241	1.00	1.67	1.28	0.15

Table 14. Descriptive Statistics: Transformed Values (Chinese Coursework)

	N	Minimum	Maximum	Mean	Std. Deviation
Average 100-level coursework outcomes	98	1.00	1.52	1.22	0.14
Average 200-level coursework outcomes	98	1.00	1.61	1.22	0.13
Average 300-level coursework outcomes	98	1.00	1.46	1.21	0.11

Table 15. Descriptive Statistics: Transformed Values (Korean Coursework)

	N	Minimum	Maximum	Mean	Std. Deviation
Average 100-level	75	1.00	1.55	1.19	0.12

³¹ The correction formula for AVG_100, 200, and 300_NS = SQRT(K-100-, 200-, 300-level coursework average). K is equivalent to a constant from which each score is subtracted, which is equal to the largest score for the variable of interest + 1. For the Arabic, Chinese, and Korean datasets, K was equivalent to 5.

³² The same caution in interpreting the outcomes of the AFQT and language preference self-assessment variables (noted in footnote 24) applies to the interpretation of the transformed average coursework outcome variables. Similarly, the correction for negative skew will be taken into account when interpreting the observed output in the planned path-analytic procedure.

coursework outcomes					
Average 200-level coursework outcomes	75	1.03	1.55	1.22	0.12
Average 300-level coursework outcomes	75	1.00	1.48	1.19	0.11

As can be gleaned from Table 13 through Table 15, the highest 100-level and 200-level coursework averages are associated with the Arabic language, and the highest 300-level coursework averages are associated with the Korean language. The 100- and 200-level average coursework findings differ from those established using the non-transformed values, in which Korean and Arabic were found to have the highest coursework averages, respectively. Figure 23 through Figure 31 show the transformed Average 100-, 200-, and 300-level coursework outcomes for the Arabic, Chinese, and Korean languages, as well as their associated skewness and kurtosis statistics.³³

³³ The correction formula is: $100, 200, 300_Level_GPA_NS = \sqrt{K - AVG_100, 200_300\text{-level GPA}}$. K is equivalent to a constant from which each score is subtracted, which is equal to the largest score for the variable of interest + 1. For the Arabic, Chinese, and Korean datasets, K was equivalent to 5.

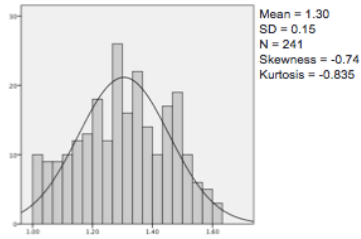


Figure 32. Arabic 100-level (T)

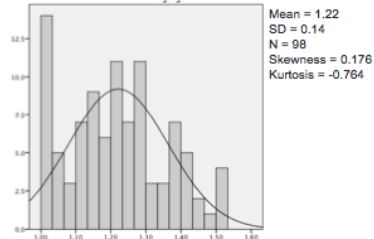


Figure 33. Chinese 100-level

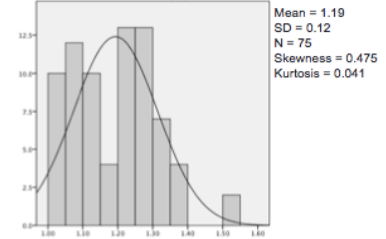


Figure 34. Korean 100-level (T)

(T)

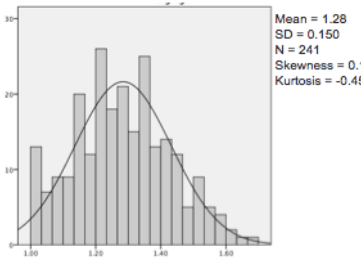


Figure 35. Arabic 200-level (T)

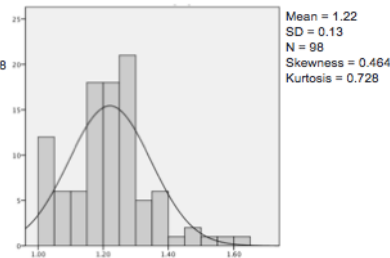


Figure 36. Chinese 200-level (T)

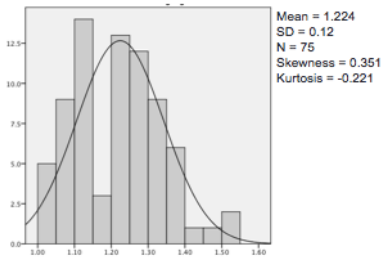


Figure 37. Korean 200-level (T)

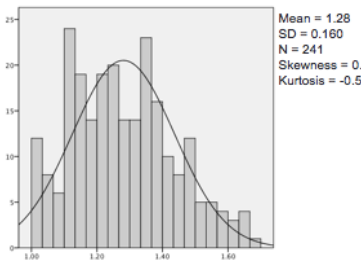


Figure 38. Arabic 300-level (T)

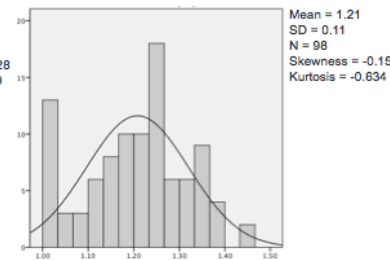


Figure 39. Chinese 300-level (T)

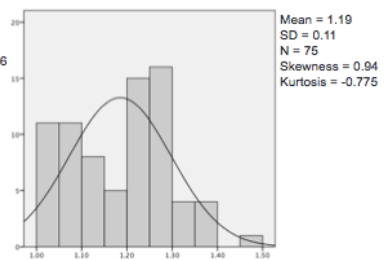


Figure 40. Korean 230-level (T)

As can be gleaned from Figure 32 through Figure 40, the 100-, 200-, and 300-level coursework transformations corrected the negative skewness observed for the non-transformed data; these values will be used as input for the planned inferential analyses.

To test for statistically significant differences in mean 100-, 200-, and 300-level coursework outcomes, one-way ANOVAs were calculated between each language group for each of the 100-, 200-, and 300-level average coursework variables using the non-transformed values, for ease of interpretability. Significant differences were found

between the 100-level ($F(2, 411) = 23.15, p < 0.001$), 200-level ($F(2, 411) = 10.58, p < 0.001$), and 300-level ($F(2, 411) = 18.98, p < 0.001$) average coursework variables. For the 100-level average coursework outcomes, Tukey post hoc tests revealed that scores were statistically significantly higher for the Korean language ($M = 3.56, SD = 0.30$) than for the Arabic language ($M = 3.27, SD = 0.40$). Chinese 100-level average coursework outcomes ($M = 3.49, SD = 0.35$) were also found to be statistically significantly higher than Arabic 100-level average coursework outcomes ($M = 3.27, SD = 0.40$). No significant differences were found between the Chinese and Korean languages. The same general patterns of mean differences were found for both the 200- and 300-level average coursework outcomes. Tukey post hoc tests revealed that 200-level average coursework outcomes were statistically significantly higher for the Korean language ($M = 3.49, SD = 0.32$) than for the Arabic language ($M = 3.33, SD = 0.39$) and for the Chinese language ($M = 3.50, SD = 0.32$) than the Arabic language ($M = 3.33, SD = 0.39$). No significant differences were found between the Chinese and Korean languages. Lastly, Tukey post hoc tests revealed that 300-level average coursework outcomes were statistically significantly higher for the Korean language ($M = 3.58, SD = 0.27$) than for the Arabic language ($M = 3.34, SD = 0.41$) and for the Chinese language ($M = 3.53, SD = 0.27$) than the Arabic language ($M = 3.34, SD = 0.41$). No significant differences were found between the Chinese and Korean languages. Figure 41 below visually confirms the findings discussed above for each of the Wave 2 variables across languages.

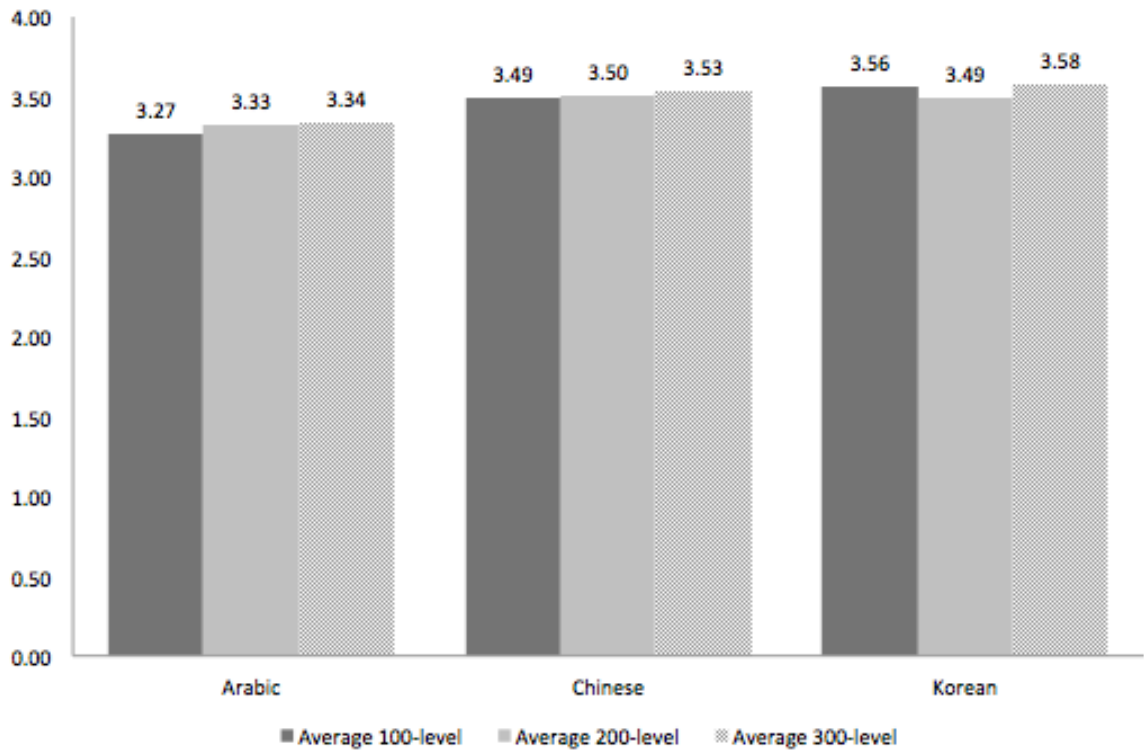


Figure 41. Non-Transformed Arabic, Chinese, and Korean 100-, 200-, and 300- average coursework outcomes

As shown in Figure 41 above, highest average coursework outcomes tend to be associated with the Korean language, followed by the Chinese and Arabic languages.

The final component of the logic model is composed of program “Outputs”, which are represented by outcomes on DLPT reading, DLPT listening, and OPI speaking proficiency test scores. Trends in these outcome measures were also examined for comparability across the Arabic, Chinese, and Korean languages. Table 16 through Table 18 display the descriptive statistics associated with each language.³⁴

³⁴ Raw DLPT and OPI outcome variables are reported on a rating scale (0, 1+, 2, 2+, 3, and 3+). The seven ordinal-level outcome variables were transformed to a continuous scale in order to compute the planned inferential analyses. Per Tabachnick and Fidell (2007, p. 7), this transformation is allowed provided that there is a minimum of seven categories associated with the ordinal-level data. The transformation is thus as follows: 0 = 0, 1 = 10, 1+ = 16, 2 = 20, 2+ = 26, 3 = 30, 3+ = 36.

Table 16. Descriptive Statistics (Arabic End-of-Program Outcomes; Scale = 0 - 30)

	N	Minimum	Maximum	Mean	Std. Deviation
DLPT Listening	241	6	30	23.04	5.45
DLPT Reading	241	6	30	23.64	5.38
OPI Speaking	241	10	26	17.31	2.24

Table 17. Descriptive Statistics (Chinese End-of-Program Outcomes; Scale = 0 - 30)

	N	Minimum	Maximum	Mean	Std. Deviation
DLPT Listening	98	6	30	24.00	5.28
DLPT Reading	98	16	30	24.04	3.60
OPI Speaking	98	16	26	18.73	2.18

Table 18. Descriptive Statistics (Korean End-of-Program Outcomes; Scale = 0 - 30)

	N	Minimum	Maximum	Mean	Std. Deviation
DLPT Listening	75	16	30	24.03	5.04
DLPT Reading	75	20	30	26.35	3.67
OPI Speaking	75	16	20	17.49	1.95

As shown in Table 16 through Table 18, comparable average DLPT and OPI outcomes are found across all three languages. Figure 42 through Figure 50 show the associated distributions for the end-of-program proficiency outcomes across the Arabic, Chinese, and Korean languages.

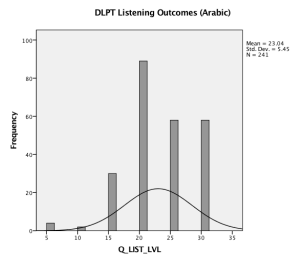


Figure 42. Arabic DLPT Listening

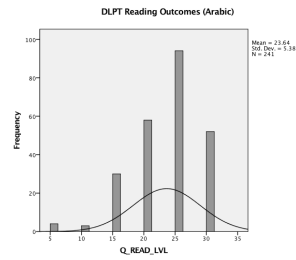


Figure 43. Arabic DLPT Reading

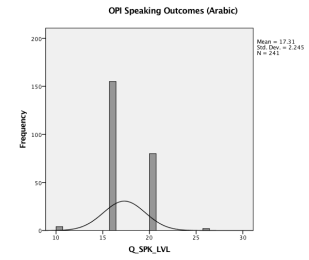


Figure 44. Arabic OPI Speaking

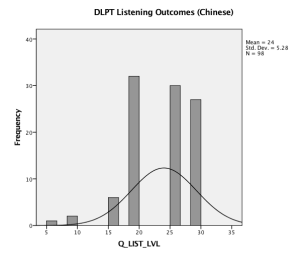


Figure 45. Chinese DLPT Listening

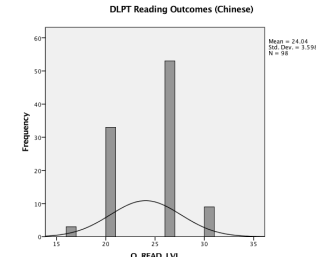


Figure 46. Chinese DLPT Reading

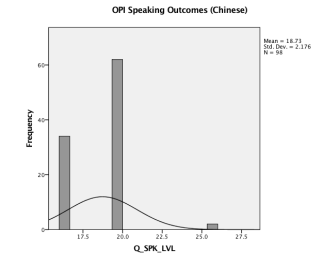


Figure 47. Chinese OPI Speaking

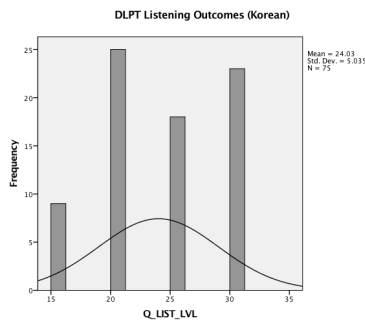


Figure 48. Korean DLPT Listening

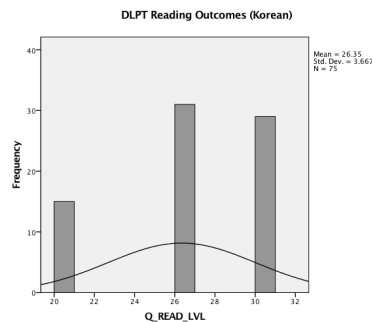


Figure 49. Korean DLPT Reading

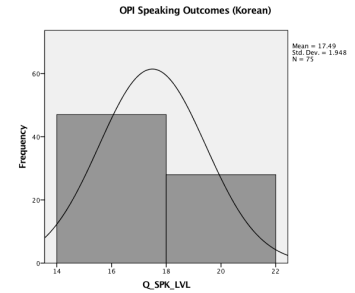


Figure 50. Korean OPI Speaking

The visual depiction of end-of-program training outcomes shows similar distributional patterns across the three languages, with the majority of learners meeting established 2, 2+, 1+ DLPT and OPI criterion outcomes (for ease of reference 2 = 20, and 1+ = 10). Since the DLPT and OPI variables will not be used as predictors in the planned analyses, no statistical transformations were required.

To test for statistically significant differences in mean DLPT listening, reading and OPI speaking outcomes, one-way ANOVAs were calculated between each language group for each of the listening, reading, and speaking skills. Significant differences were found between the reading ($F(2, 411) = 9.438, p < 0.001$), and speaking ($F(2, 411) = 15.222, p < 0.001$) proficiency test score outcomes. For the reading skills, Tukey post hoc tests revealed that Korean proficiency test score outcomes ($M = 26.35, SD = 3.667$) were statistically significantly higher than Arabic proficiency test score outcomes ($M = 23.64, SD = 5.381$). Korean reading proficiency test score outcomes ($M = 23.64, SD = 5.381$) were also found to be statistically significantly higher than Chinese reading proficiency test score outcomes ($M = 24.04, SD = 3.598$). No significant differences were found between the Arabic and Chinese languages. For the speaking skill, Tukey post hoc tests revealed statistically significantly higher Chinese OPI speaking proficiency test score outcomes ($M = 18.73, SD = 2.176$) than Arabic OPI speaking proficiency test score outcomes ($M = 17.31, SD = 2.245$). Chinese OPI speaking proficiency test score outcomes ($M = 18.73, SD = 2.176$) were also found to be significantly higher than Korean OPI speaking proficiency test score outcomes ($M = 17.49, SD = 1.948$). Figure 51 below visually confirms the findings discussed above for each of the Wave 3 end-of-program outcome variables across languages.

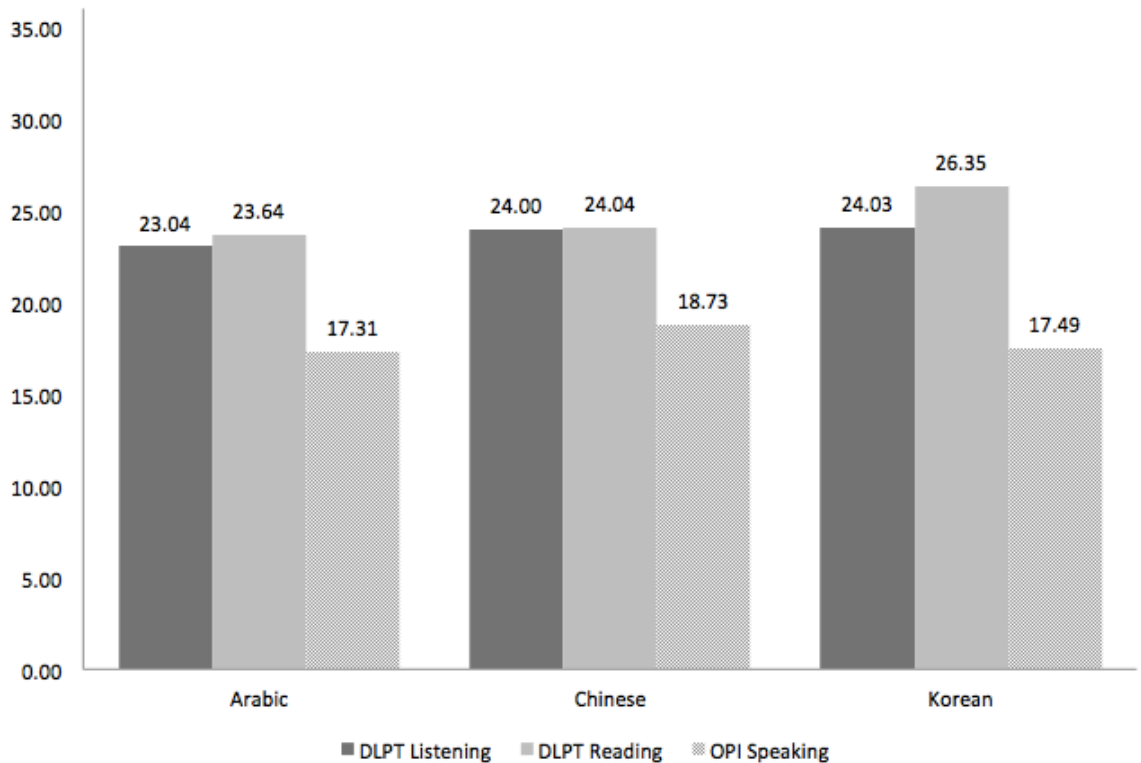


Figure 51 Transformed Arabic, Chinese, and Korean DLPT and OPI outcomes

As shown above in Figure 51, the highest DLPT listening test score outcomes are associated with the Chinese and Korean languages. Table 19 shows the percentage of learners to meet DLPT and OPI criterion cut-off scores of 2 in reading, 2 in listening, and 1+ in speaking (equivalent to 20, 20, and 16 for the current scale, as noted in footnote 34).

Table 19. Percentage of learners to meet established DLPT and OPI criterion scores

	N	% that met DLPT Reading Criterion (Level 2 = 20)	% that met DLPT Listening Criterion (Level 2 = 20)	% that met OPI Speaking Criterion (Level 1+ = 16)
Arabic	241	84.6%	81.7%	98.3%
Chinese	98	96.9%	90.8%	100%
Korean	75	100%	88.0%	100%

Overall, Korean language learners had the highest percentage of students that met the DLPT reading criterion cut-off score (100%). Chinese language learners had the highest percentage of students to meet the DLPT listening criterion cut-off score (90.85). Chinese

and Korean learners had the highest percentage of learners to meet OPI speaking criterion cut-off scores (100%). The descriptive statistics and mean comparisons of between each of the variables contained within the Arabic, Chinese, and Korean logic models play an important role in hypothesizing predictions to be made concerning how each of the variables will relate within an overall model, as well as hypothesized differences between models. The research design and hypothesized model predictions will be discussed in detail in the following section.

Research Design: Panel Study

As mentioned above, the research design associated with Study 1 took the form of a panel study. Learners at the DLIFLC progress through three main panels, or waves of progress. These waves of progress correspond to the three main components of the logic model, described in detail in Chapter 2. Wave 1 is composed of “Program Inputs,” Wave 2 is composed of “Program Activities,” and Wave 3 is composed of program “Outcomes.” This relationship is visually depicted in Figure 52.

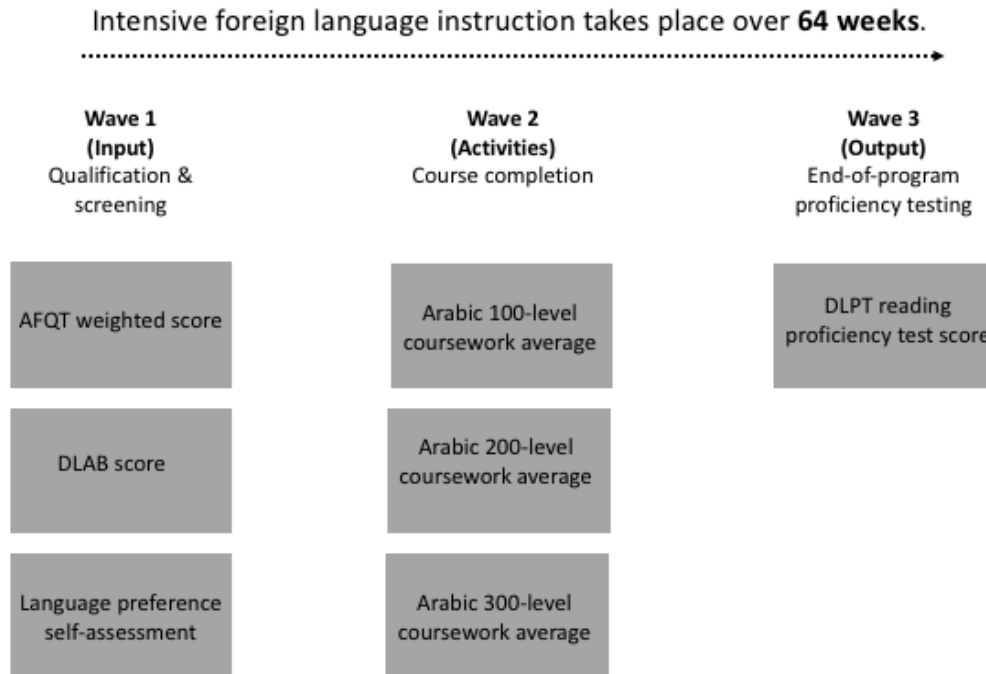


Figure 52. Visual depiction of the variables associated with each wave of the panel study

As depicted above, the first wave is referred to as the “Qualification and screening” stage, which consists of learners’ outcomes on the AFQT, DLAB, and self-reported Language Preference Self-Assessment instruments. To prepare the data for the path analyses, weighted AFQT weighted scores were manually calculated from the raw ASVAB data using the following equation: $AFQT \text{ weighted scores} = ((\text{Verbal Expression Scores} * 2) + (\text{Mathematics Knowledge} + \text{Arithmetic Reasoning}))$.³⁵

The second wave, referred to as the “Course completion” stage, consists of learners’ averages on a 4.0 grading scale as they progress through the series 100-, 200-,

³⁵ The ASVAB is composed of 10 total sections. A subset of these sections, Verbal Expression, Mathematics Knowledge, and Arithmetic Reasoning, are used to calculate Armed Forces Qualification Test scores. Established threshold AFQT scores are then used as screening criteria to determine DLAB eligibility. See: http://official-asvab.com/understand_coun.htm. This calculation is a notable change from Masters (2016) in which AFQT percentile scores were used as the predictor variable. Since AFQT percentile scores represent relative scores from a given reference group, and the current dataset is composed of AFQT scores from multiple reference groups, it was determined that the raw AFQT score would serve as a more stable predictor.

and 300-level DLIFLC coursework, described above.³⁶ The outcome data associated with the achievement-related variables took the form of average 100-, 200- and 300-level course grades learners obtained on a 4.0 grading scale. The third and final stage is referred to as the “End-of-program proficiency-testing stage” and is composed of learners’ DLPT listening, reading, and speaking proficiency test score outcomes. All measured variables are therefore ordered by temporal priority, allowing for the planned path analysis to take place. No additional manipulations were made to the DLAB or Language Preference Self-Assessment scores, and DLPT/OPI outcome data. The following section will describe in detail the path-analytic research methodology.

Path Analysis

Path analysis is a method of analysis that allows for the *simultaneous estimation* of hypothesized causal relationships between measured (observed) or latent (unobserved) variables at the individual, rather than group, level (which serves as the basis for comparison for common statistical procedures, such as ANOVAs) . Path analysis is considered a special type of structural equation modeling (SEM) that involves multiple regression analyses of a given set of variables (Tabachnick & Fidell, 2007, p. 676). Fundamental to the completion of a path analysis is the creation of a path diagram. Path diagrams are visual depictions of a hypothesized set of relationships between observed (measured) or latent (unmeasured) variables and serve as representations of the underlying structural equations required for the analysis. In a path diagram, observed variables (also referred to as indicators or manifest variables) are denoted by squares or

³⁶ Although learners progress through each of their courses in succession, the author decided to situate course-achievement outcomes within a single wave in order to create a more parsimonious model.

rectangles. Unobserved variables (also referred to as latent variables or constructs) are represented by circles or ovals.

Two types of arrows are used between variables to denote relationships between variables within a path model. Directional (or single-headed) arrows are used to represent hypothesized direct, or causal, effects. Non-directional (or double-headed) arrows are used to represent non-structural (or noncausal) variation or covariation. Independent (or predictor) variables only have arrows pointing out of them and are referred to as *exogenous* variables. Dependent variables have arrows pointing into them and are referred to as *endogenous* variables (Byrne, 2001; Hancock, 2011; Tabachnick & Fidell, 2007). The absence of directional or bidirectional arrows indicates that there is no hypothesized relationship between the variables in the model.

As denoted by use of rectangles in Figure 52 above, the current investigation is composed only of measured variables given that observed scores are associated with each rectangle. Starting with Wave 1, AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores take the place of the left-most predictor variables, or exogenous variables, shown below in Figure 53. For all figures used to visually depict the path-analytic outcomes within the current investigation, grey lines denote hypothesized, but not significant, causal pathways between waves. Dark lines denote significant causal pathways between waves. Dotted lines denote significant negative causal pathways between waves. For ease of visual inspection of each model outcome, the Wave 2 concurrent variation modeled between each of the 100-, 200-, and 300-level average coursework outcomes was statistically modeled, but it is not shown in the path-analytic diagrams.

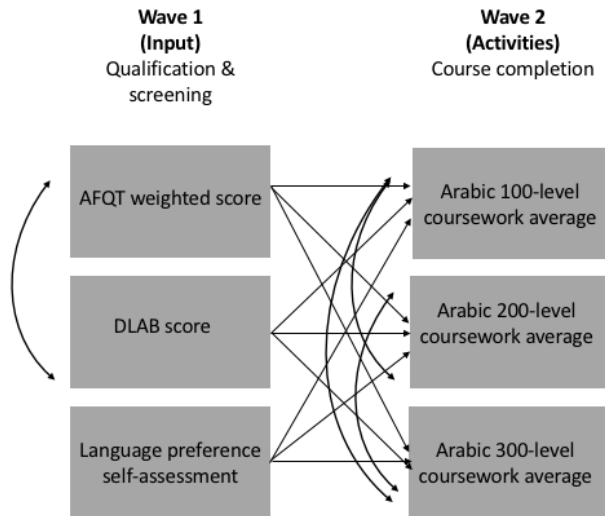


Figure 53. Example Wave 1 to Wave 2 Partial Path Model

The position of the Wave 1 AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores, shown in Figure 53, are indicative of their temporal priority relative to the Wave 2 course completion variables (Average 100-level course outcomes, Average 200-level course outcomes, Average 300-level course outcomes) since the Wave 1 outcomes are collected first, before any of the other variables within the model.³⁷ Moving from left to right, and indicative of the temporal causal flow between waves, outcomes on average 100-, 200-, and 300-level classroom achievement variables are expected to be predicted by Wave 1 variables, AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores, as well as to concurrently covary with the achievement outcome variables within the same wave.

As can be gleaned from Figure 53 three single headed arrows are associated with each of the Wave 1 exogenous variables to each of the Wave 2 course achievement

³⁷ In addition to the creation of a more parsimonious model, although 100-, 200-, and 300-level courses are taken sequentially, they are modeled as being taken concurrently in order to account for the observed high correlations between each of the average 100-, 200-, and 300-level course outcomes (see Appendix D). These high correlations would likely suppress their hypothesized causal influence on the Wave 3 predictor variables, to be discussed next.

variables. This notation denotes a hypothesized causal relationship between AFQT weighted scores, DLAB scores, and Language-Preference Self-Assessment scores and average 100-, 200-, and 300-level coursework outcomes. Four double-headed arrows can also be observed in Figure 53: one in Wave 1 and three in Wave 2. The Wave 1 double-headed arrow denotes the covariation between AFQT scores and DLAB scores given that potential students are recommended to take the DLAB based upon whether minimum AFQT score thresholds are met. It also reflects the observed covariation between these variables for each language (detailed in the correlation matrices in Appendix D). The Wave 2 double-headed arrows denote observed high covariation between average 100-, 200-, and 300-level coursework outcomes.

Continuing to progress through the model from left to right, controlling for all antecedent and concurrent variables, Wave 3 variable DLPT reading proficiency test score outcomes (used as an example) are expected to be predicted by Wave 2 variables average 100-, 200-, and 300-level coursework outcomes, as well as all other modeled concurrent covariation between each of the endogenous course achievement variables. Figure 54 below visually depicts the hypothesized relationship between Wave 2 variables and Wave 3 variables.

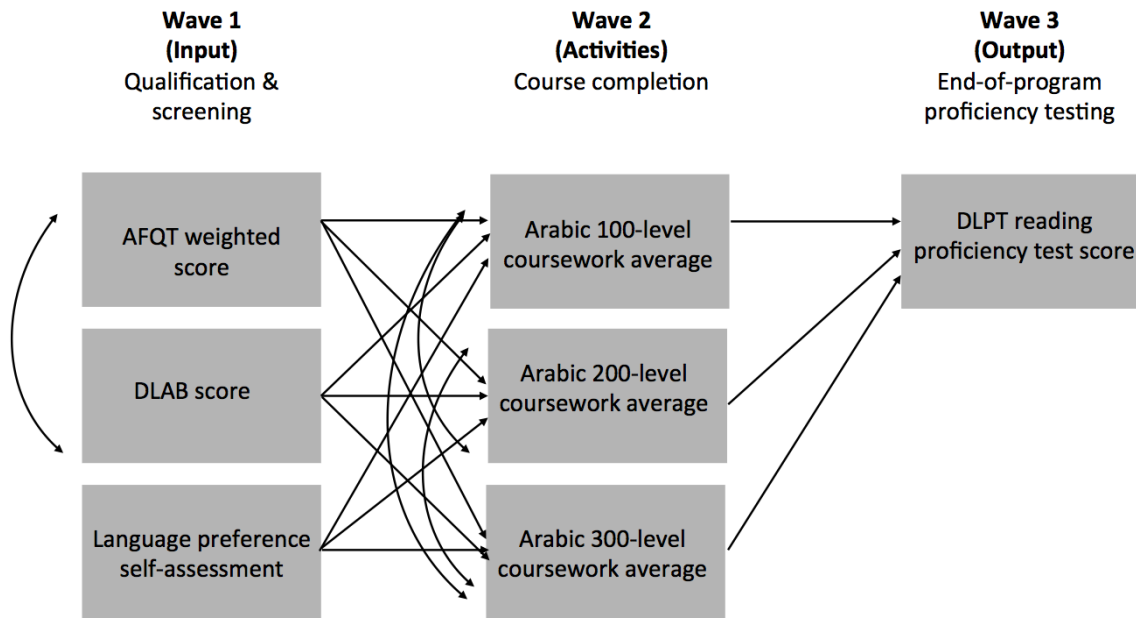


Figure 54. Example Wave 2 to Wave 3 Partial Path Model

As can be gleaned from Figure 54 above, a single-headed arrow can be observed from each 100-, 200-, and 300-level course outcome variable to DLPT reading proficiency test score outcomes (used as an example), denoting a hypothesized causal relationship between each observed Wave 2 variable and end-of program proficiency test score outcomes as well as the modeled concurrent covariation between each of the three endogenous course achievement variables.

Lastly, given the brief amount of time in which DLIFLC learners are enrolled in intensive language training (just 64 weeks), and building upon previous findings by Masters (2016), the author postulated that Wave 1 qualification and screening variables could potentially have a direct effect on Wave 3 end-of-program proficiency test score outcomes. This hypothesized relationship is depicted in Figure 55, which represents the full path diagram.

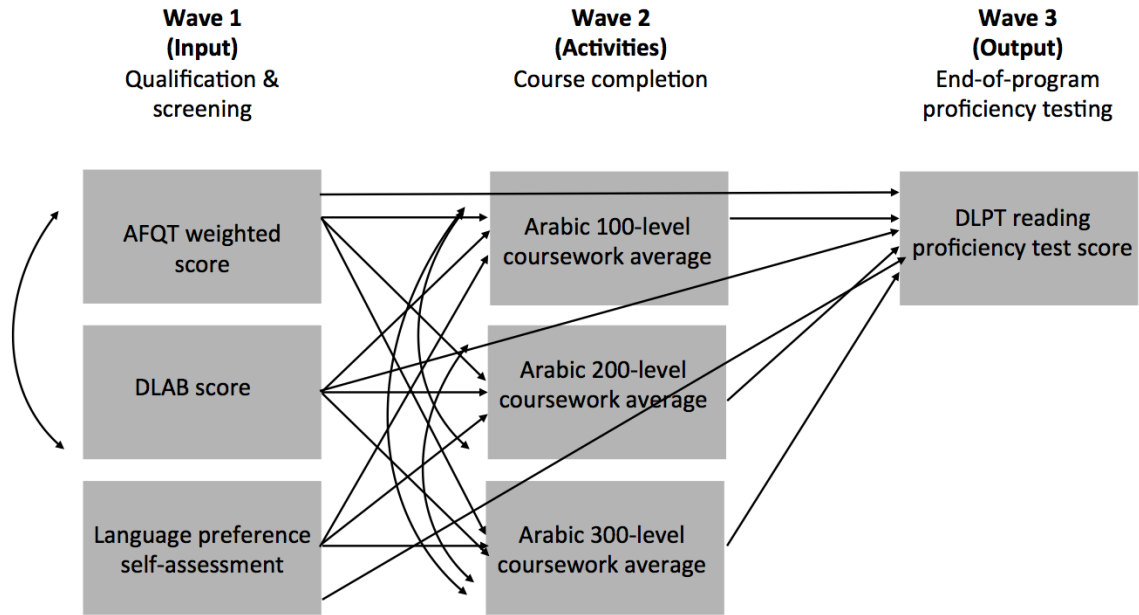


Figure 55. Complete path-analytic diagram

As shown above in Figure 55, AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores are not only hypothesized to indirectly influence DLPT and OPI proficiency test score outcomes through 100-, 200-, and 300-level average course proficiency outcomes, but are also hypothesized to have a direct causal effect on end-of-program proficiency test score outcomes, as denoted by the single-headed arrows. Figure 55 above represents the complete path-analytic diagram, serving as the foundation for the structural equations to be computed for Study 1.³⁸

Recommendations regarding sample for size requirements for running path analysis within an SEM framework vary widely in the literature. In their investigation of sample size recommendations through a series of simulations, Wolf et al. (2015) found the standard rule of thumb of ten cases per variable to be acceptable provided the model

³⁸ Not pictured in Figure 55 but modeled in the path-analytic estimations, are the associated error terms for each of the four endogenous variables (100-level coursework average, 200-level coursework average, 300-level coursework average, and the end-of-program proficiency test score outcomes (DLPT reading, DLPT listening, and OPI speaking)). Also of note regarding model specification is that exogenous variables are not assigned error terms.

outcomes do not produce wide variation in the observed magnitude of the standardized regression weights within the model. Very strong or very weak relationships may require larger sample sizes (p. 11). Following this recommendation, the path-analytic models discussed above require a minimum sample size of 70 cases since there are seven observed variables in each model. The Arabic, Chinese, and Korean languages meet these requirements ($n = 241, 98, \text{ and } 75$, respectively). As can be gleaned by Figure 55, the complete path model contains three exogenous variables (AFQT weighted score, DLAB score, and Language Preference Self-Assessment score) and four endogenous variables (100-level coursework average, 200-level coursework average, 300-level coursework average, and DLPT reading, DLPT listening, or OPI speaking scores, depending on the language skill for which the model is run) for a total of seven measured variables.

After a path-analytic model is specified, it is important to determine whether or not a given model is “identified.” A model is identified if a unique numerical solution can be estimated for each of the parameters in the model, which is necessary to produce an estimated population covariance matrix (Hancock, 2011; Tabachnick & Fidell, 2007, p. 709). To identify a model, the total number of observations within the model must be determined. This is calculated by multiplying the total number of measured variables in the model by the number of measured variables in the model plus one. This number is then divided by two. For the model specified in Figure 55 the number of observations is equal to the following:

$$7 (\text{total number of measured variables}) \times 8 (\text{total number of measured variables plus one}) / 2 = 28$$

Equation 1. Total number of path-analytic observations

Equation 1 reveals that the total number of observations for our hypothesized path diagram equals 28.

The next step is to calculate the total number of model parameters. Model parameters are determined by counting each exogenous variable, each single- or double-headed arrow in the path diagram, and each error term associated with the endogenous variables in the model. Referencing **Figure 55**, the current model contains 15 single-headed arrows, four double-headed arrows, four error terms, and three exogenous variables, equaling a total of 26 parameters. The number of model parameters is then subtracted from the total number of model observations. This is equivalent to the following equation:

$$28 (\text{total number of observations}) - 26 (\text{total number of parameters}) = 2$$

Equation 2. Degrees of freedom associated with path-analytic model

The outcome from Equation 2 represents the number of degrees of freedom (df) associated with the path model. The model is thus over-identified because the degrees of freedom are greater than zero, allowing for the planned statistical analyses to take place.³⁹

The primary goal of the path-analytic procedure within SEM is to test statistically the goodness-of-fit between a hypothesized relational model and a sample dataset (Byrne, 2001). Since the purpose of the current investigation is to explore empirically the a priori

³⁹ Other possible outcomes include “just identified” models (df = 0) or models with negative degrees of freedom. Negative degrees of freedom indicate that the model will not function.

hypothesized causal relationship between Wave 1 and Wave 2 variables on end-of-course proficiency outcomes, a model containing all hypothesized causal relationships and covariations is run. Output from the analyses is then analyzed in terms of the extent to which model estimates meet established model fit indices. The three that will be referenced in the current investigation include: (1) Chi Square Statistics (p value should be greater than 0.05), (2) Confirmatory Fit Index (CFI should be less than 0.90), and (3) Root Mean Square Error of Approximation (RMSEA should be less than 0.06).⁴⁰ The output from these indices compare the hypothesized model to either a computer-generated fully saturated model or an independent model, depending on the statistic used. Byrne (2001) suggests conceptualizing each of the three models (the hypothesized model, the fully saturated model, and the independent model) as points along a continuum, “with the independent model at one extreme, the saturated model at the other extreme, and the hypothesized model somewhere in between” (p. 79).⁴¹ Models that meet the established criterion indicate that there are minimal differences between the hypothesized model and fully saturated or independent models. Models that do not meet the established criterion indicate that it will be necessary to modify the hypothesized relationships between variables in order to address previously unaccounted for relationships.⁴² The entire analytical process will be conducted separately for the listening, reading, and speaking

⁴⁰ For a detailed review of the suite of goodness-of-fit statistics available when determining the adequacy of model specification, see Byrne (2001), pp. 79-88.

⁴¹ The independent model assumes complete independence of all variables in the model (in that correlations among variables are hypothesized to be zero). The saturated model reflects a “just identified” model (see footnote 36, above) in which the number of estimated parameters is equivalent to the number of data points (Byrne, 2001, p. 79).

⁴² Byrne (2001), in her discussion of using fit indices when judging model utility, cautions that, “exclusive reliance on model fit indices is unacceptable. Fit indices yield information bearing only on the model’s *lack of fit*. More importantly, they can in no way reflect the extent to which the model is plausible; this judgment rests squarely on the shoulders of the researcher” (p. 88). This guidance will be considered when interpreting established path-analytic outcomes.

modalities for the Arabic, Chinese, and Korean languages, resulting in nine separate analyses. This approach thus allows for the simultaneous estimation of both indirect and direct causal influences between upstream and downstream variables.

Study 1 was run on Analysis of Moment (AMOS) software associated with the SPSS statistical package. To perform the path analyses for Study 1, as mentioned above, only observed longitudinal records were extracted for each language. The records contained within the observed datasets represented 70% of all data for the Arabic language (241 observed cases of 345 total cases), 61% of all data for the Chinese language (98 observed cases of 161 total cases), and 64% of all data for the Korean language (75 observed cases of 118 total cases). As mentioned previously, path diagrams, as shown in Figure 55, are visual representations of the structural equations underlying each hypothesized theoretical relationship between measured variables (Hancock, 2011). As shown in Figure 55, there are four structural equations specified in the current model, represented by each of the four endogenous variables. Each endogenous variable is expressed as a function of all elements having a direct structural effect or covariation with it. The structural equations associated with each of the endogenous variables are outlined in Table 20.

Table 20. Structural equations for each endogenous variable in model

Variable	Structural Equation
Arabic 100	= constant + $\beta(\text{AFQT_Weighted}) + \beta(\text{DLAB}) + \beta(\text{AFQT*DLAB}) + \beta(\text{LangPref}) + \text{error}$
Arabic 200	= constant + $\beta(\text{AFQT_Weighted}) + \beta(\text{DLAB}) + \beta(\text{AFQT*DLAB}) + \beta(\text{LangPref}) + \text{error}$
Arabic 300	= constant + $\beta(\text{AFQT_Weighted}) + \beta(\text{DLAB}) + \beta(\text{AFQT*DLAB}) + \beta(\text{LangPref}) + \text{error}$
DLPT Listening	= constant + $\beta(\text{AFQT_Weighted}) + \beta(\text{DLAB}) + \beta(\text{AFQT*DLAB}) + \beta(\text{LangPref}) + \beta(\text{AD100}) + \beta(\text{AD200}) + \beta(\text{AD300}) + \beta(\text{AD100*AD200}) + \beta(\text{AD100*AD300}) + \beta(\text{AD200*AD300}) + \text{error}$
DLPT Reading	= constant + $\beta(\text{AFQT_Weighted}) + \beta(\text{DLAB}) + \beta(\text{AFQT*DLAB}) + \beta(\text{LangPref}) + \beta(\text{AD100}) + \beta(\text{AD200}) + \beta(\text{AD300}) + \beta(\text{AD100*AD200}) + \beta(\text{AD100*AD300}) + \beta(\text{AD200*AD300}) + \text{error}$
OPI Speaking	= constant + $\beta(\text{AFQT_Weighted}) + \beta(\text{DLAB}) + \beta(\text{AFQT*DLAB}) + \beta(\text{LangPref}) + \beta(\text{AD100}) + \beta(\text{AD200}) + \beta(\text{AD300}) + \beta(\text{AD100*AD200}) + \beta(\text{AD100*AD300}) + \beta(\text{AD200*AD300}) + \text{error}$

There is no modeled covariation between the error terms (also referred to as residuals) associated with each of the endogenous variables.

Study 1 predicted results: Wave 1 to Wave 2

If the LDC system is accurate, and languages grouped within the same category display coherence, the number and magnitude of the observed path coefficients will be comparable across languages. A number of hypothesized relationships are predicted to be stronger between waves and/or languages. Beginning with Wave 1 (depicted in Figure 55), for all three languages included in the subset of Category IV language analyses, one would anticipate the aptitude-related variables (i.e., the AFQT and DLAB variables) to most strongly predict initial course achievement outcomes, given their use as screening and selection tools for attendance at the DLIFLC and for language category assignment once enrolled. Consistent with what was found in Masters (2016) and Lett (1990), of the AFQT and DLAB variables, it is anticipated that the DLAB variable will most strongly predict outcomes in initial (100-level) coursework success, given its use as a selection tool for language difficulty category assignments, but will wane over time as learners

progress through their program of study. Consistent with previous research by Masters (2016), the Language Preference Self-Assessment variable is expected to predict learners' coursework success throughout their studies, but to predict 100-level average coursework outcomes most strongly given that other, highly contextual variables, likely influence learners' academic progress as they progress through coursework. The author makes these predictions with caution since the role of language preference and motivation have been found to be highly variable across learners as they progress through coursework. Although the Language Preference Self-Assessment variable is expected to predict all 100-, 200-, and 300-level average course outcomes, its influence will likely be confounded by other highly contextual variables not accounted for in the current model (such as likelihood of deployment into war zones, acculturation into military culture, curricular differences between languages, classroom language-learning climates, or the physical training demands also made at the DLIFLC), potentially impacting the strength of its predictive power.

Referencing the results of the one-way ANOVAs conducted between languages, for the AFQT predictor variable, no significant differences are expected to be found between languages. For the DLAB predictor variable, observed path coefficients are expected to be highest for the Chinese language and comparable for the Arabic and Korean languages. Lastly, for the Language Preference Self-Assessment variables, path coefficients are expected to be highest for the Chinese language, followed by the Arabic and Korean languages.

Study 1 predicted results: Wave 2 to Wave 3

Moving to Wave 2 within the path diagram, while all achievement-related course averages are likely to predict the development of proficiency, consistent with Masters (2016), it is hypothesized that only 300-level course achievement variables will significantly predict DLPT end-of-program outcomes for all languages, given that final coursework is likely designed to prepare students to successfully meet established DLPT and OPI proficiency standards. That is, only statistically significant path weights are expected to be found from 300-level average course outcomes to DLPT reading, DLPT listening, and OPI speaking proficiency test score outcomes. Referencing the results of the one-way ANOVAs conducted between languages, for the 300-level comparisons across language groups, established path coefficients are predicted to be highest for the Korean, then Chinese, and then Arabic languages, since higher 300-level average coursework outcomes are expected to predict higher end-of program proficiency test score outcomes. Outcomes from these analyses play a particularly important role in determining the coherence of the development of proficiency for languages grouped within the same category. By grouping the Chinese, Arabic, and Korean languages within the same category, DoD policy makers make an implicit assumption that patterns in initial acquisition proficiency development should be comparable across languages, which would serve as validity evidence to the categorization scheme. Non-comparable patterns in proficiency development might suggest a re-examination of the language categorization framework both within and across languages.

Study 1 predicted results: Wave 1 to Wave 3

Moving to the final relationships to be examined within the path model, although none of the AFQT, DLAB, or Language Preference Self-Assessment variables were

designed to predict end-of-program proficiency outcomes, Masters (2016) found significant causal pathways between the AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores and DLPT/OPI outcomes. The strongest relationships were established between DLAB outcomes and DLPT reading scores, suggesting that language-specific aptitude plays an important role, not only in predicting initial success in the development of proficiency, but also in predicting end-of-program outcomes. Given that the current study is working with a similar dataset, positive, significant causal pathways are expected to be found between the aptitude-related variables (i.e., AFQT and DLAB) and end-of-program proficiency testing. The magnitude of the path coefficients is hypothesized to be larger for the language-specific (e.g., DLAB scores) than for learners' general aptitude scores (i.e., AFQT weighted scores). AFQT weighted scores and DLAB scores are hypothesized to predict DLPT reading and listening outcomes equally across the Arabic, Chinese, and Korean languages. No significant differences in predictive power are expected across languages. Of note is that the prediction is only made for the reading and listening test score outcomes, since both instruments entail a great deal of reading comprehension and listening comprehension skills. No significant causal pathways are hypothesized between the AFQT and DLAB test score outcomes and OPI speaking proficiency test score outcomes. Consistent with Masters (2016), the Language Preference Self-Assessment variable, although not designed to predict end-of-program outcomes, was found to significantly predict Arabic listening end-of-program outcomes. It is therefore included as a significant predictor of both Wave 2 achievement variables and Wave 3 end-of-

program proficiency outcomes It is expected to predict Wave 2 achievement variables more strongly than Wave 3 end-of-program outcomes.

Study 1 Results

Research Question 1 (RQ1): For languages grouped within the same category, are similar patterns of individual differences in general aptitude, language-specific aptitude, and motivation observed in the prediction of learners' success as they progress through coursework?

Study 1 Results: Wave 1 to Wave 2 Path Analyses: Reading, Listening, and Speaking

The current section discusses the results of the Wave 1 to Wave 2 path analyses for all three skills, given that 100-, 200-, and 300-level average course outcomes were not broken down by skill. Across the Arabic, Chinese, and Korean languages, no statistically significant path coefficients were found from the AFQT scores to any of the Wave 2 100-, 200-, and 300-level average course outcomes. This finding differs from what the author hypothesized since it was posited that the AFQT scores would significantly predict average 100-level outcomes given its use as an initial screening tool for candidacy at the DLIFLC. As expected, DLAB scores were found to consistently predict 100-, 200-, and 300-level average coursework outcomes across all languages. Lastly, although the Language Preference Self-Assessment variable was hypothesized to predict 100-, 200-, and 300-level course outcomes, results of the path analyses found just one significant causal pathway. For the Arabic language only, a significant negative causal pathway was found between Language Preference Self-Assessment and 100-level average coursework outcomes. Referencing the correlation coefficients between Language Preference Self-Assessment scores and 100, 200-, and 300-level course outcomes, this finding suggests that higher Language Preference Self-Assessment scores are associated with lower

average coursework averages (see Appendix D to review the correlation matrices for each variable within the path-analytic model). It is important to note that this variable only represents learners' responses to self-reported survey data asking if he or she was assigned to their first language choice, which may not adequately distinguish the role that language preference may play as learners progress through DLIFLC coursework. Figure 56 through Figure 58 below visually depict the results of these analyses.⁴³

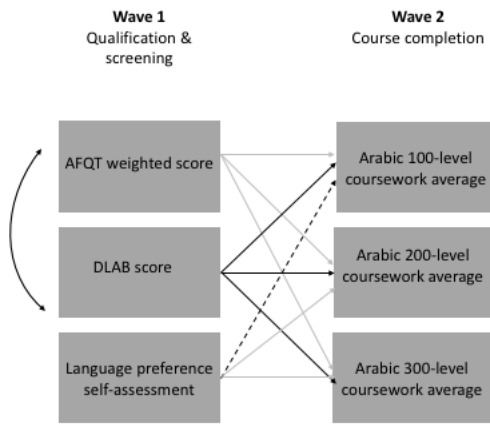


Figure 56. Wave 1 to Wave 2 Model Arabic

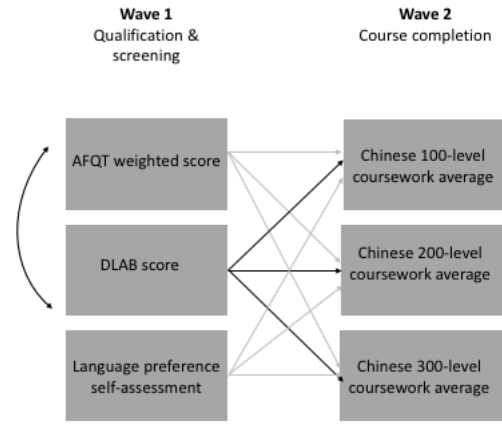


Figure 57. Wave 1 to Wave 2 Model Chinese

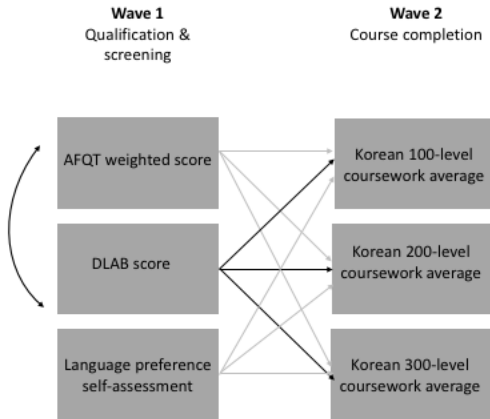


Figure 58. Wave 1 to Wave 2 Model Korean

⁴³ For all path-analytic figures, dark lines represent significant causal pathways between measured variables.. Grey lines indicate modeled, but non-significant causal pathways between measured variables. Dotted lines represent a significant, negative relationship between measured variables.

As can be gleaned from Figure 56 through Figure 58 above, the DLAB score serves as the only Wave 1 variable to play a robust role in the prediction of DLIFLC average coursework outcomes, thus providing initial Wave 1 to Wave 2 validity evidence for use of this instrument as a screening tool for the LDC framework. To assist with the interpretation of the relative magnitude of each of the causal pathways depicted in Figure 56 through Figure 58 above, Table 21 through Table 23 display the standardized regression weight, standard error, and significance level for each of the significant pathways in the above model.

Table 21. Arabic Wave 1 to Wave 2 path-analytic outcomes (n =241)

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to 100-level average coursework	0.322	0.020	p < 0.001
DLAB to 200-level average coursework	0.311	0.021	p < 0.001
DLAB to 300-level average coursework	0.288	0.022	p < 0.001
Language Preference Self-Assessment to 100-level average coursework	-0.121	0.026	p < 0.05

Table 22. Chinese Wave 1 to Wave 2 path-analytic outcomes (n = 98)

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to 100-level average coursework	0.338	0.031	p < 0.01
DLAB to 200-level average coursework	0.233	0.028	p < 0.05
DLAB to 300-level average coursework	0.282	0.025	p < 0.01

Table 23. Korean Wave 1 to Wave 2 path-analytic outcomes (n = 75)

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to 100-level average coursework	0.378	0.034	p < 0.001
DLAB to 200-level average coursework	0.350	0.033	p < 0.01
DLAB to 300-level average coursework	0.399	0.031	p < 0.001

As shown above, the DLAB score variable plays a consistent role in predicting 100-, 200-, and 300-level average course outcomes; significant causal pathways were found across all three languages. This finding provides validity evidence for use of the DLAB as a screening tool for predicting initial DLIFLC coursework success.⁴⁴ Direct effects between Wave 1 and Wave 2 predictor variables, such as those observed in Figure 56 through Figure 58 above, are equivalent to correlation coefficients. The established path coefficients between languages can be compared by transforming the observed coefficients to a z value. The outcome of these analyses are reported in Table 24, below.

Table 24. z-transformation of DLAB to 100-, 200-, and 300-level course outcomes

Path Comparison	z-value (Reference Value +/- 1.96)	Significant Difference Between Languages?
Arabic and Chinese DLAB to 100-level average coursework	0.148	Not significant
Arabic and Chinese DLAB to 200-level average coursework	0.694	Not significant
Arabic and Chinese DLAB to 300-level average coursework	0.053	Not significant
Arabic and Korean DLAB to 100-level average coursework	0.474	Not significant
Arabic and Korean DLAB to 200-level average coursework	0.365	Not significant
Arabic and Korean DLAB to 300-level average coursework	0.973	Not significant
Chinese and Korean DLAB to 100-level average coursework	0.293	Not significant
Chinese and Korean DLAB to 200-level average coursework	0.819	Not significant
Chinese and Korean DLAB to 300-level average coursework	0.846	Not significant

As shown in **Table 24**, no statistically significant differences across languages were found for the correlations between the DLAB and 100-, 200-, and 300-level average coursework outcome variables, contrary to what was hypothesized by the author based on

⁴⁴ The equation is: $z = \frac{z_{a_1} - z_{a_2}}{\sqrt{1/(N_a - 3) + 1/(N_b - 3)}}$. Values outside of the critical value +/- 1.96 indicate a statistically significant difference in observed coefficients.

the results of the one-way ANOVAs conducted between languages. This indicates that the DLAB variable robustly predicts DLIFLC coursework success for all three Category IV languages.

Research Question 2 (RQ2): For languages grouped within the same category, are similar patterns of individual differences in general aptitude, language-specific aptitude, and motivation observed in the prediction of learners' end-of-program outcomes?

Study 1 Results: Wave 2 to Wave 3 Path Analyses: Reading

The current section will discuss the path-analytic outcomes of Wave 2 exogenous predictor variables to Wave 3 endogenous outcome variables separately by skill.⁴⁵

Beginning with the predictive influence of average 100-, 200-, and 300-level coursework on DLPT reading outcomes, as predicted, average 100-level course outcomes did not significantly predict DLPT reading outcomes. Unexpectedly, for the Arabic language, 200-level average course outcomes predicted DLPT reading outcomes, indicating an alignment between 200-level course grades and end-of program proficiency test scores. As predicted, across all languages, 300-level average course outcomes predicted end-of-program DLPT reading test outcomes. Figures 59 through 61 visually depict the results of these analyses.

⁴⁵ Unlike the examination of Wave 1 to Wave 2 causal pathways, Wave 3 outcome variables differ by the listening, reading, and speaking skills and therefore will be discussed separately.

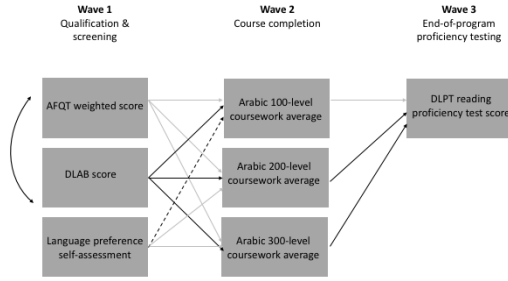


Figure 59. Wave 2 to 3 Arabic Reading

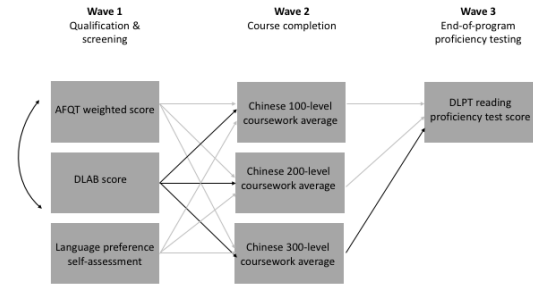


Figure 60. Wave 2 to 3 Chinese Reading

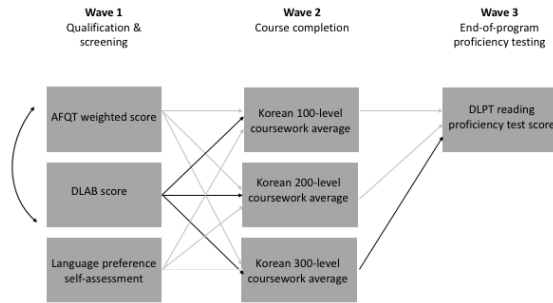


Figure 61. Wave 2 to 3 Korean Reading

As can be seen in Figure 59 through Figure 61, 300-level average coursework consistently predicts Arabic, Chinese, and Korean DLPT reading proficiency test score outcomes. This finding aligns with the author’s prediction, in that students in their end-of-program training are likely preparing for proficiency testing. To assist with the interpretation of the figures above, Tables 25 through 27 display the associated standard error, significance level, and standardized regression weights for each of the significant pathways in the above model.

Table 25. Arabic Reading: Wave 2 to Wave 3 path-analytic outcomes (n = 241)

Path	Standardized Regression Weight	Standard Error	Significance Level
200-level average coursework to DLPT Reading	0.255	4.130	p < 0.05
300-level average coursework to DLPT Reading	0.314	3.389	p < 0.001

Table 26. Chinese Reading: Wave 2 to Wave 3 path-analytic outcomes (n = 98)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Reading	0.527	4.919	p < 0.001

Table 27. Korean Reading: Wave 2 to Wave 3 path-analytic outcomes (n = 75)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Reading	0.392	4.916	p < 0.05

As shown in Table 25 through

Table 27 above, for the Arabic language, the magnitude of the standardized regression weight associated with 200- and 300-level average course outcomes and DLPT reading scores increases as learners progress through coursework (from 0.255 to 0.314). For 300-level coursework outcomes, the largest standardized regression weights were found for the Chinese (path coefficient = 0.527), then Korean (path coefficient = 0.392), then Arabic languages (path coefficient = 0.314).

Because Wave 2 to Wave 3 outcomes involve both direct and indirect effects, more sophisticated modeling methods are required to test for statistically significant differences in standardized path weights between models. Multi-group invariance testing within the AMOS statistical package can be used to determine whether the established standardized regression weights for each causal pathway are invariant across groups. As stated by Byrne (2001), “the pattern of factor loadings for each observed measure is tested for its equivalence across the groups. Once it is known which measures are group-invariant, these parameters are constrained equal while subsequent tests of the structural parameters are conducted” (p. 175). Multi-group invariance testing yields pairwise

parameter comparison in matrix form across all three language groups. Critical ratios for differences are presented in the form of a z test statistic, indicating the difference in established path coefficients between models. Similar to the testing for differences in observed correlations with use of a z transformation, absolute values outside of 1.96 are indicative of statistically significant differences between parameters of interest with p set at < 0.05 . Table 28 below numerically details the outcomes of the multi-group invariance testing between the standardized regression weights for 300-level average coursework outcomes to DLPT reading scores across all languages.

Table 28. Multi-Group Invariance Testing: Wave 2 to Wave 3: Reading

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Arabic and Chinese 300-level average coursework to DLPT Reading	-1.024	Not significant
Arabic and Korean 300-level average coursework to DLPT Reading	-0.314	Not significant
Korean and Chinese 300-level average coursework to DLPT Reading	0.611	Not significant

As shown above, outcomes from multi-group invariance testing yielded non-significant differences in the established regression weights across groups, indicating that 300-level average course outcomes play a consistent role in predicting DLPT reading outcomes for each of the Arabic, Chinese, and Korean languages. Outcomes from the multi-group invariance testing differ from what the author hypothesized based on the results of the one-way ANOVAs conducted between languages in which the path-analytic outcomes for the Korean language were predicted to be significantly stronger. However, this finding provides additional evidence concerning the coherence of observed initial acquisition patterns for languages grouped within the same category. In line with the role of the Wave 1 to Wave 2 DLAB variable, outcomes from the Wave 2 to Wave 3 path

analysis provide evidence that the 300-level average plays a comparable role in the development of proficiency across languages.

Study 1 Results: Wave 2 to Wave 3 Path Analyses: Listening

As predicted, average 100-level and 200-level course outcomes did not significantly predict DLPT listening outcomes. Also as hypothesized, across all languages, 300-level average course outcomes predicted end-of-program DLPT listening test outcomes. With the exception of a significant causal pathway between 200-level average coursework outcomes and DLPT test score outcomes for the Arabic language (which were found for reading but not for listening), the listening proficiency outcomes are identical to those found for the reading proficiency outcomes.⁴⁶ This suggests a coherent pattern in foreign language listening proficiency development across the Category IV languages. Figure 62 through Figure 64 below visually depicts the results of these analyses.

⁴⁶ An exception to this finding is the loss of a significant causal pathway from the reading skill analyses, in which a significant causal pathway was found between 200-level average course outcomes and DLPT reading test score outcomes for the Arabic language.

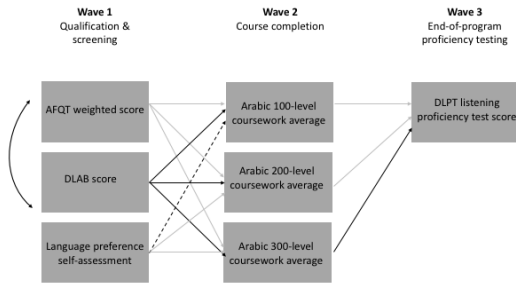


Figure 62. Wave 2 to 3 Arabic Listening

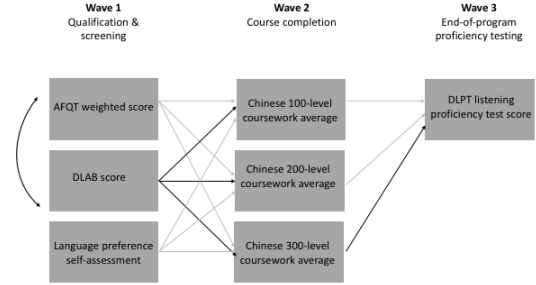


Figure 63. Wave 2 to 3 Chinese Listening

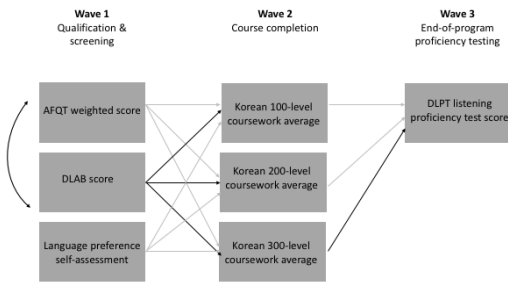


Figure 64. Wave 2 to 3 Korean Listening

To assist with the interpretation of the figures above, Table 29 through Table 31 display the associated standard error, significance level, and standardized regression weights for each of the significant pathways in the above model.

Table 29. Arabic Listening: Wave 2 to Wave 3 path-analytic outcomes (n = 241)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Listening	0.334	3.262	p < 0.001

Table 30. Chinese Listening: Wave 2 to Wave 3 path-analytic outcomes (n = 98)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Listening	0.599	6.543	p < 0.001

Table 31. Korean Listening: Wave 2 to Wave 3 path-analytic outcomes (n = 75)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Listening	0.599	6.543	p < 0.001

As shown in

Table 32, and consistent with the findings discussed above for the reading skill, 300-level average coursework consistently predicts Arabic, Chinese, and Korean DLPT listening proficiency test score outcomes, suggesting an alignment between 300-level DLIFLC coursework and the DLPT listening test. As hypothesized by the author based on the results of the one-way ANOVAs calculated between languages, outcomes from multi-group invariance testing found a statistically significant difference in the standardized regression weights between the Arabic and Korean language models ($z = 2.077$). No significant differences were found between the Arabic and Chinese path coefficients.

Table 32. Multi-Group Invariance Testing: Wave 2 to Wave 3: Listening

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Arabic and Chinese 300-level average coursework to DLPT Listening	-1.759	Approaches significance
Arabic and Korean 300-level average coursework to DLPT Listening	-2.077	Significant
Korean and Chinese 300-level average coursework to DLPT Listening	-0.096	Not significant

As shown in

Table 32, the z-value of -2.077, established from the multi-group invariance testing between the Arabic and Korean languages, falls outside the reference value of +/- 1.96. Of note is that the z-value of -1.759 between the Arabic and Chinese languages also approaches significance. The significant difference established between the Arabic and Korean 300-level average coursework to DLPT reading outcomes indicates that, while both causal pathways predict DLPT reading outcomes, the prediction is significantly stronger for the Korean language than the Arabic language.

Study 1 Results: Wave 2 to Wave 3 Path Analyses: Speaking

Similar to the findings for the reading and speaking skills, no significant causal pathways were found between 100-level average coursework outcomes and OPI speaking proficiency test scores for the Arabic, Chinese, and Korean languages. Unlike the findings for the reading and listening skills, just two significant causal pathways were found. For the Arabic language, as predicted, 300-level average coursework outcomes were found to predict OPI speaking test scores. For the Chinese language, 200-level average coursework outcomes were found to predict OPI speaking test scores. Surprisingly, no significant causal pathways were found between average 100-, 200-, and 300-level coursework outcomes and the OPI speaking test scores for the Korean language. Figure 65 through Figure 67 below visually depict the results of these analyses.

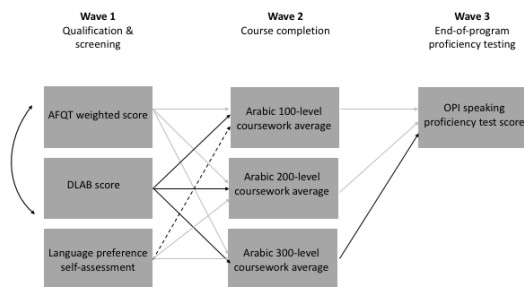


Figure 65. Wave 2 to 3 Arabic Speaking

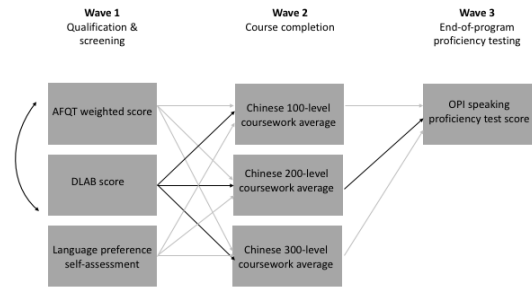


Figure 66. Wave 2 to 3 Chinese Speaking

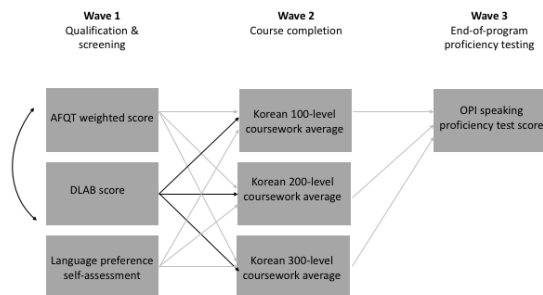


Figure 67. Wave 2 to 3 Korean Speaking

As can be gleaned from Figure 65 through Figure 67 above, inconsistent patterns are observed in the development of speaking proficiency for the Arabic, Chinese, and Korean languages. Table 33 and Table 34 below detail these results.

Table 33. Arabic Speaking: Wave 2 to Wave 3 path-analytic outcomes (n = 241)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to OPI Speaking	0.241	1.470	p < 0.05

Table 34. Chinese Speaking: Wave 2 to Wave 3 path-analytic outcomes (n = 98)

Path	Standardized Regression Weight	Standard Error	Significance Level
200-level average coursework to OPI Speaking	0.324	3.088	p < 0.05

As shown in Table 33 and Table 34 above, outcomes from the speaking skill analyses were not in line with what was hypothesized by the author. No significant causal pathways were established for the Korean language, while 300-level average course outcomes predicted Arabic OPI outcomes and 200-level average course outcomes predicted Chinese OPI outcomes. This finding suggests that the development of speaking proficiency across the Arabic, Chinese, and Korean Category IV languages differs from that of reading and listening skill development, which yielded similar patterns of development across languages. As mentioned above, it also may suggest instability in the measurement of the speaking skill itself, since the OPI is a human-rated performance-based assessment administered by DLIFLC language teachers or curriculum developers who are trained as OPI examiners. Potential 300-level instructional differences for the Korean language group may influence these findings as well. Of note is that virtually 100% of learners enrolled in the Korean program of study met OPI criterion test score

outcomes. The fact that none of the average 100-, 200-, and 300-level coursework outcomes predict OPI speaking proficiency outcomes for the Korean language, even though all learners meet the established criterion, is particularly counterintuitive, suggesting that there are unaccounted for dimensions external to the path-analytic model that likely contribute to the development of Korean speaking proficiency.

Study 1 Results: Wave 1 to Wave 3 Full Path Analyses: Reading

The current section will discuss the path-analytic outcomes of Wave 1 exogenous predictor variables to Wave 3 endogenous outcome variables separately by skill. The section will also compare the results of the full Arabic, Chinese, and Korean models across skill modalities. Beginning with the predictive influence of Wave 1 AFQT weighted scores, DLAB scores, and learners' self-reported motivation scores on DLPT reading outcomes, just three significant causal pathways were found. Consistent with what the author hypothesized, learners' AFQT weighted scores predicted end-of-program reading proficiency test score outcomes for both the Chinese and Korean languages. Figure 68 through Figure 70 visually depict these findings.

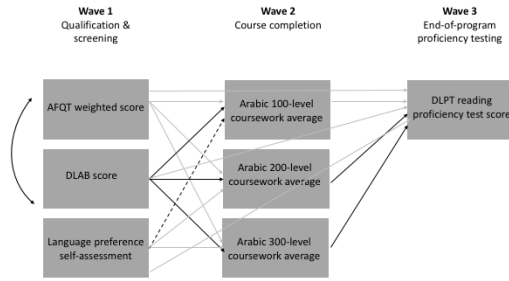


Figure 68. Full Arabic Path Model Reading (n = 241, CFI = 0.999, RMSEA = 0.034, $\chi^2 = 2.71$, p = 0.26)

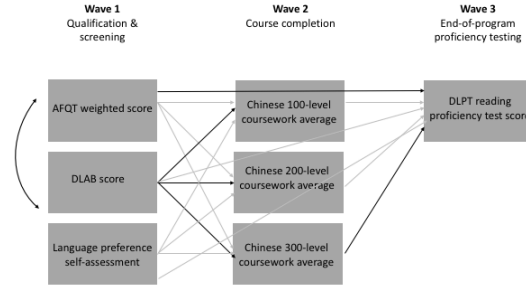


Figure 69. Full Chinese Path Model Reading (n = 98, CFI = 1.000, RMSEA = 0.000, $\chi^2 = 0.144$, p = 0.87)

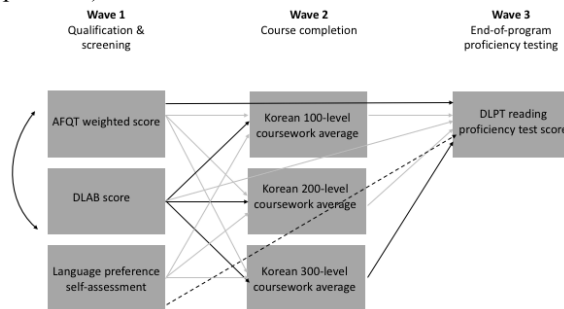


Figure 70. Full Korean Path Model Reading (n = 75, CFI = 1.000, RMSEA = 0.000, $\chi^2 = 1.55$, p = 0.46)

Acceptable model fit was found for all languages across the all skills.^{47 48} As visually depicted in Figure 68 through Figure 70, none of the Wave 1 variables were found to predict Arabic DLPT reading proficiency test scores, while AFQT scores for both the Chinese and Korean languages were found to significantly predict DLPT reading proficiency outcomes. This finding suggests that the Verbal Expression, Mathematical

⁴⁷ Identical model fit was found for the full path model across all language skills for each language. With the exception of substituting end-of-program proficiency test score outcomes with DLPT reading, DLPT listening, or OPI speaking outcomes, identical models were run for each language and skill. The variability associated with each outcome measure within each language was likely not significant enough to affect model fit across language modality.

⁴⁸ Of note is the perfect model fit found for the CFI and RMSEA indices for the Chinese and Korean languages. As noted by Bentler and Chou, 1987, perfect model fit indices have been found to be potential indicators of model misspecification. Byrne (2001) warns, “indeed, fit indices provide no guarantee whatsoever that a model is useful. In fact, it is entirely possible for a model to fit well and yet still be incorrectly specified” (p. 86). Therefore, the path-analytic findings for the Chinese and Korean data should be interpreted with caution.

Knowledge, and Arithmetic Knowledge, which compose the AFQT weighted scores and likely require considerable reading comprehension skills, positively predict end-of-program DLPT outcomes. The absence of a predicted causal pathway from the AFQT weighted score to DLPT Arabic reading outcomes remains to be explored. Surprisingly, for the Korean language, a statistically significant negative causal pathway was found between the Language Preference Self-Assessment variable and DLPT reading proficiency test score outcomes, again suggesting the potential instability of the OPI speaking outcome variable for this language.

Also of relevance to the current section is the number of learners for each language who met DLPT reading criterion standards. For the Arabic language, 84.6% of learners met criterion or better; for the Chinese language, 96.9% of learners met criterion or better; and for the Korean language, 100% of learners met criterion or better. This finding also confirms the DLIFLC “re-cycling” and “re-languaging” policies across languages, whereby only the academically strongest candidates are allowed to progress through coursework and sit for end-of-program testing. Table 35 through

Table 36 below detail the results of outcomes visualized in Figure 68 through Figure 70 above.

Table 35. Chinese Reading: Wave 1 to Wave 3 path-analytic outcomes (n = 98)

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT to DLPT Reading	0.295	0.245	p < 0.001

Table 36. Korean Reading: Wave 1 to Wave 3 path-analytic outcomes (n = 75)

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT to DLPT Reading	0.209	0.259	p < 0.05

Language Preference Self-Assessment to DLPT Reading	-0.185	0.805	p < 0.05
---	--------	-------	----------

As hypothesized by the author, results of multi-group invariance testing, comparing the path coefficients between the Korean and Chinese AFQT to DLPT reading outcomes, yielded no statistically significant differences between models, as detailed in Table 37.

Table 37. Multi-Group Invariance Testing: Wave 1 to Wave 3 Reading

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Korean and Chinese AFQT weighted scores to DLPT Reading	0.742	Not significant

In addition to the goodness-of-fit statistics, another way of exploring how well a given model explains hypothesized relationships among variables is to examine the associated squared multiple correlation indices, which are calculated for each endogenous variable within the model. These statistics denote the amount of variability accounted for by a hypothesized model as well as the amount of variability that remains to be explained.

Table 38 below reports the squared multiple correlation statistics for the Arabic, Chinese, and Korean reading models.

Table 38. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean reading models

Language	Endogenous Variable	% of Variability Accounted for	% of Variability Remaining
Arabic	100-level course outcomes	14.8%	85.2%
	200-level course outcomes	9.2%	90.8%
	300-level course outcomes	9.8%	90.2%
	DLPT Reading (Full Model)	33.9%	66.1%
Chinese	100-level course outcomes	0.9%	99.1%
	200-level course outcomes	7.5%	92.5%
	300-level course outcomes	11.4%	88.6%
	DLPT Reading (Full Model)	32.3%	67.7%
Korean	100-level course outcomes	13.0%	87.0%
	200-level course outcomes	11.2%	88.8%
	300-level course outcomes	16.5%	83.5%
	DLPT Reading (Full Model)	35.8%	64.2%

As shown above, although squared multiple correlation indices are comparable across languages, the Korean reading model accounts for the largest amount of model variability (35.8%). Across all languages, the complete path-analytic model for the reading skill accounts for between 32.3% and 33.9% of total model variability. Although this also indicates that about 60% of model variability remains to be explained by other unaccounted for factors, the author argues that this finding is acceptable given that just six measured variables account for over 30% of total model variation.

Study 1 Results: Wave 1 to Wave 3 Full Path Analyses: Listening

For the listening skill, contrary to what was expected by the author, no significant causal pathways were found between any of the Wave 1 predictor variables and Wave 3 DLPT listening outcomes for the Arabic, Chinese, and Korean languages. Compared to the findings for the reading skill, this suggests that of the Wave 1 predictor variables, AFQT weighted score is potentially more well suited as a screening instrument for the reading skill than the listening skill. Figure 71 through Figure 73 below visually depict these results.

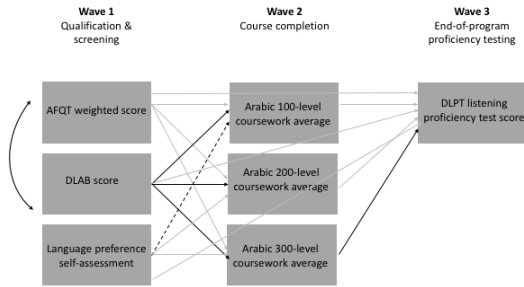


Figure 71. Full Arabic Path Model Listening

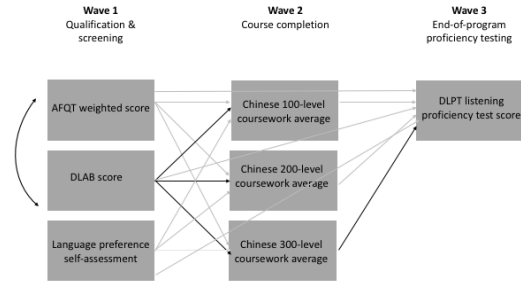


Figure 72. Full Chinese Path Model Listening

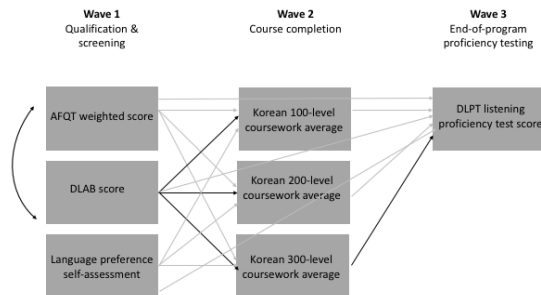


Figure 73. Full Korean Path Model Listening

For DLPT listening outcomes, 81.7% of Arabic learners met criterion scores or better, 90.8% of Chinese learners, and 88% of Korean learners met criterion scores or better. Across all languages, these percentages are lower than those found for the reading skill. Results of the comparison of squared multiple correlations across languages revealed

substantially lower squared multiple correlation indices for the Chinese language than for the Arabic and Korean languages.

Table 39. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean listening models

Language	Endogenous Variable	% of Variability Accounted for	% of Variability Remaining
Arabic	100-level course outcomes	14.8%	85.2%
	200-level course outcomes	9.2%	90.8%
	300-level course outcomes	9.8%	90.2%
	DLPT Listening (Full Model)	40.7%	59.3%
Chinese	100-level course outcomes	0.9%	99.1%
	200-level course outcomes	7.5%	92.5%
	300-level course outcomes	11.4%	88.6%
	DLPT Listening (Full Model)	29.0%	71.0%
Korean	100-level course outcomes	13.0%	87.0%
	200-level course outcomes	11.2%	88.8%
	300-level course outcomes	16.5%	83.5%
	DLPT Listening (Full Model)	40.6%	59.4%

As shown in

Table 39, the Arabic and Korean listening models accounted for about 41% of total model variation, with the Chinese model accounting for just 29%. This finding is surprising, particularly given that Chinese had the highest percentage of learners to meet DLPT criterion outcomes, suggesting that other unaccounted for programmatic variables are likely contributing to the acquisition of Chinese listening skills.

Study 1 Results: Wave 1 to Wave 3 Full Path Analyses: Speaking

Finally, for the speaking skill, contrary to what was hypothesized by the author, AFQT weighted scores and DLAB scores were found to significantly predict OPI proficiency test score outcomes for the Korean language, but in unexpected directions. Both learners' AFQT scores and self-assessed motivation scores were found to negatively predict OPI outcomes. Contrary to what was expected by the author, no statistically significant causal pathways were found for the Arabic and Chinese languages. These findings are detailed in Figure 74 through Figure 76.

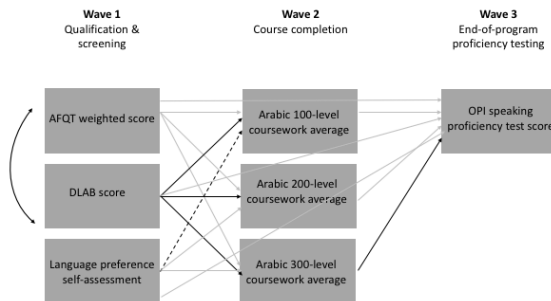


Figure 74. Full Arabic Path Model Speaking

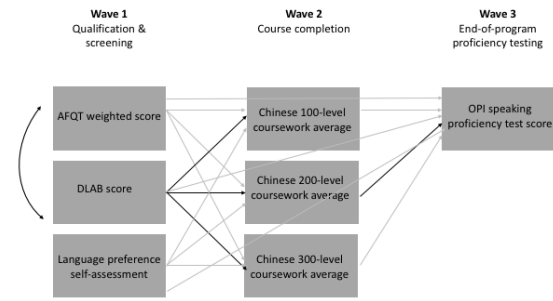


Figure 75. Full Chinese Path Model Speaking

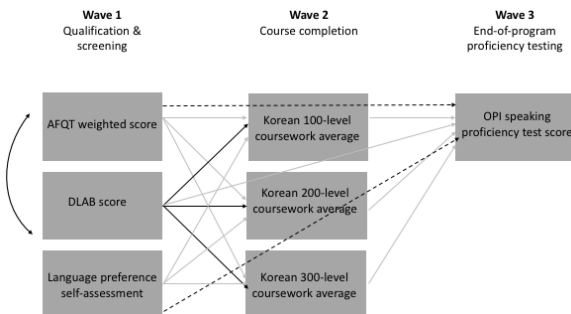


Figure 76. Full Korean Path Model Speaking

For OPI speaking outcomes, 98.3% of Arabic learners and 100% of both Chinese and Korean learners met OPI speaking criterion outcomes. The unexpected negative causal pathways between Korean learners' AFQT scores and OPI speaking scores and the fact that 100% of all Chinese and Korean learners met speaking criterion is surprising,

potentially indicative of the instability of speaking skill proficiency measurement. That is, although the author postulated that no relationships would be found between Wave 1 predictor variables and Wave 3 proficiency outcomes, the significant negative relationship was unanticipated. The path coefficients associated with the findings depicted in Figure 74 through Figure 76 above are detailed in Table 40 and Table 41 below.

Table 40. Korean Speaking: Wave 1 to Wave 3 path-analytic outcomes (n = 75)

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT Weighted score to OPI Speaking	-0.217	0.147	p < 0.05
Motivation to OPI Speaking	-0.236	0.457	p < 0.05

The negative causal relationship between Korean learners' speaking scores on the one hand, and AFQT scores and self-assessed motivation scores is a surprising finding that could potentially be attributable to instability in the administration of the OPI for Korean. Unexpected relationships such as these could also be attributable to range restriction. That is, range restriction is associated with both Wave 1 input variables, as well as Wave 3 output variables, in which 100% of learners met OPI speaking outcomes. This type of ceiling effect possibly attenuates or even distorts observed path analytic outcomes.

Table 41. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean speaking models

Language	Endogenous Variable	% of Variability Accounted for	% of Variability Remaining
Arabic	100-level course outcomes	14.8%	85.2%
	200-level course outcomes	9.2%	90.8%
	300-level course outcomes	9.8%	90.2%
	OPI Speaking (Full Model)	28.7%	71.3%
Chinese	100-level course outcomes	0.9%	99.1%

	200-level course outcomes	7.5%	92.5%
	300-level course outcomes	11.4%	88.6%
	OPI Speaking (Full Model)	32.3%	67.7%
Korean	100-level course outcomes	13.0%	87.0%
	200-level course outcomes	11.2%	88.8%
	300-level course outcomes	16.5%	83.5%
	OPI Speaking (Full Model)	28.1%	71.9%

As shown above, the Arabic and Korean languages account for comparable percentages of total model variability (~28% to 29%), while the Chinese language accounts for the largest percentage of total model variability (32.3%). This indicates that the hypothesized causal model is strongest for the Chinese language. Of note is that across language skills, the speaking model accounts for the least amount of total model variability compared to the reading and listening models, which were comparable. Despite the fact that 100% of Chinese and Korean learners met OPI speaking criterion outcomes, about 70% of total model variability remains to be accounted for within the path-analytic framework. This also suggests that there are other variables not related to aptitude, language preference, and curricular outcomes that remain to be identified and modeled.

Discussion: Study 1

Research Question 3 (RQ3): For languages grouped within the same category, are homogenous patterns of initial language acquisition patterns observed across languages?

Situating the main inputs, activities, and outcomes associated with the DLIFLC instructional paradigm within a logic model allowed for the empirical examination of the coherence and comparability of L2 initial acquisition patterns through use of a path analysis. In his discussion concerning the applicability of logic modeling to L2 acquisition, Norris (2016) states:

The potential of program logic models for advancing language education and related endeavors is tremendous, in that--if developed by educational experts, practitioners, and other insiders (i.e., versus external "logic model" experts)--they would lay bare the rationales, activities, and implicit theories that constitute language programs and thereby render them amenable to empirical confirmation (or rejection). (p. 178)

Overall, across all languages and skills, the Wave 1 DLAB variable was found to play a robust role in predicting 100-, 200-, and 300-level coursework outcomes, thus adding to the validity evidence for use of the DLAB as a selection tool at the DLIFLC. When breaking Study 1 findings down by waves, beginning with the left-most antecedent predictor variables, although the Wave 1 qualification and screening variables were postulated to potentially predict end-of-program outcomes for all languages, only a few causal pathways approached significance. For Chinese and Korean, AFQT scores predicted DLPT reading outcomes. Learners' general aptitude, measured by AFQT weighted scores, was found to significantly predict DLPT reading outcomes for both Chinese and Korean. This suggests that the heavily reading-based components of this measure (e.g., paragraph comprehension and word knowledge) perform well in predicting end-of-program reading proficiency outcomes. For the Korean language only, Language Preference Self-Assessment scores negatively predicted DLPT reading outcomes, while AFQT and Language Preference Self-Assessment scores negatively predicted OPI

speaking outcomes. This finding suggests potential incoherence in the acquisitional patterns of the Korean language given this unexpectedly negative causal relationship was not found for Arabic or Chinese.

The squared multiple correlation indices for the Korean speaking model was the lowest of all speaking models, accounting for just 28.1% of total model variance. Consistent with what has been recommended by Norris (2016), unexplained outcomes such as these may indicate that, for the Korean speaking skill, the logic model may not be adequately specified and that other unaccounted for external program factors may be mediating the observed path-analytic findings. Future models should account for additional contextual variables that could potentially explain Korean speaking proficiency development patterns, such as curricular content or heritage language status. Lastly, across almost all languages, 300-level average course outcomes were found to significantly predict DLPT reading and listening and OPI speaking outcomes.⁴⁹ This indicates an alignment between 300-level instructional content and end-of-training proficiency testing. It also may reflect an artifact of student placement practices at the DLIFLC. As noted previously, some learners within the DLIFLC are recommended by instructors for “re-linguaging”, or reassignment to a language in a lower difficulty category due to sub-par course grades. The significant causal pathways found between 300-level average course outcomes and end-of-program proficiency testing likely reflect this practice, as learners who complete the 300-level courses reflect the strongest learners in the cohort.

⁴⁹ Exceptions to this finding include Chinese speaking (in which 200-level average coursework outcomes predicted OPI speaking scores) and Korean speaking (in which no significant causal pathways were found between average course outcomes and OPI speaking scores).

With few exceptions, a great deal of coherence was found in the development of Arabic, Chinese, and Korean foreign language proficiency skills. This finding provides validity evidence for situating these languages within the same category within the current DoD classification scheme. When looking at the development of learners' proficiency within the same language, a great deal of model overlap is found across skills. For Arabic, almost identical causal pathways are found across the reading, listening, and speaking skills (the Arabic reading skill includes an additional causal pathway from 200-level average course outcomes to DLPT proficiency test score outcomes and an unexpected negative causal pathway between Language Preference Self-Assessment scores and average 100-level coursework outcomes). For Chinese, with the exception of just two paths, identical path models are found across the reading, listening, and speaking skills (a significant causal pathway was found between learners' AFQT scores and reading proficiency outcomes, and 200-level average coursework outcomes predicted OPI speaking outcomes). Korean was found to have the most within-language variability: With the exception of the significant causal pathway found between learners' AFQT scores and DLPT reading proficiency scores and the negative causal pathway between Language Preference Self-Assessment and DLPT reading scores, identical causal pathways were found between Korean reading and listening skills. The development of Korean speaking proficiency identified unexpected negative causal pathways between learners' AFQT scores and Language Preference Self-Assessment scores and OPI speaking outcomes. Unlike the reading and listening skills, no significant causal pathways were found from average 100-, 200-, and 300-level coursework and OPI

speaking outcomes. As mentioned above, this may indicate potential instability in the measurement of the performance-based OPI assessment of the Korean speaking skill.

Study 2: Validity of the Imputation Procedure for Path-Analytic Analyses Within a Large-Scale, L2 Instructional Context

As stated above, Study 1 included learners with observed longitudinal records (i.e., no missing records). Study 2, described in detail below, will follow the same methodological procedures as Study 1, but will impute data, or substitute estimated values, for missing DLIFLC achievement and/or DLPT/OPI proficiency outcome measures, given the institutional re-cycling and re-linguaging policies.

Research Question 4a: For the Arabic language, are significant differences observed between models containing observed learner records versus imputed learner records?

The purpose of Study 2 is to investigate research question four by examining the potential hidden effects of non-random attrition across learners and to determine the extent to which observed outcomes in Study 1 are influenced by the casewise deletion of missing data. The outcomes from Study 2 will be then compared to the outcomes from Study 1 across all languages and skills. Any observed differences and/or similarities in the resulting significant causal pathways for each language and skill, as well as the potential implications for working with observed versus incomplete data, will be discussed.

Missing Data: Overview

Missing data occurs when data values are not stored for observed variables within a given dataset, reducing the representativeness of the sample and potentially distorting inferences made about the larger population. The data-imputation process replaces missing data with substituted values, allowing researchers to avoid having to use casewise deletion of missing data, which has been found to result in the overestimation of statistical outcomes and often significantly reduces the n-size of a given dataset (Tabachnick & Fidell, 2007). The substituted values become *estimated values* calculated

from other available information within the dataset. There are three basic types of missing data: (1) Missing Completely at Random (MCAR), (2) Missing at Random (MAR), and (3) Missing Not at Random (MNAR) (Rubin, 1976; Little & Rubin, 2002; Enders, 2010; Jackson, 2016). In his overview of the characteristics of each the three types of missing data, Jackson (2016) describes the missingness of MCAR data to be independent of both observable and unobservable parameters of interest, occurring entirely at random and unrelated to any study-related variable of interest. Examples of MCAR data include a participant accidentally advancing a prompt without recording their response or inadvertently skipping a question on a questionnaire. For MAR data, the probability of missingness is related to some measured variable within the dataset. In other words, there is a systematic reason that accounts for the missing data. An example of MAR data might be that males typically don't answer questions on surveys that are deeply personal, such as those having to do with depression or mental health. Lastly, MNAR data is characterized by non-ignorable non-responses within a dataset. That is, as described by Jackson (2016), the value of the variable that is missing is related to the reason that it is missing. Building on the example, an example of MNAR data might be that males do not answer certain questions on a survey related to depression or mental health *because* they are experiencing certain levels of depression. Prior to engaging in the data-imputation process, it is imperative to examine the missingness of one's data since the parameters associated with the imputation procedures are influenced by how the missing data within a dataset are characterized.

To explore the nature of missing data within the current dataset, descriptive statistics concerning the number of missing cases associated with each of the 17

endogenous variables informing the path analyses were generated for the Arabic, Chinese, and Korean languages. Figure 77 through Figure 85 display the results.

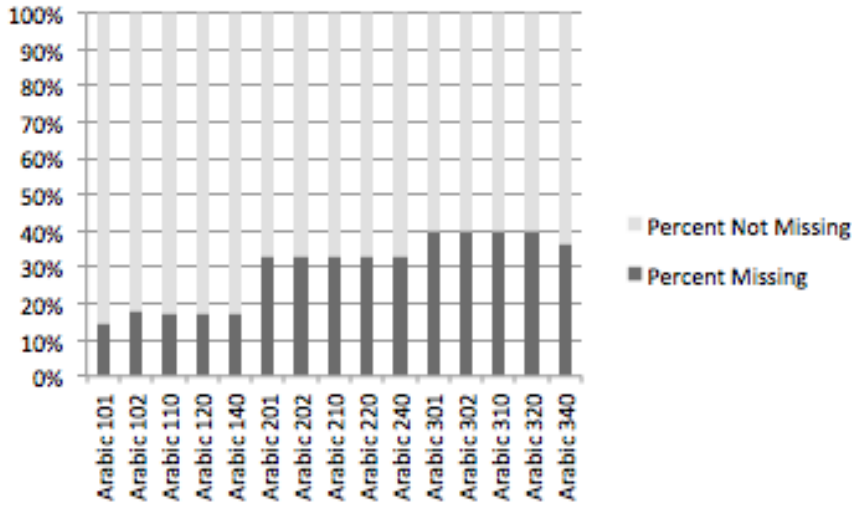


Figure 77. Arabic Pattern of Missing Data (Coursework)

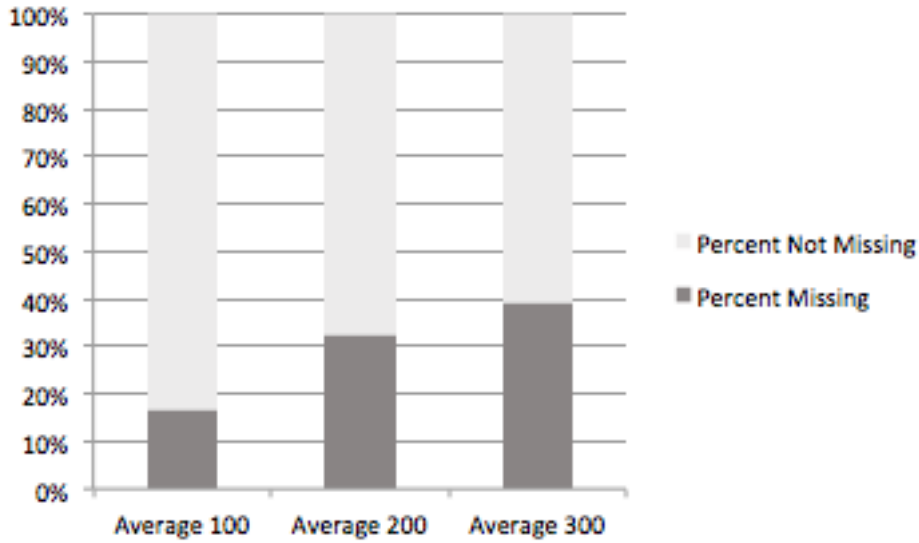


Figure 78. Arabic Missing Data: Course averages

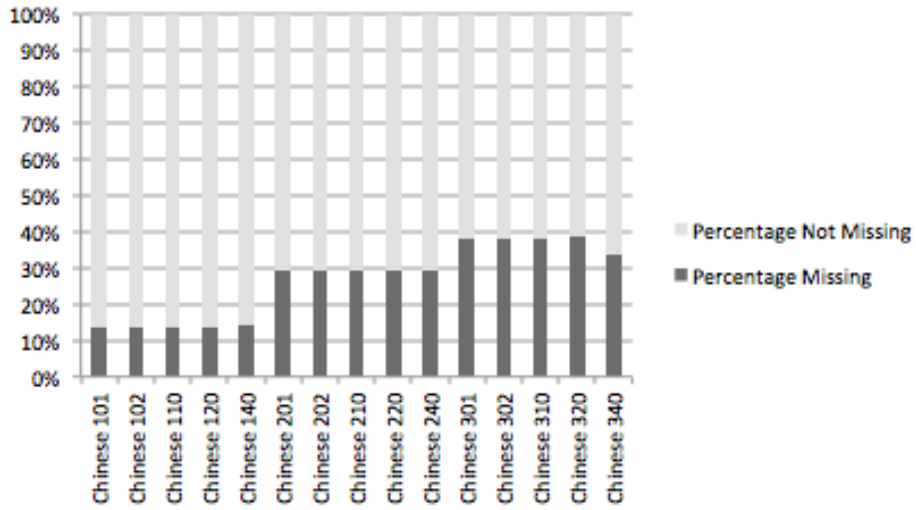


Figure 79. Chinese Pattern of Missing Data (Coursework)

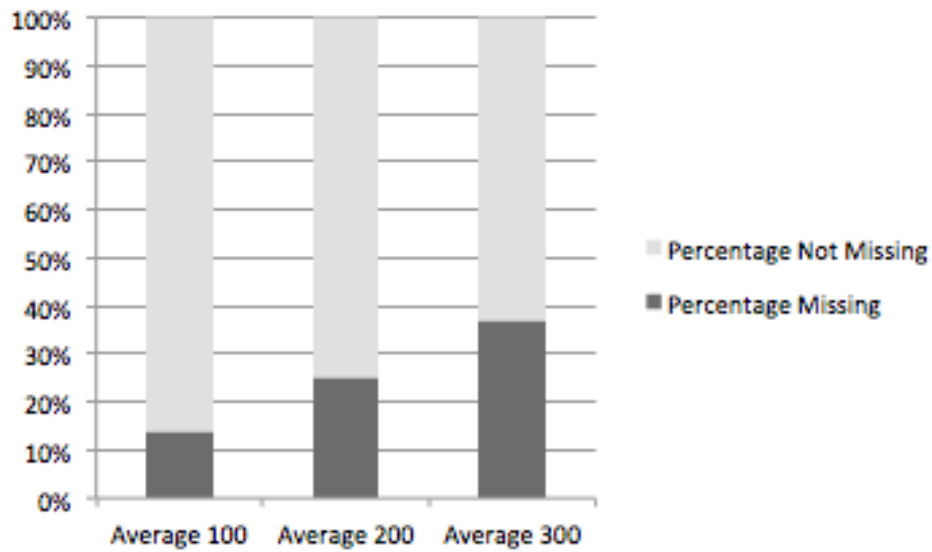


Figure 80. Chinese Missing Data: Course averages

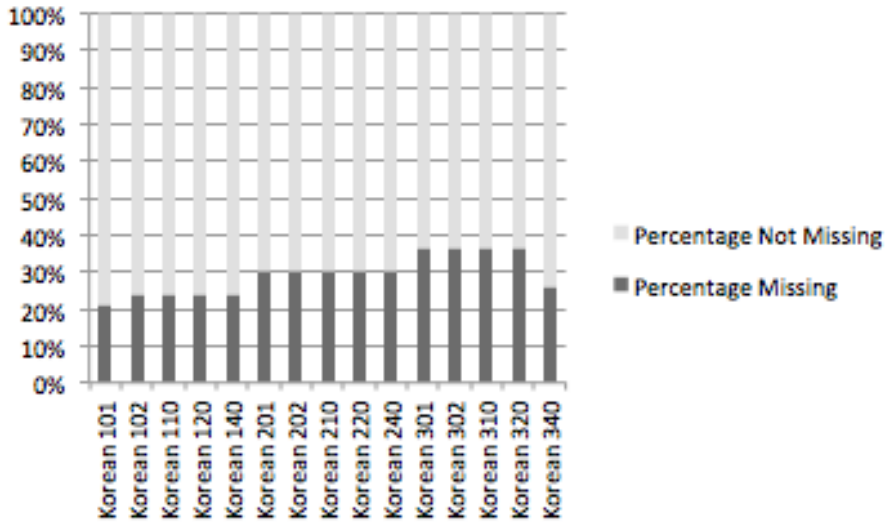


Figure 81. Korean Pattern of Missing Data (Coursework)

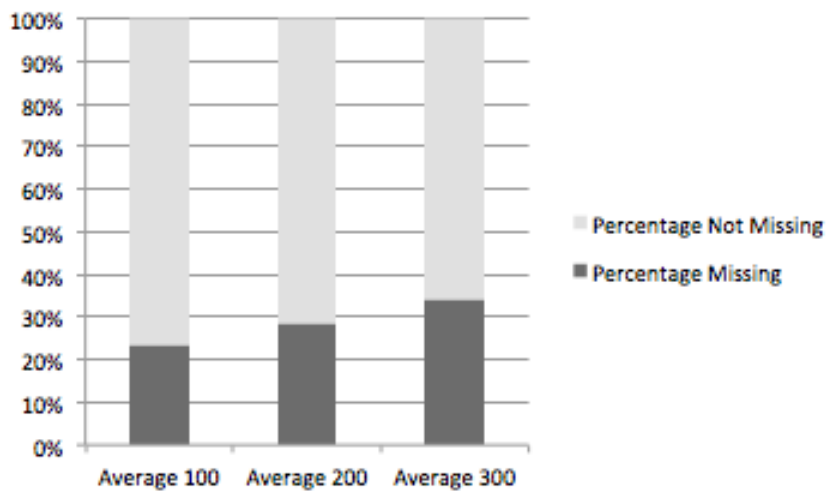


Figure 82. Korean Missing Data: Course averages

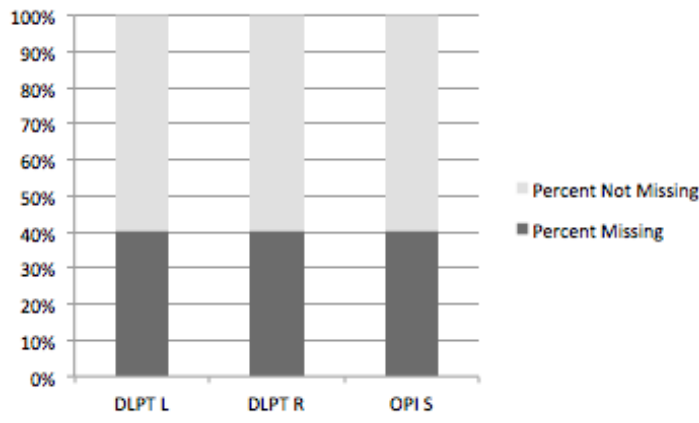


Figure 83. Arabic Wave 3 Missingness

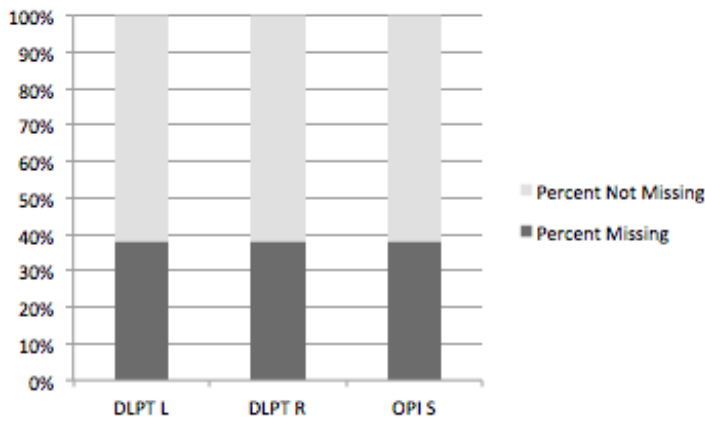


Figure 84. Chinese Wave 3 Missingness

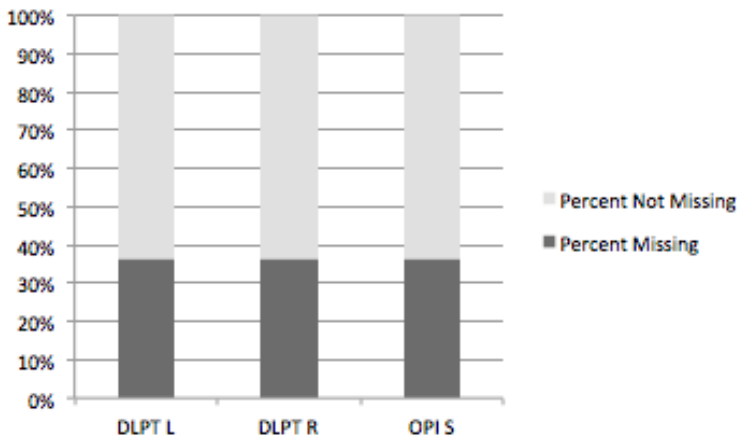


Figure 85. Korean Wave 3 Missingness

As shown in Figure 77 through Figure 85, all 18 items serving as downstream variables within the path-analytic framework (all Wave 2 and 3 variables for each language) contain missing data. Notably, learners in the Arabic, Chinese, and Korean initial acquisition courses at DLIFLC appear to attrite in systematic patterns. Another way of examining the missing value patterns across the Arabic, Chinese, and Korean languages is to model the percentage of missing data across course achievement outcomes. If languages grouped within the same difficulty category are equally difficult, it would

follow that the percentage of learners who no longer continue their language study would also be comparable, likely attributable to the systematicity with which learners are asked to begin their language study again or to completely drop out of language study by teachers or program administrators. **Figure 86** below displays the results of this analysis, organized by the average percentage of missing data aggregated within the 100, 200, and 300 course levels of study.

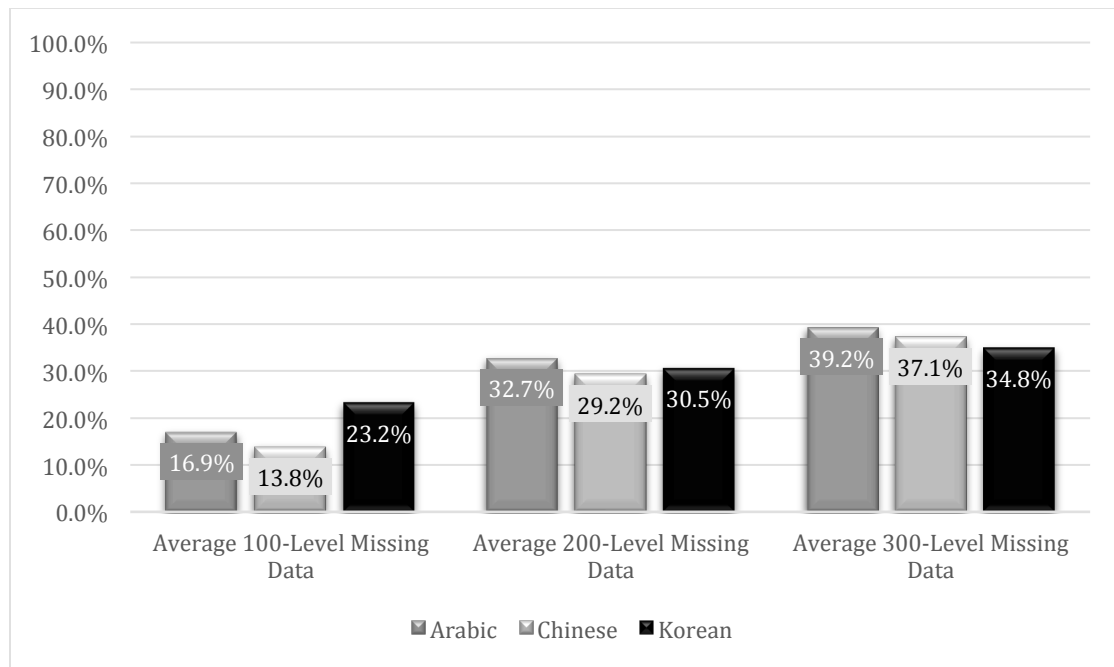


Figure 86. Patterns of Missingness across language groups

Across all languages, the fewest cases of missing data are associated with the 100-level courses and the most cases of missing data are associated with the 300-level courses. This trend is understandable given that languages associated with Category IV study are posited to be the most difficult to acquire, resulting in some learners no longer continuing with their foreign language training after participating in initial 100-level coursework. It logically follows that learners who participated in Arabic initial acquisition training (through their Arabic 240 coursework, for example) and then dropped out, will not have

achievement-related data for the Arabic 301, 302, 310, 320, and 340 courses, nor end-of-program proficiency test scores for the listening, reading, and speaking skills. This type of pattern of missing data is referred to as monotonic, since once a learner within the current dataset drops out, she or he does not later reappear in the dataset.

As shown in Figure 86, the Arabic, Chinese, and Korean languages display the same general patterns of missing data throughout initial acquisition training at DLIFLC. The largest disparities in the percentage of missing data is found between Chinese and Korean (at the 100 level), with Korean containing 10% more missing data than Chinese, as well as between Arabic and Korean (also at the 100 level), with Korean containing about six percent more missing data than Arabic. The percentages of missing data for the 200 and 300 levels of coursework are comparable across all three languages, showing less than a three percent difference in average missing data. The observed patterns in Figure 86 imply that these missing data are MNAR, since the nature of missingness can be accounted for by a “non-ignorable non-response” within the dataset: namely, a “drop-out” variable. This suggests that the same individuals who are missing in the 200 level are also missing in the 300 level. While it is typically not recommended that the imputation process be employed with MNAR data, Sinharay et al. (2001) argue that the procedure can be performed if the nature of missingness can be systematically accounted for. They state,

If one can collect information on a number of good predictor variables that might govern the missingness mechanism, the MAR assumption (and hence the results from MI [multiple imputation] becomes more plausible....If there is a variable which alone governs missingness, we will have [a] MAR situation if we collect information on that variable. This is why the common advice about MI is that one should collect information about any characteristics that might even remotely affect missingness and include those characteristics in the imputation model (p. 321).

As mentioned previously, DLIFLC adheres to a “re-linguaging” or “re-cycling policy”, in which learners who do not appear to be well-positioned to meet DLIFLC graduation requirements are asked by instructors to either: (1) begin their instruction again (referred to as “re-cycling”) or (2) begin instruction in a different, lower-category language (referred to as “re-linguaging”). Further, as mentioned in Chapter 2 and outlined in Appendix B, the current dataset is a subset of a larger dataset containing 2,263 records and 244 separate variables. This original, larger dataset contained the predictors to be used in developing the Defense Language 2 (DLAB 2) aptitude battery. The 244 variables within the larger dataset represented observed scores from five different dimensions posited to predict one’s likelihood of succeeding in intensive foreign language study at the DLIFLC. The five dimensions of language-learning attributes within the DLAB 2 dataset included: (1) Existing test scores (ASVAB, DLAB, and their subcomponents, 22 variables, see Bunting et al., pp. 3-5 – 3-27), (2) Demographic and Biographical Variables (32 variables, described on pages 3-27 – 3-50), (3) Cognitive and Perceptual Measures (25 variables, described on pages 3-51 – 3-82), (4) Personality Measures (28 variables, described on pages 3-83 – 3-119), and (5) Motivational Measures (31 variables, described on pages 3-119 – 3-161). Using a variety of data reduction procedures (such as exploratory factor analysis and confirmatory factor analysis), and data reduction procedures (such as correlational analyses, reviews of multicollinearity statistics, and factor structures), Bunting et al. (2011) reduced the 116 variables contained within dimensions two through four described above to a final set of 58 predictor variables (see Appendix B) to be used as candidates for inclusion in the DLAB 2

predictive model.⁵⁰ The author used this same set of predictor variables as input into the imputation procedure associated with the current analysis. Given that the drop-out variable can be accounted for directly by DLIFLC’s “re-cycling” and “re-languaging” policies, and that Bunting et al. (2011) used the same set of predictor variables to model the probability of learners meeting DLPT and OPI proficiency standards, the variability in learner outcomes is accounted for by including the 58 predictor variables in the imputation model. Therefore, although the data in its original form can be characterized as MNAR, because the nature of missingness can be explained by a suite of cognitive and non-cognitive variables that have already found to predict proficiency test score outcomes, it is argued that the data can be considered MAR and that the imputation process can proceed.⁵¹

As summarized succinctly by Jackson (2016), a variety of imputation methods are available to researchers working with missing data, such as hot deck, mean, or single regression.⁵² The challenge associated with these traditional methods is that they do not take into account the uncertainty, or variability, of data within the imputation model. To address this issue, Rubin (1987) argues that repeated or multiple imputations greatly improve the quality of estimates generated for missing data. As described by Van Burren (2017), multiple imputation involves three main steps. The first step entails imputing (or

⁵⁰ For a thorough overview of the candidate variables analyzed for inclusion in the DLAB 2 predictive model, please see Chapter 3 of Bunting et al., 2011. For each potential predictor, the chapter describes a brief overview of the test or question, scoring mechanisms, response distributions, and reliability of the measure.

⁵¹ Bunting et al.(2011) found that 23 of the 58 variables predicted learners’ likelihood of reaching DLPT and OPI proficiency test score criteria. For a detailed discussion of their analyses and results see pp. 6-1 – 6-36.

⁵² Little (1987) states, “commonly used procedures for imputation include *hot deck* imputation, where means from sets of recorded values are substituted; *mean* imputation, where means from sets of recorded values are substituted, and *regression* imputation, where the missing variables for a unit are estimated by predicted values from the regression on the known values on that unit” (p. 6).

filling in) missing entries of incomplete datasets not just once but multiple times, resulting in imputed values drawn from a distribution which can be different for each missing case. This process results in the generation of m complete datasets, depending on the number specified by the researcher.⁵³ The second step involves analyzing the data specified within the path-analytic framework (outlined in detail in Chapter 2) using the m unique imputed datasets. The final step involves integrating the results of the m analyses into the final set of path-analytic outcomes. The outcomes from this final, pooled set of result from each of the m analyses conducted on the separate m imputed datasets. However, a challenge associated with working from a final, pooled estimate of imputed values is that a novice researcher can be blind to variability often present across the m imputed datasets when working solely from the final pooled estimations. Figure 87, below, taken from Van Buuren (2017) visually depicts this process.

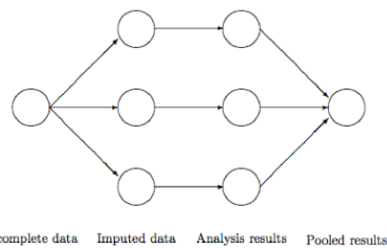


Figure 87. Visual depiction of the Multiple Imputation Process (taken from Van Buuren, 2017)

Based on this recommendation to work with the results from multiple imputation procedures, rather than from the results of a single imputation procedure, the multiple, regression-based imputation method was employed within the SPSS platform. In order to explore potential differences in path-analytic outcomes between complete versus imputed

⁵³ Generally, five to 20 separate datasets are generated during this step of the imputation process (<http://www.stefvanburren.nl/mi/MI.html>).

data, the imputation model was customized to address the monotonic pattern of missing data found for the Arabic, Chinese, and Korean languages.

The SPSS imputation model allows the user to select the desired role for each of the variables within the imputation model. These include: (1) predict only, (2), impute and predict, and (3) impute only. Since none of the Wave 1 qualification and screening variable (ASVAB, DLAB, and Language Preference Self-Assessment) scores were missing, all three variables were entered into the imputation model as predictor variables only. Next, all 58 DLAB 2 predictor variables and 12 course outcome variables were entered into the imputation model as both predictor variables and variables to be imputed. Lastly, since proficiency test score results represent the final wave of the path-analytic framework, the three end-of-program proficiency-testing variables were entered into the imputation model as variables to be imputed only. Next, the author consulted Bunting et al. (2011) in order to define the minimum and maximum constraints associated with each of the 58 DLAB 2 predictor variables. This allows for reasonable limits to be established around the mathematical algorithms used to create the m imputed datasets.⁵⁴

Following Jackson et al. (2011), the number of imputation methods was set at 10. The same path analyses completed for Study 1 were completed for Study 2, working from imputed data rather than the observed dataset as input into the path-analytic procedure. For exploratory purposes, in order to gauge the degree of variability across each of the 10 imputed datasets, rather than working from a final, pooled estimate, in

⁵⁴ During the imputation estimation process, the SPSS software required a number of constraints to be released in order to generate final estimates. Of note was the requirement to release most of the 4.0 GPA constraints across the 100-, 200-, and 300-level achievement related variables and the constraint of 36 (representing a score of 3+) for end-of-program proficiency test score outcomes. Upon model convergence, the author replaced values that exceeded or fell below these required caps with the appropriate minimum and maximum values.

which variation across imputation models would be hidden, the author employed a simplified pooling approach to compute the median standardized regression weights and standard errors associated with each of the ten separate imputed datasets for each language. Results were then compared against the outcomes from the observed datasets in order to validate the use of multiple imputation procedure for the current investigation. Observed differences between observed and imputed models were both visually and statistically inspected. Comparisons of interest for Study 1 included differences in path-analytic outcomes between the observed and imputed datasets across languages and skill modalities.⁵⁵ Overall, outcomes from Study 2 (working with the pooled estimates generated through AMOS) were then compared with the outcomes from Study 1 in order to explore differences in statistical results when working with observed datasets (using casewise deletion methods) versus using imputed data (working from the median results across 10 imputed datasets). The results of these analyses are discussed below.

Results: Study 2

Research Question 4 (RQ4b): For languages grouped within the same category, are similar patterns of individual differences in general aptitude, language-specific aptitude, and motivation observed in the prediction of learners' success as they progress through coursework for complete versus imputed datasets?

Study 2 Results: Wave 1 to Wave 2 Path Analyses: Reading, Listening, and Speaking

The current section discusses the results of the Wave 1 to Wave 2 path analyses for the reading, listening, and speaking skills concurrently, since average coursework outcomes were not broken down by skill. Outcomes of interest include both differences in imputed models across the Arabic, Chinese, and Korean languages and differences in

⁵⁵ For the imputed datasets, the groups of interest were the median path weights computed from each of the 10 separate path-analytic outcomes generated as part of the multiple imputation procedure.

observed models between complete and imputed datasets within the same language. For each language and skill, across all 10 imputations, the author identified and counted the significant causal pathways in each model. Pathways that were found to be significant in at least six of 10 models (greater than chance) were included in the synthesized model and are displayed below. Median path weights and squared multiple correlations (for the endogenous variables in each model) across all 10 models were also calculated and reported.

Beginning with the AFQT score Wave 1 predictor variable using the imputed data, just two statistically significant path coefficients were found: for the Arabic language, significant negative causal pathways were found from AFQT scores to average 100-level and 300-level coursework outcomes. Comparing the complete versus imputed Wave 1 to Wave 2 Arabic models, for the complete records, no significant causal pathways between AFQT scores and average 100-, 200-, or 300-level outcomes were found. The differences in findings between the complete and imputed models, while initially counterintuitive, provide validity evidence for the use of AFQT scores as a screening instrument. That is, while negative causal pathways would not be expected for the complete learner records containing only those learners who have not been re-learned or re-cycled, 16.9% of average course-outcome data were estimated by the imputation procedure. Given that students are asked to either start a different, lower category language or begin their instruction again only when showing signs of academic struggle, it logically follows that the estimated average course outcome scores for the imputed sample would be lower, reflecting the estimated scores of students who would have been asked to discontinue their language study due to less promising academic

performance. This potentially accounts for the negative causal relationship found between AFQT scores and Arabic average 100- and 300-level coursework outcomes.

The DLAB score predictor variable was found to consistently predict average 100-, 200-, and 300-level course outcomes across all languages for both the complete and imputed datasets. This finding provides additional corroborating evidence that the DLAB plays a robust role as a screening instrument for placing learners into a Category IV language of study, irrespective of the increased variability introduced by the imputed lower performing learners. Lastly, for the imputed datasets and Arabic language only, a positive causal pathway was found between the Language Preference Self-Assessment variable and average 100-level course outcomes. In other words, the less pleased that Arabic language learners were with their assigned language, the lower their average course outcomes. This finding contradicts what was found for the observed records where a significant negative causal relationship was found between Language Preference Self-Assessment scores and average 100-level course outcomes. This indicates that, for the observed data, the less pleased that Arabic learners were with their assigned language, the higher their average course outcomes. As postulated in the discussion of Study 1 results, it is important to note that this variable only represents learners' responses to one question on a self-reported survey asking if he or she was assigned to their first language choice, which may not adequately distinguish the role that language preference may play as the learners progress through DLIFLC coursework. Figure 88 through Figure 93 below visually depict the results of these analyses.

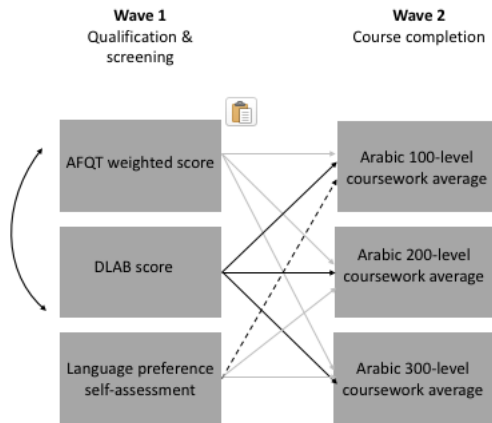


Figure 88. Wave 1 to Wave 2 Model Arabic (Observed)

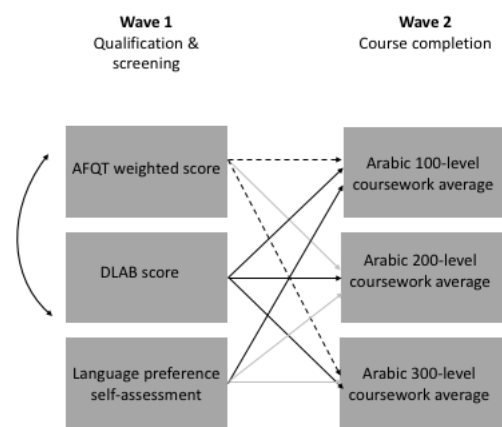


Figure 89. Wave 1 to Wave 2 Model Arabic (Imputed)

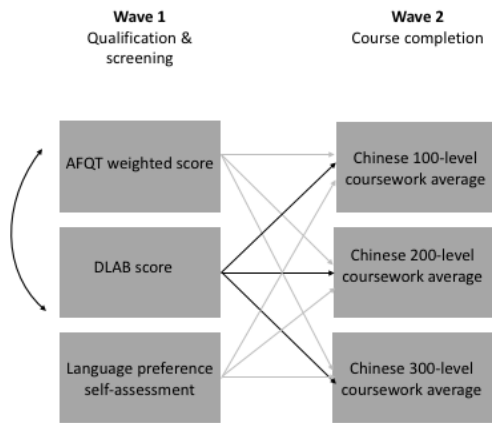


Figure 90. Wave 1 to Wave 2 Model Chinese (Observed)

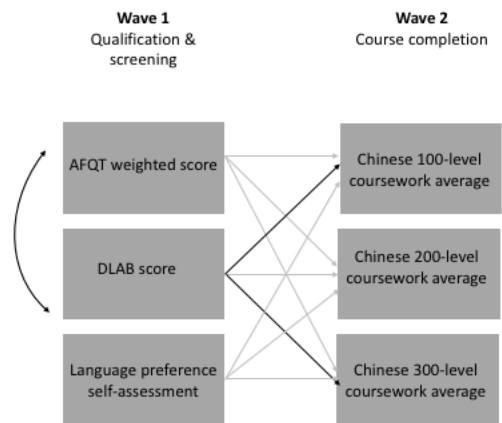


Figure 91. Wave 1 to Wave 2 Model Chinese (Imputed)

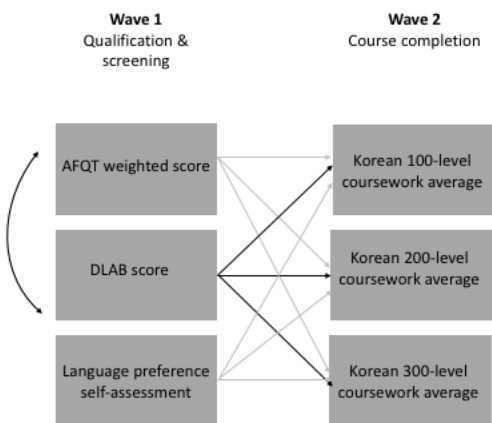


Figure 92. Wave 1 to Wave 2 Model Korean (Observed)

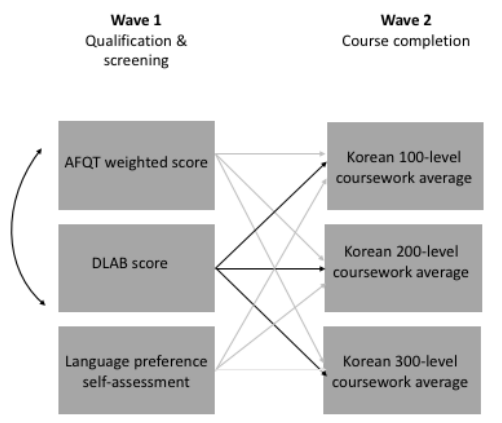


Figure 93. Wave 1 to Wave 2 Model Korean (Imputed)

To assist with the interpretation of the relative magnitude of each of the causal pathways for both the complete and imputed datasets, Table 42 through Table 44 display the standardized regression weight, standard error, and significance level for each of the significant pathways in the above models. For ease of interpretation, results from both the complete and imputed datasets are reported below as well as a note regarding the number of times a pathway was found to be significant across models.⁵⁶

Table 42. Arabic Wave 1 to Wave 2 path-analytic outcomes (Observed (n= 241) and Imputed (n = 411))

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT to 100-level average coursework	N/A (Imputed) -0.141	(Observed) N/A (Imputed) 0.007	N/A 9 significant pathways
AFQT to 300-level average coursework	N/A (Imputed) -0.114	(Observed) N/A (Imputed) 0.009	N/A 8 significant pathways
DLAB to 100-level average coursework	(Observed) 0.322 (Imputed) 0.314	(Observed) 0.020 (Imputed) 0.020	(Observed) $p < 0.001$ 10 significant pathways
DLAB to 200-level average coursework	(Observed) 0.311 (Imputed) 0.292	(Observed) 0.021 (Imputed) 0.023	(Observed) $p < 0.001$ 10 significant pathways
DLAB to 300-level average coursework	(Observed) 0.288 (Imputed) 0.262	(Observed) 0.022 (Imputed) 0.028	(Observed) $p < 0.001$ 10 significant pathways
Language Preference Self-Assessment to 100-level average coursework	(Observed) -0.121 (Imputed) 0.111	(Observed) 0.026 (Imputed) 0.023	(Observed) $p < 0.05$ 9 significant pathways

⁵⁶ Significance levels for the imputed data are not reported, because they were not calculated across models.

Table 43. Chinese Wave 1 to Wave 2 path analytic outcomes (Observed (n = 98) and Imputed (n = 161))

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to 100-level average coursework	(Observed) 0.338 (Imputed) 0.268	(Observed) 0.031 (Imputed) 0.030	(Observed) p < 0.01 9 significant pathways
DLAB to 200-level average coursework	(Observed) 0.233 (Imputed) 0.174	(Observed) 0.028 (Imputed) 0.033	(Observed) p < 0.05 5 significant pathways
DLAB to 300-level average coursework	(Observed) 0.282 (Imputed) 0.224	(Observed) 0.025 (Imputed) 0.033	(Observed) p < 0.01 9 significant pathways

Table 44. Korean Wave 1 to Wave 2 path analytic outcomes (Observed (n = 75) and Imputed (n = 118))

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to 100-level average coursework	(Observed) 0.378 (Imputed) 0.241	(Observed) 0.034 (Imputed) 0.035	(Observed) p < 0.001 7 significant pathways
DLAB to 200-level average coursework	(Observed) 0.350 (Imputed) 0.227	(Observed) 0.033 (Imputed) 0.050	(Observed) p < 0.01 8 significant pathways
DLAB to 300-level average coursework	(Observed) 0.399 (Imputed) 0.315	(Observed) 0.031 (Imputed) 0.038	(Observed) p < 0.001 9 significant pathways

As can be gleaned by Table 42 through **Table 44** above, and consistent with the findings from Study 1, comparable path coefficients can be found across all three languages for the imputed datasets. This finding suggests that the DLAB plays a robust role in predicting coursework success, irrespective of the variability introduced when including both high- and low-performing learners within a path-analytic model. To test for statistically significant differences in common path weights between the Observed and imputed dataset, the author employed a z transformation to test for differences in correlations.⁵⁷

⁵⁷ Consistent with the procedures employed in Study 1, direct effects between Wave 1, Wave 2, and Wave 3 predictor variables within the same language for the observed and imputed datasets are equivalent to correlation coefficients. The path coefficients can be compared by transforming the observed coefficients to a z-value. The equation is: $z = \frac{z_{oa} - z_{ib}}{\sqrt{1/(N_a - 3) + 1/(N_b - 3)}}$. Values outside of the critical value +/- 1.96 indicate a statistically significant difference in observed coefficients.

Table 45. Differences in Observed versus imputed correlations

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Arabic DLAB to 100-level average coursework (Observed and Imputed)	0.109	Not significant
Arabic DLAB to 200-level average coursework (Observed and Imputed)	0.256	Not significant
Arabic DLAB to 300-level average coursework (Observed and Imputed)	0.345	Not significant
Arabic Language Preference to 100-level average coursework (Observed and Imputed)	-2.857	Significant
Chinese DLAB to 100-level average coursework (Observed and Imputed)	0.594	Not significant
Chinese DLAB to 300-level average coursework (Observed and Imputed)	0.742	Not significant
Korean DLAB to 100-level average coursework (Observed and Imputed)	1.010	Not significant
Korean DLAB to 200-level average coursework (Observed and Imputed)	0.894	Not significant
Korean DLAB to 300-level average coursework (Observed and Imputed)	0.641	Not significant

As noted above in

Table 45, outcomes from the statistical comparison of common path weights between observed and imputed models yielded just one significant difference: Arabic Language Preference Self-Assessment to 100-level average coursework outcomes. This finding suggests that the established standardized regression weight from the Arabic Language Preference Self-Assessment score to 100-level average coursework outcomes were statistically significantly lower for the imputed than the observed dataset.

Study 2 Results: Wave 2 to Wave 3 Path Analyses: Reading

The current section will discuss both the observed and imputed path-analytic outcomes of Wave 2 exogenous predictor variables to Wave 3 endogenous outcome variables separately by skill.⁵⁸ Beginning with the predictive influence of average 100-, 200-, and 300-level coursework on DLPT reading outcomes, across almost all languages and models, 300-level average course reading outcomes consistently predict end-of-program DLPT reading test outcomes. Figure 94 through Figure 99 below visually depict the results of these analyses.

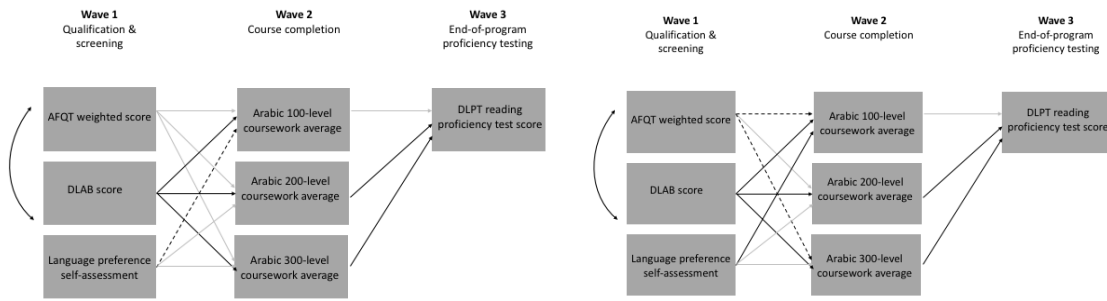


Figure 94. Wave 2 to 3 Arabic Reading (Observed)

Figure 95. Wave 2 to 3 Arabic Reading (Imputed)

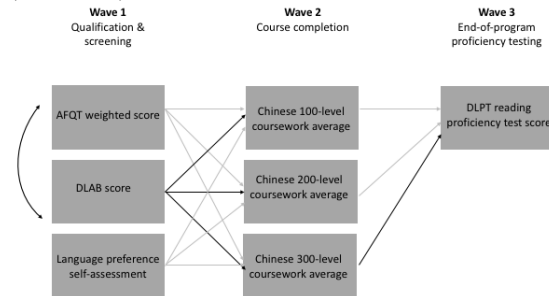


Figure 96. Wave 2 to 3 Chinese Reading (Observed)

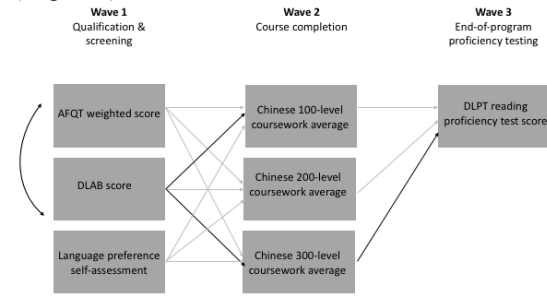


Figure 97. Wave 2 to 3 Chinese Reading (Imputed)

⁵⁸ Consistent with the discussion of findings for Study 1, Wave 3 outcome variables differ by the listening, reading, and speaking skills and therefore will be discussed separately.

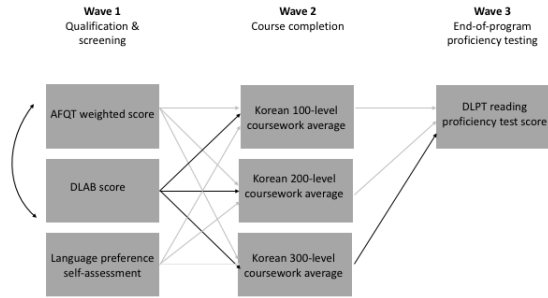


Figure 98. Wave 2 to 3 Korean Reading (Observed)

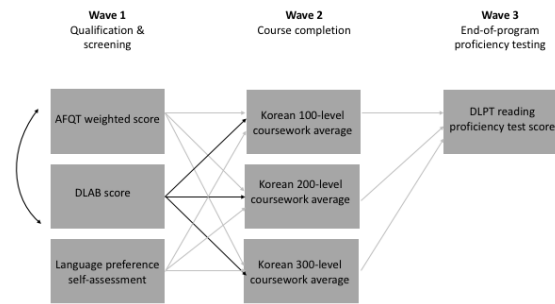


Figure 99. Wave 2 to 3 Korean Reading (Imputed)

As shown in Figure 94 through Figure 99, identical Wave 2 to Wave 3 significant causal pathways were found for the Arabic and Chinese languages for both the observed and imputed datasets. Absent in the imputed Korean model is a significant causal pathway from 300-level coursework averages to DLPT reading outcomes. This indicates that 300-level course outcomes are robust enough to predict DLPT reading outcomes when only the strongest Korean language learners are modeled, but they are not quite robust enough to predict success when weaker Korean learners are included in the model. Interestingly, this was not the case for the Arabic and Chinese languages, potentially suggesting that more learners than necessary were recommended to be “re-linguaged” or “re-cycled”. That is, it could be the case that instructors in the Arabic and Chinese languages were overly cautious in their recommendations to re-situate learners into other languages or programs of study.

To assist with the interpretation of the figures above, Table 46 through Table 48 display the associated standard error, significance level, and standardized regression weights for each of the significant pathways in the above model.

Table 46. Arabic Reading: Wave 2 to Wave 3 path-analytic outcomes (Observed (n = 241) and Imputed (n = 411))

Path	Standardized	Standard Error	Significance Level
------	--------------	----------------	--------------------

	Regression Weight		
200-level average coursework to DLPT Reading	(Observed) 0.255 (Imputed) 0.393	(Observed) 4.130 (Imputed) 3.373	p < 0.05 10 significant pathways
300-level average coursework to DLPT Reading	(Observed) 0.314 (Imputed) 0.299	(Observed) 3.389 (Imputed) 2.357	p < 0.001 6 significant pathways

Table 47. Chinese Reading: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 98 and Imputed (n = 161))

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Reading	(Observed) 0.527 (Imputed) 0.242	(Observed) 4.919 (Imputed) 4.026	p < 0.001 6 significant pathways

Table 48. Korean Reading: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118))

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Reading	(Observed) 0.392 (Imputed) 0.161	(Observed) 4.916 (Imputed) 0.740	p < 0.05 2 significant pathways

For the imputed data, as shown in Table 46 through Table 48 above, the magnitude of the shared 300-level path weights between Arabic and Chinese are comparable across languages.

To test for statistically significant differences in common path weights between the observed and imputed dataset, the author employed z transformations to test for differences in path-analytic outcomes between the observed and imputed datasets within the same language.⁵⁹

⁵⁹ Consistent with the procedures employed in Study 1, direct effects between Wave 1 and Wave 2 predictor variables are equivalent to correlation coefficients. The path coefficients can be compared by

Table 49. Differences in observed versus imputed correlations

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Arabic 200-level average course outcomes to DLPT Reading (Observed and Imputed)	-1.895	Not significant
Arabic 300-level average course outcomes to DLPT Reading (Observed and Imputed)	0.210	Not significant
Chinese 300-level average course outcomes to DLPT Reading (Observed and Imputed)	2.612	Significant

As noted in Table 49, outcomes from the statistical comparison of common path weights between observed and imputed models found significant differences for the Chinese language, indicating that the predictive influence of 300-level average course outcomes is markedly stronger for the observed dataset (0.527) than for the imputed dataset (0.242). This finding suggests that, although 300-level average coursework outcomes play a significant role in predicting DLPT reading proficiency outcomes, it does so most strongly when the model includes just those learners identified as strongest candidates for taking the DLPT reading test.

Study 2 Results: Wave 2 to Wave 3 Path Analyses: Listening

Consistent with the findings discussed in Study 1, relatively coherent patterns of foreign language listening proficiency development were found across the Category IV languages for both the observed and imputed datasets. For the Arabic and Chinese languages, of note is the additional causal pathway found between 200-level average coursework outcomes and DLPT listening outcomes. Similar to what was found for the Arabic reading proficiency skill, this finding suggests that more than anticipated numbers

transforming the observed coefficients to a z-value. The equation is: $z = \frac{z_{ob} - z_{im}}{\sqrt{1/(N_o - 3) + 1/(N_i - 3)}}$. Values outside of the critical value +/- 1.96 indicate a statistically significant difference in observed coefficients.

of Arabic and Chinese learners might have successfully progressed through 200-level coursework and potentially have met DLPT listening proficiency score outcomes. Figure 100 through Figure 105 below visually depict the results of these analyses. For the Korean language, identical models were established for both the observed and imputed datasets.

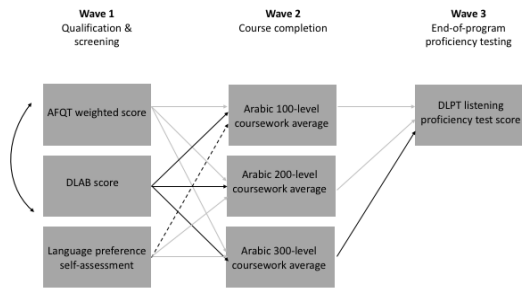


Figure 100. Wave 2 to 3 Arabic Listening (Observed)

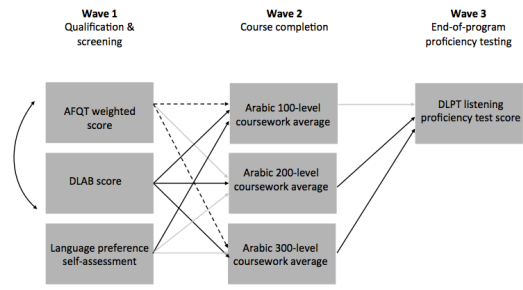


Figure 101. Wave 2 to 3 Arabic Listening (Imputed)

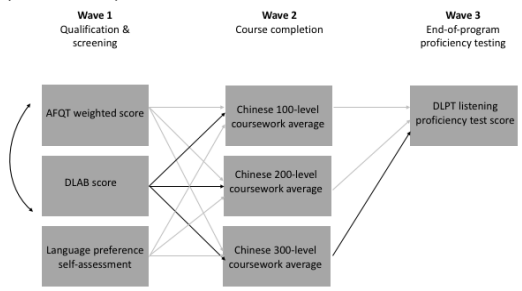


Figure 102. Wave 2 to 3 Chinese Listening (Observed)

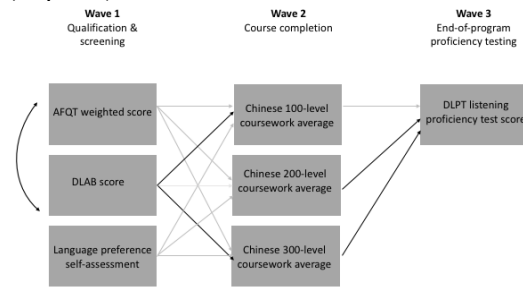


Figure 103. Wave 2 to 3 Chinese Listening (Imputed)

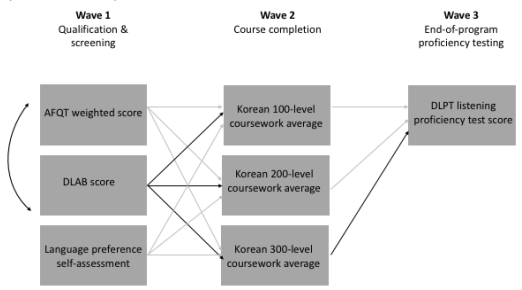


Figure 104. Wave 2 to 3 Korean Listening (Observed)

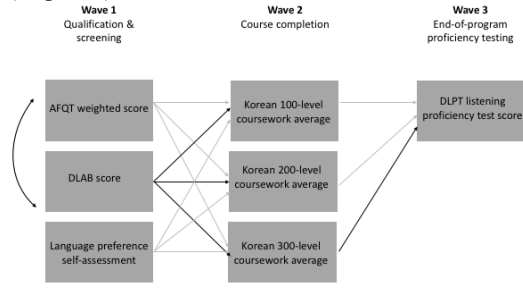


Figure 105. Wave 2 to 3 Korean Listening (Imputed)

To assist with the interpretation of the figures above,

Table 50 through Table 52 display the associated standard error, significance level, and standardized regression weights for each of the significant pathways in the above model.

Table 50. Arabic Listening: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 241 and Imputed (n = 411))

Path	Standardized Regression Weight	Standard Error	Significance Level
200-level average coursework to DLPT Listening	N/A (Imputed) 0.261	(Observed) N/A (Imputed) 3.119	N/A 10 significant pathways
300-level average coursework to DLPT Listening	(Observed) 0.334 (Imputed) 0.383	(Observed) 3.262 (Imputed) 2.237	p < 0.001 10 significant pathways

Table 51. Chinese Listening: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161))

Path	Standardized Regression Weight	Standard Error	Significance Level
200-level average coursework to DLPT Listening	N/A (Imputed) 0.077	(Observed) N/A (Imputed) 4.902	N/A 6 significant pathways
300-level average coursework to DLPT Listening	(Observed) 0.550 (Imputed) 0.259	(Observed) 7.395 (Imputed) 5.772	p < 0.001 6 significant pathways

Table 52. Korean (Observed n = 75; Imputed n = 118)

Path	Standardized Regression Weight	Standard Error	Significance Level
300-level average coursework to DLPT Listening	(Observed) 0.599 (Imputed) 0.266	(Observed) 6.543 (Imputed) 2.614	(Observed) p < 0.001 9 significant pathways

For the imputed data, as shown in

Table 50 through **Table 52** above, the magnitude of the shared 300-level path weights across languages are strongest for Arabic and comparable between Chinese and Korean. Results of significance testing of common path weights between observed and imputed path weights are reported in **Table 53** below.

Table 53. Differences in observed versus imputed correlations

Path Comparison	z-value (Reference Value +/- 1.96)	Significant Difference Between Languages?
Arabic 300-level average course outcomes to DLPT Listening (Observed and Imputed)	-0.690	Not significant
Chinese 300-level average course outcomes to DLPT Listening (Observed and Imputed)	2.720	Significant
Korean 300-level average course outcomes to DLPT Listening (Observed and Imputed)	2.789	Significant

Outcomes from the statistical comparison of common path weights between observed and imputed models found significant differences between the Chinese and Korean 300-level average coursework outcomes, indicating that the predictive influence of 300-level average course outcomes is substantially stronger for the observed datasets (0.550, 0.599) than for the imputed datasets (0.259, .266). Consistent with similar findings for the reading skill, this suggests that although 300-level average coursework outcomes play a significant role in predicting DLPT listening proficiency outcomes, they do so most strongly when the model includes just those learners identified as strongest candidates for taking the DLPT listening test.

Study 2 Results: Wave 2 to Wave 3 Path Analyses: Speaking

For the speaking skill, substantial model variability was found between the observed and imputed models. For the Arabic and Chinese languages, different causal pathways were found for the imputed datasets than for the observed datasets. For the Arabic imputed dataset, an additional, unanticipated significant causal pathway was found from 100-level average coursework outcomes to OPI speaking proficiency test scores. Similar to what was found for the Arabic reading and listening skills, this potentially indicates that more than anticipated Arabic learners might have successfully progressed through 100-level coursework and potentially have met OPI speaking proficiency test score outcomes. For the Chinese imputed dataset, in addition to the significant, positive causal pathway from 200-level average coursework outcomes to OPI speaking proficiency test scores, a significant, negative causal pathway was also found from 300-level average coursework outcomes to OPI speaking outcomes. Similar to what was found for the negative causal pathway between AFQT scores and 100- and 300-level average coursework outcomes for the Arabic language, average 300-level course outcome scores for the imputed sample are lower than for the observed learner sample, thus accounting for the negative causal relationship found between 300-level coursework outcomes and OPI speaking scores. For the Korean language, consistent with what was found for the observed dataset, no significant causal pathways were found between average 100-, 200-, and 300-level coursework outcomes and the OPI speaking test scores. Figure 106 through Figure 111 below visually depict the results of these analyses.

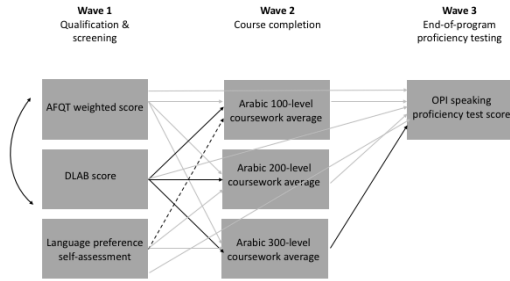


Figure 106. Wave 2 to 3 Arabic Speaking (Observed)

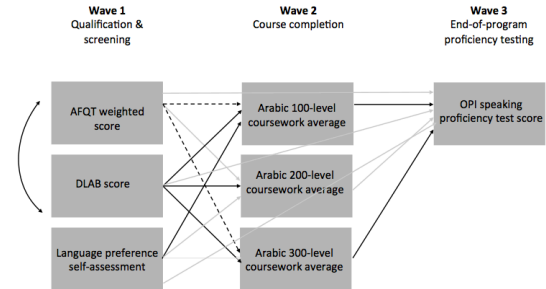


Figure 107. Wave 2 to 3 Arabic Speaking (Imputed)

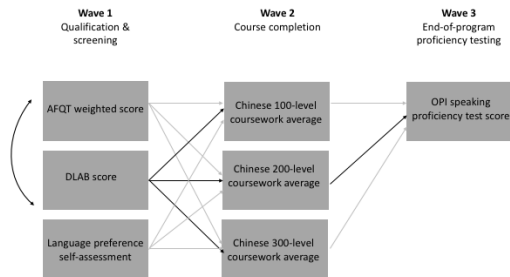


Figure 108. Wave 2 to 3 Chinese Speaking (Observed)

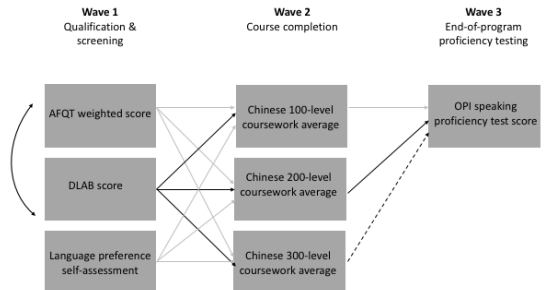


Figure 109. Wave 2 to 3 Chinese Speaking (Imputed)

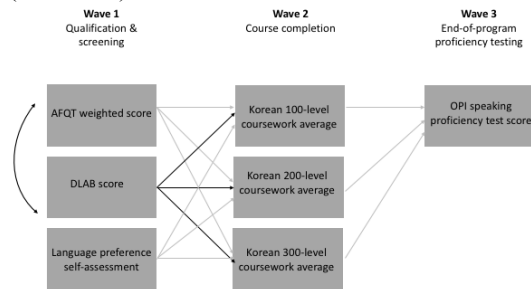


Figure 110. Wave 2 to 3 Korean Speaking (Observed)

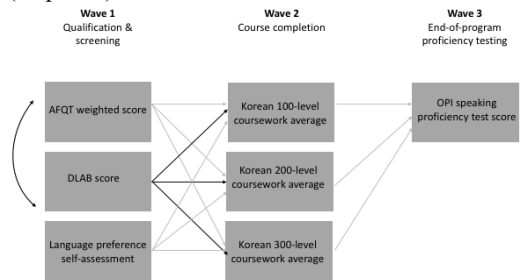


Figure 111. Wave 2 to 3 Korean Speaking (Imputed)

As can be gleaned from Figure 106 through Figure 111, inconsistent patterns are observed in the development of speaking proficiency for the Arabic, Chinese, and Korean languages for both the observed and imputed datasets. Table 54 through Table 55 below detail these results.

Table 54. Arabic Speaking: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 241) and Imputed (n = 411))

Path	Standardized Regression Weight	Standard Error	Significance Level
100-level average coursework to OPI Speaking	(Observed) N/A (Imputed) 0.190	(Observed) N/A (Imputed) 1.428	Observed (N/A) 7 significant pathways
300-level average coursework to OPI Speaking	(Observed) 0.241 (Imputed) 0.371	(Observed) 1.470 (Imputed) 1.250	(Observed) $p < 0.05$ 10 significant pathways

Table 55. Chinese Speaking: Wave 2 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161))

Path	Standardized Regression Weight	Standard Error	Significance Level
200-level average coursework to OPI Speaking	(Observed) 0.324 (Imputed) 0.363	(Observed) 3.088 (Imputed) 2.471	(Observed) $p < 0.05$ 8 significant pathways
300-level average coursework to OPI Speaking	(Observed) N/A (Imputed) -0.335	(Observed) N/A (Imputed) 2.816	(Observed) N/A 8 significant pathways

As shown in Table 54 through Table 55 above, for the Arabic and Chinese languages, outcomes from the speaking skill analyses were not consistent across observed and imputed datasets. This finding suggests that the development of speaking proficiency across the Arabic, Chinese, and Korean Category IV languages differs from reading and listening skill development, which yielded similar patterns of development across languages for both the observed and imputed datasets. As mentioned above, it also may suggest instability in the measurement of the speaking skill itself, since the OPI is a human-rated performance-based assessment. Results of significance testing of common path weights between observed and imputed path weights are reported in Table 56.

Table 56. Differences in observed versus imputed correlations

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Arabic 300-level average course outcomes to OPI Speaking (Observed and Imputed)	-1.762	Not significant
Chinese 200-level average course outcomes to DLPT Listening (Observed and Imputed)	-0.341	Not significant

The fact that none of the average 100-, 200-, and 300-level coursework outcomes predict OPI speaking proficiency outcomes for the Korean language, even though all learners within the imputed dataset would have met established criteria, is particularly counterintuitive, suggesting that there are unaccounted for contextual or programmatic variables that likely contribute to the development of Korean speaking proficiency.

Study 2 Results: Wave 1 to Wave 3 Full Path Analyses: Reading

Research Question 4c (RQ4c): For languages grouped within the same category, *are homogenous patterns of initial language acquisition observed across languages as learners progress through the DLIFLC program of study?*

The current section will discuss differences in path-analytic outcomes of Wave 1 exogenous predictor variables to Wave 3 endogenous variables for both the observed and imputed path-analytic models separately by skill. For the imputed datasets, beginning with the predictive influence of Wave 1 AFQT weighted scores, DLAB scores, and Language Preference Self-Assessment scores on DLPT reading outcomes, just two significant causal pathways were found. For the Chinese language, AFQT weighted scores and Language Preference Self-Assessment scores predicted DLPT reading outcomes. Consistent with what was found for the observed dataset, this finding suggests that the Verbal Expression, Mathematical Knowledge, and Arithmetic Knowledge, which

compose the AFQT weighted scores and likely require considerable reading comprehension skills, positively predict end-of-program DLPT reading outcomes. For the Korean language, a negative causal pathway was found between Language Preference Self-Assessment scores and DLPT reading outcomes. This finding is consistent with the observed path-analytic outcomes associated with the observed Korean language records. Figure 112 through Figure 117 visually depict these findings.

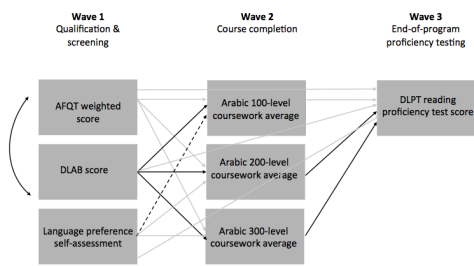


Figure 112. Full Arabic Path Model Reading (Observed: n = 241, CFI = 0.999, RMSEA = 0.034)

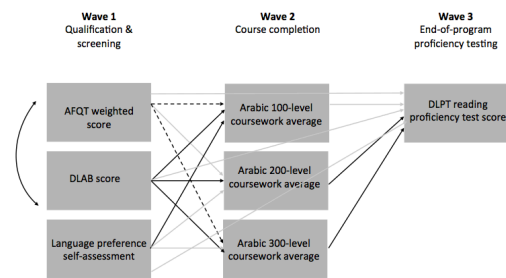


Figure 113. Full Arabic Path Model Reading (Imputed: n = 411, CFI = 1.00, RMSEA = 0.000 (all models))

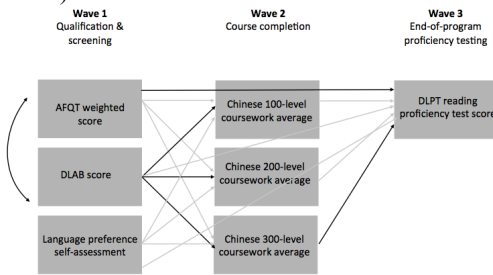


Figure 114. Full Chinese Path Model Reading (Observed: n = 98, CFI = 1.000, RMSEA = 0.000)

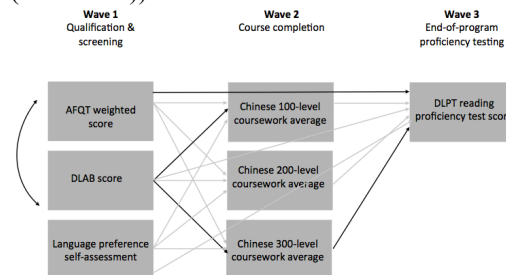


Figure 115. Full Chinese Path Model Reading (Imputed: n = 161, CFI = 1.000, RMSEA = 0.000)

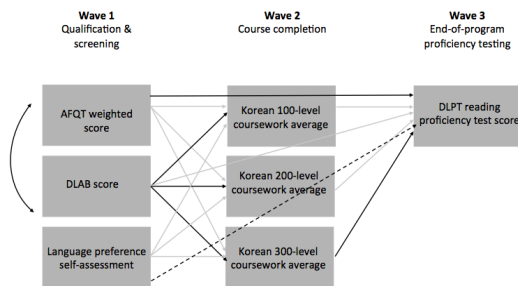


Figure 116. Full Korean Path Model Reading (Observed: n = 75, CFI = 1.000, RMSEA = 0.000)

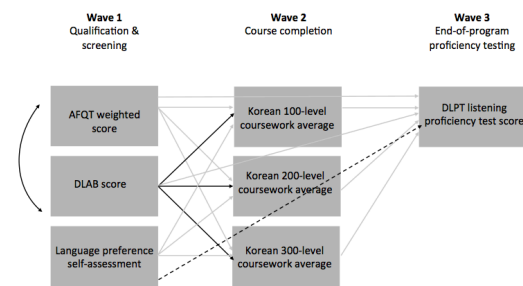


Figure 117. Full Korean Path Model Reading (Imputed: n = 118, CFI = 1.000, RMSEA = 0.000)

Visual inspection of the observed and imputed models shows little within-language variation between models. The Arabic and Chinese observed and imputed models are identical while the Korean imputed model lacks the significant causal pathway found from AFQT scores to DLPT reading outcomes in the observed model. This indicates that, for the Korean language, AFQT score outcomes may only be robust enough to predict DLPT reading outcomes for high-potential learners. In order to determine the number of learners who would have met criterion DLPT reading outcomes for the imputed datasets, representative imputation models were selected to allow for the calculation of associated descriptive statistics for each language.⁶⁰ Models projected that 62.7% of Arabic learners, 65.7% of Chinese learners, and 100% of Korean learners met DLPT reading proficiency criterion. Table 57 and Table 58 below detail the results of outcomes visualized above.

Table 57. Chinese Reading: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 98) and Imputed (n = 161))

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT to DLPT Reading	(Observed) 0.295 (Imputed) 0.154	(Observed) 0.245 (Imputed) 0.342	p < 0.001 7 significant pathways

Table 58. Korean Reading: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118))

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT to DLPT Reading	(Observed) 0.209 (Imputed) -0.157	(Observed) 0.259 (Imputed) 0.254	(Observed) p < 0.05 3 significant pathways
Language Preference Self-Assessment to DLPT Reading	(Observed) -0.185 (Imputed) -0.169	(Observed) 0.805 (Imputed) 0.740	(Observed) p < 0.05 6 significant pathways

⁶⁰ For the Arabic language, imputation Model 3 was selected for the reading and listening skills and Model 10 was selected for the speaking skill. For the Chinese language, imputation Model 10 was selected for the reading skill and Model 8 was selected for the listening and speaking skills. For the Korean reading skill, no exact imputation model was identified. Model 2 was the closest fitting imputation model for Korean reading skill. Exact matches for the Korean listening and speaking skills were identified: Models 1 and 3 were selected, respectively. To examine the path analytic outcomes across all ten imputed datasets, the author can be contacted upon request.

Results of significance testing of common path weights are reported in Table 59.

Table 59. Differences in observed versus imputed correlations

Path Comparison	z-value (Reference Value +/- 1.96)	Significance
Chinese AFQT to DLPT Reading (Observed and Imputed)	1.146	Not significant
Korean Language Preference Self- Assessment to DLPT Reading (Observed and Imputed)	-0.110	Not significant

As shown in Table 59, no significant differences in common path weights were found for the Chinese and Korean languages. To ascertain the total amount of model variability accounted for by each language, Table 60 displays the results of the median squared multiple correlation indices across the 10 imputed datasets for the Arabic, Chinese, and Korean models. For ease of reference, the squared multiple correlation indices for the observed datasets are also included. As shown below, the Arabic imputed dataset accounts for substantially more variability than the model containing only observed learner records. This indicates that including learners with a wider range of academic achievement and DLPT reading outcomes strengthens the overall path-analytic model.

Table 60. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean reading models

Language	Endogenous Variable	% Model Variability Accounted For	% of Variability Remaining
Arabic	100-level course outcomes	(Observed) 14.8% (Imputed) 16.2%	(Observed) 85.2% (Imputed) 83.8%
	200-level course outcomes	(Observed) 9.2% (Imputed) 12.4%	(Observed) 90.8% (Imputed) 87.6%
	300-level course outcomes	(Observed) 9.8% (Imputed) 11.1%	(Observed) 90.2% (Imputed) 88.9%
	DLPT Reading (Full Model)	(Observed) 33.9% (Imputed) 51.2%	(Observed) 66.1% (Imputed) 33.9%
Chinese	100-level course outcomes	(Observed) 0.9% (Imputed) 8.2%	(Observed) 99.1% (Imputed) 91.8%
	200-level course outcomes	(Observed) 7.5% (Imputed) 5.1%	(Observed) 92.5% (Imputed) 94.9%
	300-level course outcomes	(Observed) 11.4% (Imputed) 6.6%	(Observed) 88.6% (Imputed) 93.4%
	DLPT Reading (Full Model)	(Observed) 32.3% (Imputed) 30.3%	(Observed) 67.7% (Imputed) 69.7%
Korean	100-level course outcomes	(Observed) 13.0% (Imputed) 7.4%	(Observed) 87.0% (Imputed) 92.6%
	200-level course outcomes	(Observed) 11.2% (Imputed) 6.0%	(Observed) 88.8% (Imputed) 94.0%
	300-level course outcomes	(Observed) 16.5% (Imputed) 12.4%	(Observed) 83.5% (Imputed) 87.6%
	DLPT Reading (Full Model)	(Observed) 35.8% (Imputed) 24.5%	(Observed) 64.2% (Imputed) 75.5%

As shown above, for the Arabic language, the imputed model accounted for substantially more overall model variability than the observed model, suggesting that the increased sample size associated with the Arabic dataset strengthened the overall predictive model. For the Chinese language, comparable amounts of model variability are found for the Chinese observed and imputed datasets, suggesting that the increase in sample size and imputed missing values did not necessarily increase the strength of the path-analytic model. For the Korean language, the imputed model accounted for less overall model variability than the observed model. This finding can likely be attributable to the data used as input into the imputation model. That is, since the imputed models rely on the observed learner records as input into the creation of the imputed models, and 100% of

learners in the observed model met criterion reading outcomes, predicted imputed outcomes are restricted by the observed variation from each observed dataset. The new cases generated for the Korean imputation dataset represent GPA outcomes for learners that typically would not have been allowed to progress through the DLIFLC Korean program of study. These cases simply introduce variability into a model that is constrained by 100% of observed cases meeting criterion DLPT reading outcomes.

Study 2 Results: Wave 1 to Wave 3 Full Path Analyses: Listening

For the listening skill, just two significant causal pathways were established for the imputed datasets. For both the Arabic and Korean languages, DLAB scores were found to predict DLPT listening outcomes. The path-analytic outcomes for both the observed and imputed datasets are visualized in Figure 118 through Figure 123, below.

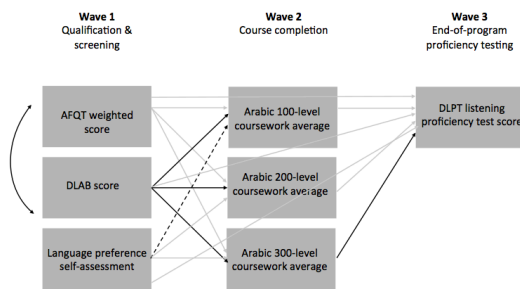


Figure 118. Full Arabic Path Model Listening (Observed: $n = 241$, CFI = 0.999, RMSEA = 0.034)

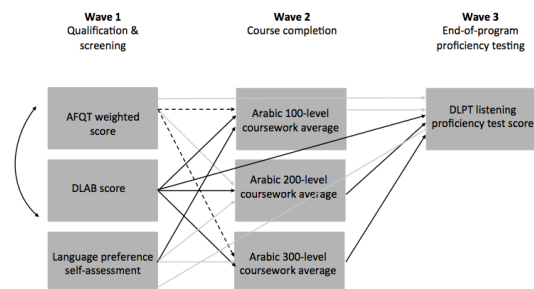


Figure 119. Full Arabic Path Model Listening (Imputed: $n = 411$, CFI = 1.00, RMSEA = 0.000 (all models))

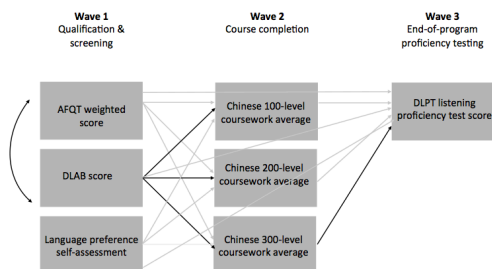


Figure 120. Full Chinese Path Model Listening (Observed: $n = 98$, CFI = 1.000, RMSEA = 0.000)

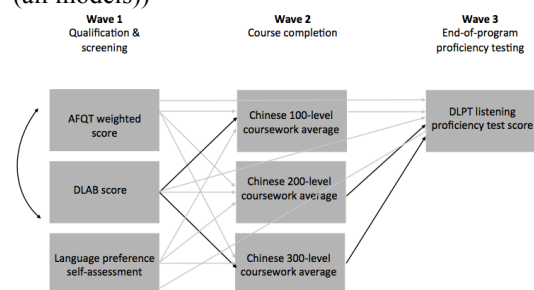


Figure 121. Full Chinese Path Model Listening (Imputed: $n = 161$, CFI = 1.000, RMSEA = 0.000)

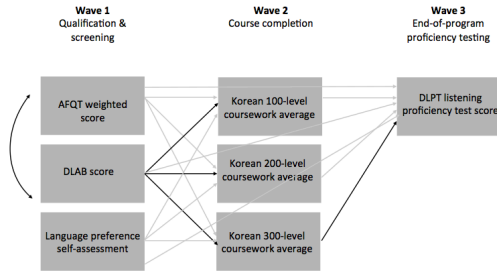


Figure 122. Full Korean Path Model Listening (Observed: n = 75, CFI = 1.000, RMSEA = 0.000)

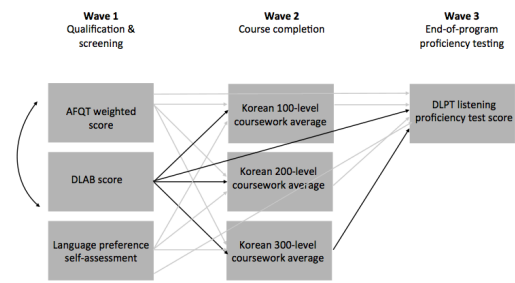


Figure 123. Full Korean Path Model Listening (Imputed: n = 118, CFI = 1.000, RMSEA = 0.000)

The significant causal pathways found between DLAB scores and DLPT listening outcomes for the Arabic and Korean languages suggests that, while DLAB is strongest in predicting average 100-, 200-, and 300-level coursework success for both the observed and imputed datasets, it is also robust enough to predict DLPT listening score outcomes when increased n-sizes are included in the model.

Table 61. Arabic Listening: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 241 and Imputed (n = 411))

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to DLPT Listening	(Observed) N/A (Imputed) 0.098	(Observed) N/A (Imputed) 0.635	(Observed) N/A 8 significant pathways

Table 62. Korean Listening: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118))

Path	Standardized Regression Weight	Standard Error	Significance Level
DLAB to DLPT Listening	(Observed) N/A (Imputed) 0.174	(Observed) N/A (Imputed) 0.967	(Observed (N/A) 7 significant pathways

As shown in Table 61 and Table 62, although smaller in magnitude than the path weights found for average 100-, 200-, and 300-level average coursework outcomes, DLAB scores were still found to significantly predict DLPT listening test score outcomes for the Arabic and Korean imputed datasets. This finding was previously hidden when only observed

learner data was modeled for these languages. For the listening skill, 67.3% of Arabic learners, 95.7% of Chinese learners, and 90.0% of Korean learners would have met DLPT listening criterion. Outcomes from squared multiple correlations reveal a substantial increase in the amount of variability accounted for in the Arabic imputed model versus the model containing only observed learner records. This indicates that the learner variability within the imputed Arabic dataset strengthens the overall model.

Table 63. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean listening models

Language	Endogenous Variable	% Model Variability Accounted For	% of Variability Remaining
Arabic	100-level course outcomes	(Observed) 14.8% (Imputed) 16.2%	(Observed) 85.2% (Imputed) 83.8%
	200-level course outcomes	(Observed) 9.2% (Imputed) 12.4%	(Observed) 90.8% (Imputed) 87.6%
	300-level course outcomes	(Observed) 9.8% (Imputed) 11.1%	(Observed) 90.2% (Imputed) 88.9%
	DLPT Listening (Full Model)	(Observed) 33.9% (Imputed) 55.4%	(Observed) 66.1% (Imputed) 44.6%
Chinese	100-level course outcomes	(Observed) 0.9% (Imputed) 8.2%	(Observed) 99.1% (Imputed) 91.8%
	200-level course outcomes	(Observed) 7.5% (Imputed) 5.1%	(Observed) 92.5% (Imputed) 94.9%
	300-level course outcomes	(Observed) 11.4% (Imputed) 6.6%	(Observed) 88.6% (Imputed) 93.4%
	DLPT Listening (Full Model)	(Observed) 32.3% (Imputed) 18.3%	(Observed) 67.7% (Imputed) 81.7%
Korean	100-level course outcomes	(Observed) 13.0% (Imputed) 7.4%	(Observed) 87.0% (Imputed) 92.6%
	200-level course outcomes	(Observed) 11.2% (Imputed) 6.0%	(Observed) 88.8% (Imputed) 94.0%
	300-level course outcomes	(Observed) 16.5% (Imputed) 12.4%	(Observed) 83.5% (Imputed) 87.6%
	DLPT Listening (Full Model)	(Observed) 35.8% (Imputed) 25.5%	(Observed) 64.2% (Imputed) 74.5%

As shown in

Table 63, for Chinese and Korean, the squared multiple correlations are reduced for the imputed versus observed datasets. Similar to what was found for the reading skill, this finding is likely attributable to the fact that there is less overall end-of-program variability modeled as input from the observed datasets. For the Chinese and Korean

languages, at least 89% of learners were projected to meet DLPT listening criterion outcomes for the observed (90.8% and 88.0%, respectively) and imputed (95.7% and 90.7%, respectively) datasets.

Study 2 Results: Wave 1 to Wave 3 Full Path Analyses: Speaking

The final analysis in the current investigation compares Wave 1 and Wave 3 path-analytic outcomes between observed and imputed datasets for the speaking skill. Across the Arabic, Chinese, and Korean models, just one significant causal pathway was established: For the Korean language, DLAB scores were found to significantly predict OPI speaking proficiency test scores. Figure 124 through Figure 129 below visualize these findings.

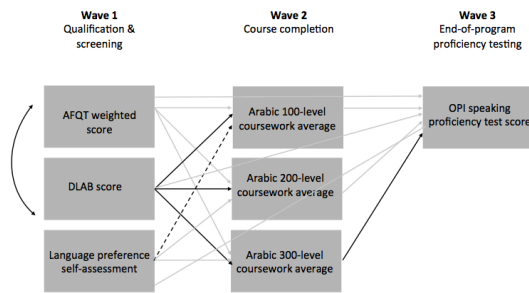


Figure 124. Full Arabic Path Model Speaking (Observed: $n = 241$, CFI = 0.999, RMSEA = 0.034)

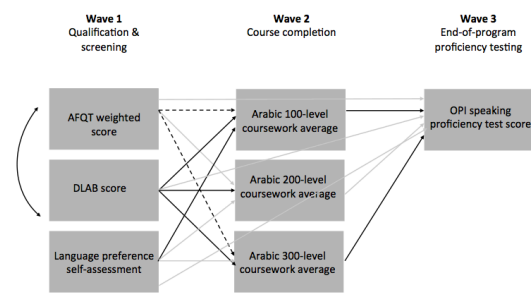


Figure 125. Full Arabic Path Model Speaking (Imputed: $n = 411$, CFI = 1.00, RMSEA = 0.000 (all models))

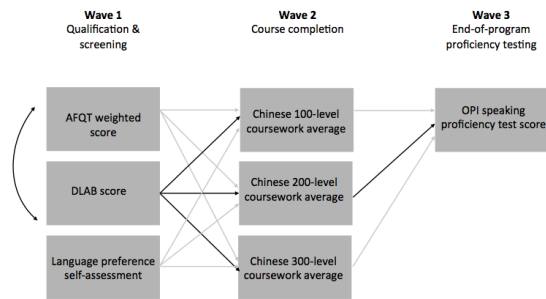


Figure 126. Full Chinese Path Model Speaking (Observed: $n = 98$, CFI = 1.000, RMSEA = 0.000)

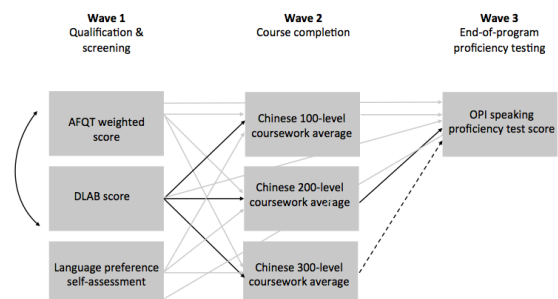


Figure 127. Full Chinese Path Model Speaking (Imputed: $n = 161$, CFI = 1.000, RMSEA = 0.000)

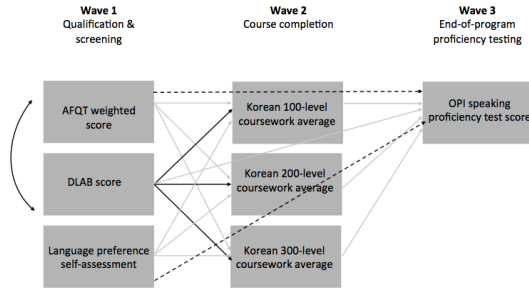


Figure 128. Full Korean Path Model Speaking (Observed: n = 75, CFI = 1.000, RMSEA = 0.000)

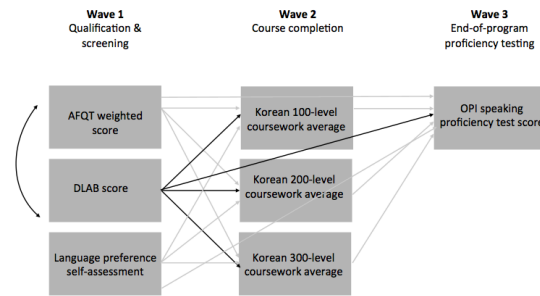


Figure 129. Full Korean Path Model Speaking (Imputed: n = 118, CFI = 1.000, RMSEA = 0.000)

As shown above, none of the Wave 1 predictor variables were found to have a significant influence on Wave 3 OPI speaking proficiency outcomes for the Arabic or Chinese languages in either the observed and imputed models. For the speaking skill, 76.5% of Arabic learners, 92.5% of Chinese learners, and 100% of Korean learners would have met OPI speaking criterion thresholds. The findings above are detailed in Table 64 below.

Table 64. Korean Speaking: Wave 1 to Wave 3 path analytic outcomes (Observed (n = 75) and Imputed (n = 118))

Path	Standardized Regression Weight	Standard Error	Significance Level
AFQT to OPI Speaking	(Observed) -0.217 (Imputed) -0.137	(Observed) 0.147 (Imputed) 0.122	(Observed) p < 0.05 2 significant pathways
Language Preference Self-Assessment to OPI Speaking	(Observed) -0.236 (Imputed) -0.141	(Observed) 0.457 (Imputed) 0.351	(Observed) p < 0.05 2 significant pathways
DLAB to OPI Speaking	(Observed) N/A (Imputed) 0.259	(Observed) N/A (Imputed) 0.376	(Observed) N/A 10 significant pathways

As shown above, the Korean language shows the most variability in path-analytic outcomes between the observed and imputed datasets. For the observed dataset, negative causal pathways were found from AFQT weighted scores and Language Preference Self-Assessment scores to OPI speaking proficiency test score outcomes. For the imputed dataset, these relationships were absent and were replaced by a positive causal pathway between DLAB scores and OPI proficiency test score outcomes. As displayed in Table 65

below, consistent with DLPT reading and listening score outcomes, results of the squared multiple correlation analyses show that the Arabic imputed model accounts for substantially more model variability than the model containing only observed learner records, also suggesting that including both high- and low-achieving learners within the path analysis strengthens the overall predictive model.

Table 65. Squared Multiple Correlation Indices for the Arabic, Chinese, and Korean speaking models

Language	Endogenous Variable	% Model Variability Accounted For	% of Variability Remaining
Arabic	100-level course outcomes	(Observed) 14.8% (Imputed) 16.2%	(Observed) 85.2% (Imputed) 83.8%
	200-level course outcomes	(Observed) 9.2% (Imputed) 12.4%	(Observed) 90.8% (Imputed) 87.6%
	300-level course outcomes	(Observed) 9.8% (Imputed) 11.1%	(Observed) 90.2% (Imputed) 88.9%
	OPI Speaking (Full Model)	(Observed) 33.9% (Imputed) 49.9%	(Observed) 66.1% (Imputed) 50.1%
Chinese	100-level course outcomes	(Observed) 0.9% (Imputed) 8.2%	(Observed) 99.1% (Imputed) 91.8%
	200-level course outcomes	(Observed) 7.5% (Imputed) 5.1%	(Observed) 92.5% (Imputed) 94.9%
	300-level course outcomes	(Observed) 11.4% (Imputed) 6.6%	(Observed) 88.6% (Imputed) 93.4%
	OPI Speaking (Full Model)	(Observed) 32.3% (Imputed) 30.6%	(Observed) 67.7% (Imputed) 69.4%
Korean	100-level course outcomes	(Observed) 13.0% (Imputed) 7.4%	(Observed) 87.0% (Imputed) 92.6%
	200-level course outcomes	(Observed) 11.2% (Imputed) 6.0%	(Observed) 88.8% (Imputed) 94.0%
	300-level course outcomes	(Observed) 16.5% (Imputed) 12.4%	(Observed) 83.5% (Imputed) 87.6%
	OPI Speaking (Full Model)	(Observed) 35.8% (Imputed) 23.0%	(Observed) 64.2% (Imputed) 77.0%

As can be gleaned from Table 65 above, the squared multiple correlation indices for the Chinese language observed and imputed models account for comparable percentages of model variability, while a substantial lower amount of model variability is accounted for by the Korean language imputed dataset than the observed dataset. These findings could also be attributable to the restricted range associated with the OPI speaking proficiency outcomes. That is, because 100 % of learners met OPI Korean speaking criterion

outcomes in the observed data, the imputation model was also constrained by the lack of variability in speaking test outcomes. The imputed model included learners with lower achievement-related outcomes than the observed dataset, but was forced to model all learners as meeting speaking criterion outcomes. This injects additional uncertainty into the overall model, which is reflected in the lower squared multiple correlation indices than the observed data. Consistent with what was found for the reading and listening skills, the imputed cases for the Chinese and Korean languages likely introduce variability into the path-analytic model, thus decreasing the squared multiple correlation indices for these skills and languages.

Discussion: Study 2

Research Question 4a-c (RQ4a-c): For languages grouped within the same category, are similar patterns of individual differences in general aptitude, language-specific aptitude, and motivation observed in the prediction of learners' success as they progress through coursework for observed versus imputed datasets?

The purpose of Study 2 was to examine the potential hidden effects of non-random attrition across learners and to determine the extent to which the observed outcomes from Study 1 were influenced by the systematic, casewise deletion of data associated with students who were “re-linguaged” or “re-cycled” within the DLIFLC program of study. Results of the multiple imputation procedure revealed that the DLAB variable played a consistent and robust role in predicting 100-, 200-, and 300-level average coursework outcomes across the Arabic, Chinese, and Korean languages for both the observed and imputed datasets. In contrast, the AFQT and Language Preference Self-Assessment scores exhibited variability in their predictive influences on 100-, 200-, and 300-level average coursework outcomes across languages. For Chinese and Korean, AFQT and

Language Preference Self-Assessment variables were not found to predict 100-, 200-, or 300-level average coursework outcomes. For Arabic, AFQT scores were found to negatively predict 100- and 300-level average coursework outcomes for the imputed dataset but not the observed dataset, and Language Preference Self-Assessment scores were found to negatively predict 100-level average coursework outcomes for the observed dataset, but to positively predict 100-level average coursework outcomes for the imputed dataset. Findings such as these underscore the importance of accounting for the nature and pattern of missing data when building complex models. While AFQT outcomes were not found to predict 100-, 200-, or 300-level average coursework outcomes for any of the observed models, they were found to negatively predict 100- and 300-level average coursework outcomes for the imputed Arabic models. This can likely be attributable to the lower GPA outcomes associated with the imputed dataset (100-level average coursework outcomes = 3.27 for the observed data and 3.05 for the imputed data; 300-level average coursework outcomes = 3.35 for the observed data and 3.00 for the imputed data). It could also be argued that although the ASVAB is used to determine enlistment eligibility, to assign applicants to military jobs, and to aid students in career explorations, the components of the ASVAB used to calculate an AFQT weighted score (Mathematics Knowledge, Arithmetic Reasoning, and Verbal Expression) are not robust enough to causally predict L2 language development for Category IV languages. The contradictory findings associated with the Language Preference Self-Assessment score and 100-level average coursework outcomes across models suggests the unreliability of using this one-question instrument as a predictor variable. Previous research by Lett (1990) found most correlations between language preference self-assessments to be non-

significant, and to vary substantially across target languages and subskills, as well as from year-to-year, likely attributable to the changing economic and political climates that engender the need for trained linguists within a particular language. These differential findings provide evidence for considering the importance of accounting for the nature and purpose of the data serving as input into path-analytic models.

When examining the causal patterns of the observed and imputed datasets between 100-, 200-, and 300-level average coursework outcomes and end-of-program proficiency test score outcomes, a great deal of coherence can be found across languages, skills, and models. With few exceptions, Arabic, Chinese, and Korean exhibit almost identical patterns in the significant causal pathways established between the observed and imputed datasets for the reading and listening skills, with 300-level coursework outcomes strongly predicting DLPT outcomes across all languages. This finding indicates that, although lower 100-, 200-, and 300-level average coursework outcomes were associated with the imputed datasets, the imputed coursework variables were robust enough to predict end-of-program DLPT proficiency test score outcomes, providing validity evidence for the imputation procedure for the reading and listening skills. However, for the speaking skill, comparisons in path-analytic outcomes between the observed and imputed datasets across languages yielded inconsistent patterns in Wave 2 to Wave 3 proficiency development. For Arabic, a causal relationship was found between 100-level average coursework outcomes and OPI speaking outcomes, for Chinese, an unanticipated negative causal relationship was found between 300-level average coursework outcomes and OPI speaking outcomes, and for Korean, no significant causal relationships were found between average coursework outcomes and OPI speaking outcomes. As alluded to

in the discussion of Wave 2 to Wave 3 speaking results for Study 1, these findings can potentially be attributed to the instability of the human-scored, performance-based OPI ratings. While the DLPT reading and listening proficiency tests are computer-scored and administered in standardized test administration settings, the OPI is human-rated and administered by DLIFLC personnel, some of whom may be familiar with the examinee, potentially influencing (either negatively or positively) their OPI ratings.

The path-analytic outcomes for Korean are particularly unexpected. For both the observed and imputed datasets, 100% of learners met (or were predicted to meet) OPI speaking criterion outcomes even though no significant causal relationships were found between 100-, 200-, or 300-level average coursework outcomes and OPI speaking proficiency test scores. This finding indicates that additional contextual variables, such as established learning communities, curricular practices, and more robust learner motivation variables would likely shed additional light on the currently hidden influence of learner context on observed outcomes. As Norris (2016) states,

Given an adequately specified program logic model, it may then be possible to set about collecting evidence regarding not only (a) whether the program actually achieves targeted outcomes, but more importantly, (b) whether all dimensions of the model are realized and implemented as intended, (c) which factors in particular may be moderating outcomes achievement (including possibly factors external to the program *per se*), (d) whether or not the model as a complex, interactive whole is viable (p. 177).

Building on Norris' recommendation concerning adequate model specification, the lack of significant causal pathways established for the Korean speaking skill for both the observed and imputed datasets could indicate that the path-analytic model for this language and skill is not viable.

In terms of path-analytic model comparisons between languages for the AFQT, DLAB, and Language Preference Self-Assessment scores on end-of-program proficiency-testing outcomes, AFQT scores were found to consistently predict Chinese reading outcomes for both the observed and imputed datasets. For Korean, AFQT scores were found to predict DLPT reading outcomes for the observed, but not imputed dataset. Further, for Korean, Language Preference Self-Assessment scores were found to negatively predict DLPT reading outcomes. For the listening skill, DLAB scores were found to significantly predict DLPT listening outcomes for the Arabic and Korean imputed datasets, but not the observed datasets. Lastly, for the speaking skill, no AFQT, DLAB, or Language Preference Self-Assessment outcomes were found to significantly predict OPI speaking outcomes for Arabic and Chinese. However, for Korean, AFQT scores and Language Preference Self-Assessment scores were found to negatively predict OPI speaking outcomes for the observed dataset while DLAB scores were found to positively predict OPI speaking outcomes for the imputed dataset. The variability observed across the path-analytic models for each language and skill may be attributable to the variability of the data contained within the observed learner records used to inform the imputation procedure. For the observed Chinese and Korean data, 97% to 100% of learners met DLPT reading and speaking outcomes, thereby constraining the end-of-program outcomes for the imputed path-analytic models. Since these models contained imputed learners who were predicted to have lower average 100-, 200-, and 300-level coursework outcomes, the multiple correlation indices for the Chinese and Korean reading and speaking skills were found to be lower for the imputed datasets (containing more learners) than the observed datasets.

Overall, Arabic language appeared to be most amenable to the path-analytic imputation procedure, thus validating its use, with caution, when working with systematic patterns of missing data in the L2 instructional context. For the Arabic imputed datasets, 62% to 76% of learners were modeled to meet end-of-program proficiency test score criteria, indicative of the variability within the observed learner data serving as input into the imputed models. Across all skills, the squared multiple correlation indices were higher for the Arabic imputed datasets than the observed datasets (51.2% for reading, 55.4% for listening, and 49.9% for speaking), suggesting that including both high- and low-achieving learners within a path-analytic framework strengthens the overall predictive model across all skills. The ceiling effects associated with Chinese and Korean were found to constrain the path analytic imputation procedure for the reading and speaking skills. The imputation procedure itself could potentially be improved by incorporating other, achievement-related, rather than proficiency-based outcomes into the imputation model, potentially allowing for more variability in learner outcomes to be modeled, thereby increasing the predictive power of the imputed models. While the measured variables within the current model account for about one-third (for the observed datasets) or about 50% (for the Arabic imputed datasets) of total model variability, consistent with Frechtling's (2007) perspective, additional input is needed from the stakeholder community to assess, value, and interpret the observed findings.

Conclusions and Directions for Future Research

Although the LDC framework has been almost universally adopted across USG agencies over the past 60 years, there is very little research that empirically examines the patterns of proficiency development for languages grouped within the same category. While early efforts to categorize language within a hierarchical framework were not made capriciously, and were based on years of subject matter experts' insights and observations, the need for robust empirical evidence concerning how well the categorization functions is essential, particularly given the high-stakes decisions associated with its application. At its most basic level, the categorization system works under the assumption that the same amount of input (in terms of weeks of intensive language training) yields the same amount of output (in terms of minimum DLPT and OPI proficiency test score criteria) for languages grouped within the same category. Lacking within the LDC framework is robust empirical evidence to support this assumption, despite the significant investment of resources across all levels of stakeholders, from tax payer dollars, to the DLIFLC instructional setting, to learners, to teachers, to curriculum developers, and to language testers, just to name a few. The current investigation examined individual patterns in foreign language achievement and how both cognitive (e.g., general and language-specific aptitude) and non-cognitive (e.g., language preference self-assessment scores) individual difference variables jointly influenced initial acquisition proficiency development for three Category IV languages at the DLIFLC (Arabic, Chinese, and Korean). In order to compare the coherence in patterns of proficiency development across languages, a contrastive-analytic approach was employed within the framework of a logic model—a novel method applied in the L2

instructional context, within which observed outcomes between waves of proficiency development were statistically modeled. Data mining can be effective for observing longitudinal trends within a given program of interest, although it can be imperfect in its ability to account for programmatic inputs that are not systematically calculated or documented. In order to examine the coherence with which both cognitive and non-cognitive variables contribute to the development of foreign language achievement and proficiency score outcomes, a sample of course achievement and testing records, as well as other aptitude- and personality-related records was aggregated from four systematically maintained databases to create observed and imputed learner models. If languages grouped within the same category truly require the same amount of instructional input to meet end-of-program proficiency test score criteria, statistical examination of proficiency development patterns would be invariant across languages (Arabic, Chinese, and Korean) and skill modalities (reading, listening, and speaking).

Overall, a great deal of coherence was found in the development of Arabic, Chinese, and Korean foreign language proficiency, providing initial validity evidence for the current LDC framework. Consistent with Masters (2016), across all languages, skills, and observed and imputed datasets, the Wave 1 DLAB variable played a robust role in predicting 100-, 200-, and 300-level average coursework outcomes. This finding could also be attributable to an alignment with the analytical components of the DLAB instrument and DLIFLC teaching practices. That is, the curricular focus on functional language proficiency at DLIFLC aligns well with the analytical nature of DLAB. The overall generalizability of the causal influence of DLAB outcomes to 100-, 200-, and 300-level course achievement outcomes would likely be constrained should DLIFLC

instructional practices shift. Also consistent with Masters (2016), Wave 3 300-level coursework averages consistently predicted end-of-program proficiency test score outcomes across all skills. This finding suggests a strong alignment, across all languages and skills, between 300-level instructional content and language proficiency test content.

Noted exceptions to the patterns of proficiency development noted above are the path-analytic outcomes found for the observed and imputed datasets for the Korean speaking skill. Although 100% of learners met (or were modeled to have met) OPI speaking criterion outcomes, no significant causal pathways were found between 100-, 200-, or 300-level average coursework outcomes and OPI speaking proficiency test score outcomes. The lack of achievement-related significant causal pathways leading to OPI test score outcomes suggests that the Korean speaking program of study may not have evolved logically for the sample of learners associated with the current analysis. That is, other external program factors, unrelated to Korean coursework, are likely influencing the development of Korean speaking proficiency. Atypical patterns could also shed some light on findings previously reported by Bloomfield et al. (2012), who, in their examination of changes in proficiency test scores over time, found that, while reading and listening skills exhibited overall patterns of improvement over time, speaking skills exhibited both a higher incidence and faster rate of loss. Their analysis involved the testing and training records of approximately 1,100 DoD language analysts who had already reached minimum DLPT and OPI proficiency test score criteria, the majority of whom were former DLIFLC graduates. The overall decline in OPI speaking proficiency test scores observed by Bloomfield et al. (2012) may be related to the asymmetrical pattern of speaking proficiency development for the Korean language in that it brings into

question the stability of OPI test score outcomes for this sample. That is, as the pressure placed on instructors by program administrators to meet established proficiency test score outcomes becomes higher and higher, the incentives placed on instructors to meet end-of-program criteria also increases, potentially to a point where it can be counter-productive. This finding could suggest that military linguists enter the workforce with superficial, rather than sustained, levels of language proficiency. To systematically examine this supposition, future research should link initial acquisition training at the DLIFLC with career-long testing and training records maintained by the DoD.

The findings associated with the current investigation corroborated the results of Masters' (2016) research which revealed striking differences in the development of proficiency between learners of languages grouped in different difficulty categories, namely Spanish and Arabic. The difference in observed proficiency acquisition patterns for languages grouped within different difficult categories (Arabic and Spanish) together with the findings associated with the current investigation, in which a great deal of coherence was established for languages grouped within the same category (Arabic, Chinese and Korean) corroborates the DLIFLC policy requiring higher aptitude scores for Category IV languages than Category I languages.

It is hoped that the current investigation can serve as a benchmark from which evidence-based comparisons can be made. As the DLIFLC considers increasing its end-of-program graduation criteria (from 2/2/1+ to 2+/2+/2) and plans on fully operationalizing a new version of the DLAB (DLAB 2, which contains both cognitive and non-cognitive measures in its estimation of language-learning aptitude), outcomes from future research investigations, replicating (and likely improving upon) the path-

analytic procedure, can be compared with the findings established from the current investigation. One might expect to see increased variability in the path-analytic outcomes for the Arabic, Chinese, and Korean languages, as learners and instructors will likely implement a variety of strategies and instructional techniques to meet the updated proficiency-testing criteria. Of particular interest for future research efforts is the modeling of OPI speaking proficiency outcomes for Korean. The fact that 100% of learners met OPI speaking criterion suggests either instability in the measurement of the OPI speaking skill or that the 1+ proficiency criterion standard is too low for this language. The introduction of a higher speaking skill standard would likely introduce more variability in the Korean speaking outcomes, since it logically follows that fewer learners would meet the increased standard within the allotted instructional time. After the DLAB 2 is fully operationalized, one might expect to see even more robust significant causal pathways established between DLAB 2 and 100-, 200-, and 300-level average coursework outcomes, particularly since the DLAB 2 includes previously unaccounted-for contextual learner constructs, such as motivation and personality. While the DLAB was found to consistently predict 100-, 200-, and 300-level coursework outcomes across Arabic, Chinese, and Korean, much stronger path weights would be expected between the DLAB 2 and 100-level average coursework outcomes, since it is likely that the previously unmodeled contextual variables contained within the DLAB 2 are highly influential in the identification of learners with favorable language-learning strategies that predict initial learning success, which, in turn, subsequently sustain them throughout their language studies. Future research should replicate the analyses within the current investigation with a more recent sample of DLIFLC learners and build on

Masters' (2016) research to compare the development of Category IV foreign language skills with languages grouped in other Categories.

The current data involved learners from a 2009 cohort of study, a time during which both program administrators and language learners alike were adjusting to an updated testing form referred to as the DLPT 5. Substantially different from its predecessor, the DLPT IV, which contained scripted, studio-based item specimens, the DLPT 5 incorporated genuine audio and written input from modern sources, such as authentic interviews, real-time conversations in public spaces (inclusive of background noise), podcasts, websites, and email correspondences. In their time series and impact analyses examining the efficacy of foreign language training programs for DLPT test score outcomes, Bloomfield et al. (2016) found a drastic decline in DLPT test scores upon introduction of the DLPT 5 in 2007, followed by an eventual recovery in test scores upon acclimation to the new test format around late 2009/early 2010, depending on the organization and skill modality. As the data used as input into the current investigation was from 2009, and the DLPT 5 was made available for operational use for Arabic, Chinese, and Korean between 2006 and 2007, observed variations in significant DLPT-related path coefficients are likely influenced by instructor and learner adjustments to the new testing formats, as well as to the potential variation in difficulty of the initial DLPT 5 test forms across languages. It is important to note that the current analysis only investigated short-term program outputs within the logic model. Additional research, linking short-term program outputs with long-term program outcomes would provide program administrators with the ability to model longitudinal changes in patterns of proficiency development, from the stage of initial acquisition through long-term, career-

long, proficiency sustainment. Systematic, disciplined analyses such as these, which examine the coherence of acquisition patterns for languages grouped within the same category, could potentially lead to increased efficiency with acquiring a given L2, potentially yielding both short- and long-term cost savings across the DoD.

The use of logic modeling to empirically examine initial L2 acquisition patterns within large-scale instructional contexts can be a helpful framework within which to make implicit theories concerning language teaching and learning explicit. Applying the path-analytic procedure to a logic model, which has been specified a priori, can be useful in establishing empirical program baselines and to examine the convergence of expected and observed patterns. Norris (2016) states, “of course, doing so may be particularly threatening to language teachers, curriculum developers, material designers, and program administrators, as it opens up the very real possibility that they are simply not functioning with any type of program logic in mind, never mind the likelihood that expectations for how programs are functioning will not meet observable realities” (p. 177). However, it is argued that the cost of *not* taking the time to model empirically the expected and observed outcomes within a national, high-stakes L2 instructional program is dire, particularly when the cost of not succeeding or arriving at unstable proficiency levels can be devastating, in terms of tax payer investment, resource investment, and overall foreign-language-enabled personnel readiness.

Limitations

The analyses were limited in seven key ways. First, it is important to note that the sample upon which the analyses were based did not represent a full range of values that one might find in the population at large. By the time learners begin instruction at the DLIFLC, they have already been twice selected, first from their scores on the ASVAB and second from their scores on the DLAB. Further, as noted in Chapter 3, almost all of the initial screening variables exhibited non-normal distributions. As observed by Lett and O'Mara (1990), the restricted range of the sample likely had an impact on the predictive power of both the cognitive and non-cognitive variables for both the observed and imputed models. Second, it is possible that contextual differences, rather than cognitive or non-cognitive differences alone, may have contributed to the observed differences found between the Category IV languages in each set of analyses. That is, as noted in Lett and O'Mara (1990), despite the general homogeneity of the language-learning context at the DLIFLC, it is unlikely that the programs are pedagogically equivalent either within or across Category IV initial acquisition courses. Third, the proposed analyses did not separate DLIFLC course outcome data by skill, but rather, treated each 100-, 200-, and 300-level course outcome as an overall skill-level average. As it is likely that some courses within each level specifically focused on the development of a particular modality, future analyses could separate out these courses by skill modality, rather than grouping all courses together and creating a single average. Fourth, the data used in the investigation were roughly ten years old. While ideally more up-to-date testing and training records should be examined, these types of data are difficult to obtain, requiring a significant investment of time and effort that would likely

delay the completion of the investigation by years. Since no major shifts in policy with respect to the initial acquisition training and testing processes has occurred since 2009, it is argued that the date from which the data were pulled are unlikely to substantially to effect observed outcomes. Fifth, as is the case with all data-mining efforts, the current analyses were constrained by the data made available within existing datasets. All analyses assume an underlying accuracy of institutional data entry and database maintenance over time. Sixth, while the n-sizes associated with the current investigation met minimum criteria for structural equation modeling, it is likely that larger sample sizes, particularly for Chinese and Korean, would increase the robustness of the inferences drawn from the path-analytic procedure. Lastly, the nature of DLPT reading and listening proficiency tests limits the generalizability of the path-analytic findings, particularly for overall predictive influence found for DLAB to 100-, 200-, and 300-, level average coursework outcomes as well as 300-level average coursework averages to end-of-program proficiency test score outcomes. These findings could potentially be an artifact of an alignment between analytically oriented aptitude batteries, L2 instructional practices, and proficiency tests. As shown in Appendix F, DLPT test item specimen tend to require examinees to analyze written or oral input and select the most appropriate item stem (written in English). Working backwards from DLPT test item specimen, DLIFLC curricular content is likely to be reversed-engineered to reflect analytically oriented L2 instructional practices, which also aligns with the nature of the DLAB. The coherence in the significant causal pathways established across Arabic, Chinese, and Korean could simply be an artifact of analytically oriented testing and teaching practices, thereby

restricting the range of possible outcomes within each wave of the path-analytic framework.

Appendices

Appendix A: Overview of the Contrastive Analysis Hypothesis

Inaugural research investigating systematically the cognitive processes through which learners acquire a second language was generally constrained to the descriptive analyses of isolated language features. Theoretical perspectives were dichotomized into two main opposing schools of thought. The first perspective maintained an “essential identity” of learners’ first language (L1) (Jakobovits, 1969; Ervin-Tripp, 1974; Burt and Dulay, 1975). The term essential identity implied that the structural characteristics of learners’ L1 had no bearing on their ability to acquire an L2. That is, the cognitive processes associated with acquiring a first language were observedly separate from the processes associated with acquiring subsequent languages.

The opposing perspective, known as the “contrastive hypothesis,” maintained that the structure of the first language affected the acquisition of the second (Fries, 1945; Lado, 1957; Weinreich, 1953). Influenced heavily by the fields of structural linguistics (Bloomfield, 1933) and behavioral psychology (Skinner, 1957), Lado maintained that “we can *predict* and *describe* the patterns that will cause difficulty in learning, and those that will not cause difficulty, by comparing systematically the language and culture to be learned with the native language and culture of the student” (1957, p. vii, emphasis mine). Thus was born the Contrastive Analysis Hypothesis (CAH), a term denoting both the theory of second language learning and the method through which similarities and differences could be examined (Larsen-Freeman & Long, 1991; Yang, 1992). Borrowing from key components of structuralism, a methodology maintaining that all elements of human cognition can be understood in terms of their relationship to a larger overarching

system, is the assumption that there is a “finite structure of a given language that can be documented and compared with another language” (Yang, 1992). Lado (1957) explicitly states:

In the comparison between native and foreign language lies the key to ease or difficulty in foreign language learning...individuals tend to transfer the forms and meanings and the distribution of forms and meanings of their native language and culture to the foreign language and culture-both productively when attempting to speak the language and receptively when attempting to grasp and understand the language and culture as practiced by natives (p. 1-2).

He goes on to state:

those elements that are similar to [the learners’] native language will be simple for him, and those elements that are different will be difficult (p. 2).

Since more sophisticated data-analytic techniques were not typically applied to early research in the social sciences, scientists in the 1940s through 1960s relied on descriptive analyses of isolated language features to systematically identify similarities and differences between two or more languages. As noted by Larsen-Freeman and Long (1991), Lado borrowed the conceptualization of the CAH from Charles Fries, a prominent applied linguist who was researching the teaching and learning of English as a foreign language. Fries (1945) stated:

The most efficient materials are those that are based on a scientific description of the language to be learned, carefully compared with a parallel description of the native language of the learner. It is not enough to simply have the results of such a thorough-going analysis; these results must be organized into a satisfactory system for teaching and implemented with adequate specific practice materials through which the learner may master the sound system, the structure, and the most useful lexical materials of the foreign language (p. 9).

The pedagogical aspects of Fries’ work perhaps prompted Lado to suggest that a contrastive analysis between two languages should be employed when developing foreign

language instructional materials.⁶¹ The assertion noted above, that “those elements that are similar to [the learner’s] native language will be simple for him, and those elements that are different will be difficult” (p. 2), had major implications for the establishment of foreign language learning instructional frameworks. Based on Lado’s assertions, through the systematic comparison between two languages, instructors could *predict* where learners might experience ease or difficulty with learning various components of a foreign language and thereby plan their curriculum accordingly. As stated by Larsen-Freeman and Long (1991), “Where two languages were similar, positive transfer would occur; where they were different, negative transfer, or interference, would result” (p. 53). That is, the greater the disparities between one’s native language and target language, the more difficult the language would be for a learner to acquire.

Lado’s CAH gave rise to a variety of contrastive analyses in the 1960s through late 1970s. In their seminal series, Stockwell et al. (1965) applied the use of the CAH to systematically outline both grammatical and phonological differences of the five main foreign languages taught within the United States at the time: (1) French, (2) German, (3), Italian, (4) Russian, and (5) Spanish.⁶² They state:







The Center for Applied Linguistics, in undertaking this series of studies, has acted on the conviction held by many linguists and specialists in language teaching that one of the major problems in the learning of a second language is the interference caused by the structural differences between the native language of the learner and the second language... a careful contrastive analysis of the two languages offers an excellent basis for the preparation of instructional materials, the planning of courses, and the development of actual classroom techniques (p. v.).

⁶¹ Also noted by Larsen-Freeman and Long (1991) was that Robert Lado had been a former student of Charles Fries.

⁶² The series of 10 books (two volumes for each language, one focused on phonology and the other focused on grammar) began in 1959 and was later funded by the Center for Applied Linguistics.

The research by Stockwell et al. (1965) not only expanded the CAH to include dichotomous characterizations of structural differences, but also outlined both functional and semantic correspondences. As noted by Larsen-Freeman and Long (1991, pp. 53-54), Stockwell et al. (1965) introduce the use of a hierarchy of difficulty based on five main characteristics of grammatical difficulty type: (1) split, (2) new, (3) absent, (4) coalesced, and (5) correspondence. They state, “We seek, then, to determine what kinds of failures to match may exist in the possible set of choices, and to arrange these in a hierarchy from more to less difficult” (p. 283). Table 1 below, taken from Larsen-Freeman and Long (1991, p. 54), provides an English L1/Spanish L2 example of Stockwell et al.’s (1965) hierarchy of difficulty from hypothesized hardest (1) to easiest (5) features of L2 acquisition.

Table 1. Hierarchy of Difficulty (taken from Stockwell et al. (1965))

<i>Type of Difficulty</i>	<i>L1: English, L2: Spanish</i>	<i>Example</i>
1. Split		
2. New		marking grammatical gender
3. Absent		<i>Do</i> as a tense carrier
4. Coalesced		his/her realized as a single form <i>su</i>
5. Correspondence		<i>-ing</i> = <i>ndo</i> as a complement with verbs of perception

As can be gleaned from Table 1, the research by Stockwell et al. expands upon Lado’s (1954) earlier work by positing that L2 learners will have greatest difficulty acquiring aspects of the L2 in which a binary choice has to be made between two possibilities. For example, the “Split” category in Table 1 predicts that it will be most difficult for L1

English learners of Spanish to choose between two different forms of the preposition “for,” which has a single form in English. Lado’s previous work had postulated that the “New” or “Absent” categories would be most difficult for foreign language learners.

Research informing the establishment of difficulty hierarchies for L2 acquisition was taking place during a time when L2 instructional methods were heavily influenced by Skinnerian behaviorism, which maintained that not only external behaviors and events, but also internal cognitive changes were largely brought about through habit formation. Pedagogical methods were thus largely designed around the notion of “drill and kill” activities. In the appendix of their seminal work entitled “Pedagogy,” Stockwell et al. (1965) write:

In order to master the complicated structure of language efficiently, the student’s attention should be drawn to one-and only one-new point at a time. Not only should grammatical patterns be presented in their simplest forms, but there should be enough drill for control and sufficient review for mastery. A student needs the experience that will enable him to call on any pattern in his repertory, fill it in with any appropriate vocabulary item he has learned, and place sentences in a logical sequence without any thought of analysis. The experience (for a second language) can be given only in drill sessions which exhaust a large proportion of the possibilities of sentence formation that exist at any point in the student’s progress. That is, we need more and simpler drills, carried out with dispatch and efficiency, with the range of choices confined to a single point in each drill. (p. 294)

Fluency, at the time, was thought of not as a learner’s ability to effortlessly and automatically apply learned concepts to novel situations, but as one’s ability to efficiently fill-in-the-blank of previously memorized instructional material. The instructor’s job, therefore, was to present the learner with as many L2 drill patterns as necessary for the student to commit them to memory and reproduce them when prompted. The authors

provide the following as examples for instructors to use when teaching the structural relationship between articles and object pronouns (p. 295):

Quiero la pluma que tienes. [I want the pen that you have.]	Quiero la que tienes. [I want what you have.]	La quiero. [I want it.]
Quiero las sillas que tienes. [I want the seats that you have.]	Quiero las que tienes. [I want what you have.]	Las quiero. [I want them.]
Quiero los libros que tienes. [I want the books that you have.]	Quiero los que tienes. [I want what you have.]	Los quiero. [I want them.]

The purpose of these drill-based activities was to “provide sufficient repetition in meaningful context to establish correct habitual responses” (p. 295). In this vein, L2 learners were considered successful when they could quickly reproduce memorized features of the target language when prompted. As noted by Larsen-Freeman and Long (1991),

The behaviorists held that language acquisition was a product of habit formation. Habits were constructed through the repeated association between some stimulus and some response, which would become bonded when positively reinforced. Second language learning, then, was viewed as a process of overcoming the habits of the native language in order to acquire the new habits of the target language. The Contrastive Analysis Hypothesis was important to this view of language-learning since, if trouble spots in the target language could be anticipated, errors might be prevented or at least held to a minimum. In this way, the formation of bad habits could be avoided” (p. 55).

That is, the CAH not only informed instructors about which features of the target language a learner would likely have trouble acquiring, but also predicted the *amount of time* L2 instructors should spend on a given language feature. The dominant school of thought was that the larger the disparity between the L1 and L2, the greater the predicted interference between the two languages, and the longer the amount of time that instructors would have to spend practicing drill- and pattern-based activities. Wardhaugh

(1970) later proposed a distinction between strong and weak versions of the CAH wherein the strong version of the CAH *predicted* learner errors a priori, before they occurred, and the weak version of the CAH *explained* learner errors a posteriori, after they occurred.

Appendix B: List of DLAB 2 Predictor and Outcome Variables
(bolded and italicized variables indicate inclusion in the imputation model)

Dimensionality of Potential Predictor Variables				
Existing Tests (ASVAB and DLAB)	Demographic and Biographical Variables	Cognitive and Perceptual Measures	Personality Measures	Motivational Measures
ASVAB Verbal Expression	Age	<i>Task switching</i>	<i>Measure of ambiguity tolerance</i>	<i>General perceived self-efficacy scale</i>
ASVAB Arithmetic Reasoning	Sex	<i>Antisaccade-analogue</i>	Need for closure-close-mindedness	<i>Achievement Goal Inventory (AGI)-Normative outcome</i>
ASVAB Mathematical Knowledge	Pay Grade	<i>AFOQT reading comprehension</i>	Need for closure-decisiveness	<i>AGI-Ability outcome</i>
ASVAB General Science	Years of Military Service	<i>Vocabulary test</i>	Need for closure-discomfort with ambiguity	<i>AGI-Challenge mastery</i>
ASVAB Mechanical Comprehension	Marital Status	<i>Inference tests</i>	Need for closure-lie scale	<i>AGI-Learning</i>
ASVAB Electronics Information	Education	<i>Explicit induction test</i>	Need for closure-preference for order scale	AGI-Ability goals
ASVAB Auto and Shop Information	Motivation toward FL Learning	<i>Serial reaction time</i>	Need for closure-preference for predictability	AGI-outcome goals
ASVAB Assembling Objects	Prior informal language experience	<i>Paired associates test</i>	<i>Self-monitoring scale</i>	<i>Patterns of Adaptive Learning (PALS)-Personal mastery</i>
DLAB Biographical Data Scale	Prior formal language experience	<i>Available long-term memory-Synonyms test</i>	Social desirability scale	PALS-Personal performance-Approach
DLAB-I Scaled Individual Items	<i>Prior language proficiency</i>	<i>Task switching baseline reaction time</i>	<i>Tailored Adaptive Personality Assessment Scale (TAPAS)-Aesthetics</i>	<i>PALS Personal performance-Avoidance</i>
DLAB I-A	<i>Early home exposure to Non-English</i>	<i>Serial reaction time baseline reaction time</i>	<i>TAPAS -Curiosity</i>	PALS Classroom mastery
DLAB I-B	Number of languages heard as a child	<i>Running memory span test</i>	<i>TAPAS-Dominance</i>	PALS Classroom performance-Approach
DLAB I-C	<i>Age of first FL listening exposure</i>	Non-word span test	<i>TAPAS-Even Tempered</i>	PALS Classroom performance-Avoidance
DLAB I-D	<i>Best early language self-rated listening proficiency</i>	Phonemic discrimination-English contrastive	<i>TAPAS-Achievement</i>	PALS Academic efficacy
DLAB I-E	<i>Number of languages spoken as a child</i>	<i>Phonemic discrimination-Hindi, English noncontrastive</i>	<i>TAPAS-Intellectual Efficiency</i>	PALS Academic self-handicapping
DLAB I-F	Best early language self-rated speaking proficiency	<i>Phonemic discrimination-Hindi, English Pseudo Contrastive</i>	<i>TAPAS-Adjustment (No anxiety)</i>	PALS Avoiding novelty

Dimensionality of Potential Predictor Variables...cont

DLAB I-G	Average best early language self-rated across two skills	Musical experience/aptitude	<i>TAPAS-Order</i>	PALS-Cheating behavior
DLAB Stress Patterns (DLAB II)	Number of FLs studied previously	<i>Difficulty singing</i>	<i>TAPAS-Physical condition</i>	<i>PALS-Disruptive behavior</i>
DLAB FL Grammar-1	Best self-rated reading ability in any FL	Difficulty matching pitch	<i>TAPAS-Responsibility</i>	<i>PALS-Self-presentation of low achievement</i>
DLAB FL Grammar-2	Best self-rated writing ability in any FL	Number of musical instruments studied	<i>TAPAS-Socialability</i>	<i>PALS-Skepticism about the relevance of school for future performance</i>
DLAB FL Grammar-3	Best self-rated listening ability in any FL	<i>Months of musical training</i>	<i>TAPAS-Generosity</i>	<i>Stress and Coping Scale (SCOPE)-Academic planning</i>
DLAB FL Grammar-4	Best self-rated speaking ability in any FL	<i>Frequency of listening to music</i>	<i>TAPAS-Tolerance</i>	<i>SCOPE-Academic disengagement</i>
DLAB Concept Formation	<i>Average best self-rated ability across four skills</i>	Frequency of getting a tune stuck in one's head	<i>TAPAS-Non-delinquency (Traditionalism)</i>	SCOPE-Active study coping
	<i>Earliest age of exposure to current DLI target language</i>		<i>TAPAS-Cooperation</i>	SCOPE-Denial
	<i>Total number of years of exposure to current DLI target language (18)</i>		<i>TAPAS-Excitement Seeking</i>	SCOPE-Efficacy
			TAPAS-Optimism	<i>SCOPE-Emotional vetting</i>
				SCOPE-General active coping
				SCOPE-General emotional support

Appendix C: Correlation Matrices for Average Coursework Outcomes

Average 100-, 200-, and 300-level course outcomes

Arabic Correlation Matrix (Observed Data)

		Grade_101	Grade_102	Grade_110	Grade_120	Grade_140	Grade_201	Grade_202	Grade_210	Grade_220	Grade_240	Grade_301	Grade_302	Grade_310	Grade_320	Grade_340
Grade_101	Pearson Correlation	1	.682**	.651**	.636**	.495**	.484**	.531**	.584**	.545**	.345**	.289**	.487**	.463**	.398**	.389**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_102	Pearson Correlation	.682**	1	.615**	.605**	.400**	.665**	.685**	.545**	.588**	.436**	.492**	.616**	.487**	.405**	.342**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_110	Pearson Correlation	.651**	.615**	1	.636**	.497**	.453**	.515**	.717**	.517**	.403**	.352**	.409**	.600**	.401**	.392**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_120	Pearson Correlation	.636**	.605**	.636**	1	.310**	.454**	.483**	.553**	.746**	.337**	.331**	.466**	.544**	.598**	.366**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_140	Pearson Correlation	.495**	.400**	.497**	.310**	1	.221**	.350**	.395**	.238**	.318**	.138**	.266**	.250**	.141**	.240**
	Sig. (2-tailed)	.000	.000	.000	.000		.001	.000	.000	.000	.000	.032	.000	.000	.028	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_201	Pearson Correlation	.484**	.665**	.453**	.454**	.221**	1	.637**	.528**	.539**	.293**	.443**	.568**	.414**	.350**	.345**
	Sig. (2-tailed)	.000	.000	.000	.000	.001		.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_202	Pearson Correlation	.531**	.685**	.515**	.483**	.350**	.637**	1	.554**	.551**	.349**	.564**	.709**	.479**	.366**	.300**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_210	Pearson Correlation	.584**	.545**	.717**	.553**	.395**	.528**	.554**	1	.527**	.375**	.311**	.472**	.601**	.354**	.435**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_220	Pearson Correlation	.545**	.588**	.517**	.746**	.238**	.539**	.551**	.527**	1	.366**	.378**	.627**	.494**	.706**	.355**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_240	Pearson Correlation	.345**	.436**	.403**	.337**	.318**	.293**	.349**	.375**	.366**	1	.264**	.394**	.437**	.165**	.416**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.010	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_301	Pearson Correlation	.289**	.492**	.352**	.331**	.138**	.443**	.564**	.311**	.378**	.264**	1	.593**	.431**	.165**	.224**
	Sig. (2-tailed)	.000	.000	.000	.000	.032	.000	.000	.000	.000	.000		.000	.000	.010	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_302	Pearson Correlation	.487**	.616**	.409**	.466**	.266**	.568**	.709**	.472**	.627**	.394**	.593**	1	.535**	.369**	.415**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_310	Pearson Correlation	.463**	.487**	.600**	.544**	.250**	.414**	.479**	.601**	.494**	.437**	.431**	.535**	1	.292**	.584**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_320	Pearson Correlation	.398**	.405**	.401**	.598**	.141**	.350**	.366**	.354**	.706**	.165**	.165**	.369**	.292**	1	.108
	Sig. (2-tailed)	.000	.000	.000	.000	.028	.000	.000	.000	.000	.010	.010	.000	.000		.094
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241
Grade_340	Pearson Correlation	.389**	.342**	.392**	.366**	.240**	.345**	.300**	.435**	.355**	.416**	.224**	.415**	.584**	.108	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.094
	N	241	241	241	241	241	241	241	241	241	241	241	241	241	241	241

** . Correlation is significant at the 0.01 level (2-tailed).
 * . Correlation is significant at the 0.05 level (2-tailed).

Chinese Correlation Matrix (Observed Data)

		Grade_101	Grade_102	Grade_110	Grade_120	Grade_140	Grade_201	Grade_202	Grade_210	Grade_220	Grade_240	Grade_301	Grade_302	Grade_310	Grade_320	Grade_340
Grade_101	Pearson Correlation	1	.637**	.597**	.388**	.164	.556**	.465**	.382**	-.094	-.085	.547**	.594**	.310**	.180	-.081
	Sig. (2-tailed)		.000	.000	.000	.106	.000	.000	.000	.359	.406	.000	.000	.002	.076	.429
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_102	Pearson Correlation	.637**	1	.569**	.377**	.345**	.766**	.595**	.441**	.105	.098	.642**	.590**	.462**	.297**	.106
	Sig. (2-tailed)	.000		.000	.000	.001	.000	.000	.000	.301	.338	.000	.000	.000	.003	.301
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_110	Pearson Correlation	.597**	.569**	1	.128	.289**	.482**	.343**	.754**	.159	.206	.397**	.420**	.617**	.025	.153
	Sig. (2-tailed)	.000	.000		.211	.004	.000	.001	.000	.118	.042	.000	.000	.000	.806	.133
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_120	Pearson Correlation	.388**	.377**	.128	1	.290**	.218	.111	-.006	.007	-.026	.239	.315**	.039	-.056	.038
	Sig. (2-tailed)	.000	.000	.211		.004	.031	.275	.950	.945	.797	.018	.002	.705	.583	.708
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_140	Pearson Correlation	.164	.345**	.289**	.290**	1	.339**	.138	.202	.131	.166	.285**	.134	.268**	.132	.238
	Sig. (2-tailed)	.106	.001	.004	.004		.001	.177	.046	.198	.103	.004	.188	.008	.195	.018
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_201	Pearson Correlation	.556**	.766**	.482**	.218	.339**	1	.644**	.494**	.073	.079	.756**	.544**	.433**	.246	.124
	Sig. (2-tailed)	.000	.000	.000	.031	.001		.000	.000	.476	.442	.000	.000	.000	.015	.224
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_202	Pearson Correlation	.465**	.595**	.343**	.111	.138	.644**	1	.480**	.175	.092	.677**	.743**	.373**	.229	.091
	Sig. (2-tailed)	.000	.000	.001	.275	.177	.000		.000	.084	.366	.000	.000	.000	.023	.372
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_210	Pearson Correlation	.382**	.441**	.754**	-.006	.202	.494**	.480**	1	.242	.249	.419**	.469**	.682**	.103	.229
	Sig. (2-tailed)	.000	.000	.000	.950	.046	.000	.000		.016	.013	.000	.000	.000	.312	.024
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_220	Pearson Correlation	-.094	.105	.159	.007	.131	.073	.175	.242	1	.711**	-.053	-.018	.009	-.101	.781**
	Sig. (2-tailed)	.359	.301	.118	.945	.198	.476	.084	.016		.000	.601	.860	.933	.321	.000
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_240	Pearson Correlation	-.085	.098	.206	-.026	.166	.079	.092	.249	.711**	1	-.057	-.001	.090	-.128	.848**
	Sig. (2-tailed)	.406	.338	.042	.797	.103	.442	.366	.013	.000		.575	.996	.377	.210	.000
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_301	Pearson Correlation	.547**	.642**	.397**	.239	.285**	.756**	.677**	.419**	-.053	-.057	1	.631**	.419**	.093	-.003
	Sig. (2-tailed)	.000	.000	.000	.018	.004	.000	.000	.000	.601	.575		.000	.000	.361	.980
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_302	Pearson Correlation	.594**	.590**	.420**	.315**	.134	.544**	.743**	.469**	-.018	-.001	.631**	1	.465**	.329**	-.039
	Sig. (2-tailed)	.000	.000	.000	.002	.188	.000	.000	.000	.860	.996	.000		.000	.001	.700
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_310	Pearson Correlation	.310**	.462**	.617**	.039	.268**	.433**	.373**	.682**	.009	.090	.419**	.465**	1	.083	.011
	Sig. (2-tailed)	.002	.000	.000	.705	.008	.000	.000	.000	.933	.377	.000	.000		.419	.912
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_320	Pearson Correlation	.180	.297**	.025	-.056	.132	.246	.229	.103	-.101	-.128	.093	.329**	.083	1	-.095
	Sig. (2-tailed)	.076	.003	.806	.583	.195	.015	.023	.312	.321	.210	.361	.001	.419		.352
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Grade_340	Pearson Correlation	-.081	.106	.153	.038	.238	.124	.091	.229	.781**	.848**	-.003	-.039	.011	-.095	1
	Sig. (2-tailed)	.429	.301	.133	.708	.018	.224	.372	.024	.000	.000	.980	.700	.912	.352	
	N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98

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

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

Korean Correlation Matrix (Observed Data)

		Grade_101	Grade_102	Grade_110	Grade_120	Grade_140	Grade_201	Grade_202	Grade_210	Grade_220	Grade_240	Grade_301	Grade_302	Grade_310	Grade_320	Grade_340
Grade_101	Pearson Correlation	1	.669**	.658**	.427**	.312**	.624**	.588**	.566**	.251	.257	.637**	.432**	.440**	.295	-.012
	Sig. (2-tailed)		.000	.000	.000	.006	.000	.000	.000	.030	.026	.000	.000	.000	.010	.918
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_102	Pearson Correlation	.669**	1	.620**	.575**	.281	.782**	.772**	.660**	.355**	.370**	.710**	.494**	.667**	.477**	.045
	Sig. (2-tailed)	.000		.000	.000	.015	.000	.000	.000	.002	.001	.000	.000	.000	.000	.699
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_110	Pearson Correlation	.658**	.620**	1	.188	.148	.504**	.453**	.752**	.152	.307**	.481**	.361**	.568**	.327**	.095
	Sig. (2-tailed)	.000	.000		.107	.206	.000	.000	.000	.193	.007	.000	.001	.000	.004	.419
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_120	Pearson Correlation	.427**	.575**	.188	1	.037	.461**	.480**	.231	.680**	.149	.531**	.338**	.256	.214	-.271
	Sig. (2-tailed)	.000	.000	.107		.751	.000	.000	.046	.000	.203	.000	.003	.027	.066	.018
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_140	Pearson Correlation	.312**	.281	.148	.037	1	.326**	.358**	.312**	.016	.201	.361**	.094	.263	.053	.039
	Sig. (2-tailed)	.006	.015	.206	.751		.004	.002	.006	.890	.083	.001	.425	.022	.649	.741
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	
Grade_201	Pearson Correlation	.624**	.782**	.504**	.461**	.326**	1	.785**	.659**	.306**	.280	.703**	.593**	.553**	.463**	.111
	Sig. (2-tailed)	.000	.000	.000	.000	.004		.000	.000	.007	.015	.000	.000	.000	.000	.341
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_202	Pearson Correlation	.588**	.772**	.453**	.480**	.358**	.785**	1	.680**	.403**	.257	.763**	.702**	.669**	.558**	.061
	Sig. (2-tailed)	.000	.000	.000	.000	.002	.000		.000	.000	.026	.000	.000	.000	.000	.606
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_210	Pearson Correlation	.566**	.660**	.752**	.312**	.312**	.659**	.680**	1	.258	.261	.551**	.447**	.681**	.412**	.033
	Sig. (2-tailed)	.000	.000	.000	.046	.006	.000	.000		.026	.024	.000	.000	.000	.000	.776
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_220	Pearson Correlation	.251	.355**	.152	.680**	.016	.306**	.403**	.258	1	.112	.427**	.266	.308**	.119	-.206
	Sig. (2-tailed)	.030	.002	.193	.000	.890	.007	.000	.026		.339	.000	.021	.007	.311	.076
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_240	Pearson Correlation	.257	.370**	.307**	.149	.201	.280	.257	.261	.112	1	.420**	.108	.262	.220	.068
	Sig. (2-tailed)	.026	.001	.007	.203	.083	.015	.026	.024	.339	.000	.000	.356	.023	.057	.562
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_301	Pearson Correlation	.637**	.710**	.481**	.531**	.361**	.703**	.763**	.551**	.427**	.420**	1	.599**	.619**	.467**	-.020
	Sig. (2-tailed)	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000		.000	.000	.000	.863
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_302	Pearson Correlation	.432**	.494**	.361**	.338**	.094	.593**	.702**	.447**	.266	.108	.599**	1	.459**	.576**	-.038
	Sig. (2-tailed)	.000	.000	.001	.003	.425	.000	.000	.000	.021	.356	.000		.000	.000	.743
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_310	Pearson Correlation	.440**	.667**	.568**	.256	.263	.553**	.669**	.681**	.308**	.262	.619**	.459**	1	.318**	.220
	Sig. (2-tailed)	.000	.000	.000	.027	.022	.000	.000	.000	.007	.023	.000	.000		.005	.058
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_320	Pearson Correlation	.295	.477**	.327**	.214	.053	.463**	.558**	.412**	.119	.220	.467**	.576**	.318**	1	.062
	Sig. (2-tailed)	.010	.000	.004	.066	.649	.000	.000	.000	.311	.057	.000	.000	.005		.599
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75
Grade_340	Pearson Correlation	-.012	.045	.095	-.271	.039	.111	.061	.033	-.206	.068	-.020	-.038	.220	.062	1
	Sig. (2-tailed)	.918	.699	.419	.018	.741	.341	.606	.776	.076	.562	.863	.743	.058	.599	
	N	75	75	75	75	75	75	75	75	75	75	75	75	75	75	75

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

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

Appendix D: Correlation Matrices for Path-Analytic Model (All Languages)

Arabic Correlation Matrix (Observed Data)

		AFQT	Q_DLAB	I_MOTIVE	AVG_100	AVG_200	AVG_300	Q_LIST_LVL	Q_READ_LVL	Q_SPK_LVL
AFQT	Pearson Correlation	1	.469**	.056	.250**	.148*	.172**	.182**	.193**	.114
	Sig. (2-tailed)		.000	.390	.000	.022	.008	.005	.003	.078
	N	241	241	241	241	241	241	241	241	241
Q_DLAB	Pearson Correlation	.469**	1	.100	.367**	.303**	.296**	.304**	.272**	.274**
	Sig. (2-tailed)	.000		.122	.000	.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241
I_MOTIVE	Pearson Correlation	.056	.100	1	.148*	.041	-.041	-.052	.034	.009
	Sig. (2-tailed)	.390	.122		.022	.522	.529	.418	.604	.888
	N	241	241	241	241	241	241	241	241	241
AVG_100	Pearson Correlation	.250**	.367**	.148*	1	.784**	.692**	.516**	.455**	.450**
	Sig. (2-tailed)	.000	.000	.022		.000	.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241
AVG_200	Pearson Correlation	.148*	.303**	.041	.784**	1	.835**	.578**	.539**	.483**
	Sig. (2-tailed)	.022	.000	.522	.000		.000	.000	.000	.000
	N	241	241	241	241	241	241	241	241	241
AVG_300	Pearson Correlation	.172**	.296**	-.041	.692**	.835**	1	.595**	.548**	.490**
	Sig. (2-tailed)	.008	.000	.529	.000	.000		.000	.000	.000
	N	241	241	241	241	241	241	241	241	241
Q_LIST_LVL	Pearson Correlation	.182**	.304**	-.052	.516**	.578**	.595**	1	.431**	.354**
	Sig. (2-tailed)	.005	.000	.418	.000	.000	.000		.000	.000
	N	241	241	241	241	241	241	241	241	241
Q_READ_LVL	Pearson Correlation	.193**	.272**	.034	.455**	.539**	.548**	.431**	1	.360**
	Sig. (2-tailed)	.003	.000	.604	.000	.000	.000	.000		.000
	N	241	241	241	241	241	241	241	241	241
Q_SPK_LVL	Pearson Correlation	.114	.274**	.009	.450**	.483**	.490**	.354**	.360**	1
	Sig. (2-tailed)	.078	.000	.888	.000	.000	.000	.000	.000	
	N	241	241	241	241	241	241	241	241	241

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

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

Chinese Correlation Matrix (Observed Data)

		AFQT	Q_DLAB	I_MOTIVE	AVG_100	AVG_200	AVG_300	Q_LIST_LVL	Q_READ_LVL	Q_SPK_LVL
AFQT	Pearson Correlation	1	.368**	.011	.166	.149	.176	.134	.357**	-.083
	Sig. (2-tailed)		.000	.917	.103	.142	.082	.188	.000	.417
	N	98	98	98	98	98	98	98	98	98
Q_DLAB	Pearson Correlation	.368**	1	-.044	.335**	.259*	.313**	.128	.199*	.024
	Sig. (2-tailed)	.000		.664	.001	.010	.002	.210	.049	.817
	N	98	98	98	98	98	98	98	98	98
I_MOTIVE	Pearson Correlation	.011	-.044	1	-.046	-.095	-.018	.082	-.013	.078
	Sig. (2-tailed)	.917	.664		.654	.352	.857	.422	.901	.444
	N	98	98	98	98	98	98	98	98	98
AVG_100	Pearson Correlation	.166	.335**	-.046	1	.537**	.679**	.363**	.291**	.337**
	Sig. (2-tailed)	.103	.001	.654		.000	.000	.000	.004	.001
	N	98	98	98	98	98	98	98	98	98
AVG_200	Pearson Correlation	.149	.259*	-.095	.537**	1	.759**	.365**	.356**	.442**
	Sig. (2-tailed)	.142	.010	.352	.000		.000	.000	.000	.000
	N	98	98	98	98	98	98	98	98	98
AVG_300	Pearson Correlation	.176	.313**	-.018	.679**	.759**	1	.517**	.486**	.423**
	Sig. (2-tailed)	.082	.002	.857	.000	.000		.000	.000	.000
	N	98	98	98	98	98	98	98	98	98
Q_LIST_LVL	Pearson Correlation	.134	.128	.082	.363**	.365**	.517**	1	.360**	.402**
	Sig. (2-tailed)	.188	.210	.422	.000	.000	.000		.000	.000
	N	98	98	98	98	98	98	98	98	98
Q_READ_LVL	Pearson Correlation	.357**	.199*	-.013	.291**	.356**	.486**	.360**	1	.091
	Sig. (2-tailed)	.000	.049	.901	.004	.000	.000	.000		.373
	N	98	98	98	98	98	98	98	98	98
Q_SPK_LVL	Pearson Correlation	-.083	.024	.078	.337**	.442**	.423**	.402**	.091	1
	Sig. (2-tailed)	.417	.817	.444	.001	.000	.000	.000	.373	
	N	98	98	98	98	98	98	98	98	98

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

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

Korean Correlation Matrix (Observed Data)

		AFQT	Q_DLAB	I_MOTIVE	AVG_100	AVG_200	AVG_300	Q_LIST_LVL	Q_READ_LVL	Q_SPK_LVL
AFQT	Pearson Correlation	1	.284*	-.123	-.034	-.050	.113	.100	.286*	-.178
	Sig. (2-tailed)		.014	.293	.770	.668	.334	.395	.013	.126
	N	75	75	75	75	75	75	75	75	75
Q_DLAB	Pearson Correlation	.284*	1	-.100	.345**	.311**	.397**	.308**	.247*	.236*
	Sig. (2-tailed)	.014		.396	.002	.007	.000	.007	.033	.041
	N	75	75	75	75	75	75	75	75	75
I_MOTIVE	Pearson Correlation	-.123	-.100	1	-.039	-.003	-.101	-.068	-.240*	-.228*
	Sig. (2-tailed)	.293	.396		.737	.981	.390	.564	.038	.049
	N	75	75	75	75	75	75	75	75	75
AVG_100	Pearson Correlation	-.034	.345**	-.039	1	.822**	.679**	.477**	.398**	.392**
	Sig. (2-tailed)	.770	.002	.737		.000	.000	.000	.000	.000
	N	75	75	75	75	75	75	75	75	75
AVG_200	Pearson Correlation	-.050	.311**	-.003	.822**	1	.763**	.463**	.442**	.363**
	Sig. (2-tailed)	.668	.007	.981	.000		.000	.000	.000	.001
	N	75	75	75	75	75	75	75	75	75
AVG_300	Pearson Correlation	.113	.397**	-.101	.679**	.763**	1	.616**	.531**	.215
	Sig. (2-tailed)	.334	.000	.390	.000	.000		.000	.000	.064
	N	75	75	75	75	75	75	75	75	75
Q_LIST_LVL	Pearson Correlation	.100	.308**	-.068	.477**	.463**	.616**	1	.512**	.294*
	Sig. (2-tailed)	.395	.007	.564	.000	.000	.000		.000	.011
	N	75	75	75	75	75	75	75	75	75
Q_READ_LVL	Pearson Correlation	.286*	.247*	-.240*	.398**	.442**	.531**	.512**	1	.320**
	Sig. (2-tailed)	.013	.033	.038	.000	.000	.000	.000		.005
	N	75	75	75	75	75	75	75	75	75
Q_SPK_LVL	Pearson Correlation	-.178	.236*	-.228*	.392**	.363**	.215	.294*	.320**	1
	Sig. (2-tailed)	.126	.041	.049	.000	.001	.064	.011	.005	
	N	75	75	75	75	75	75	75	75	75

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

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

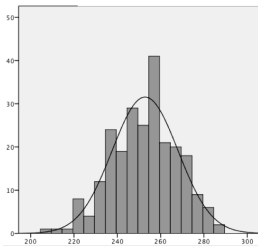
Appendix E: Descriptive Statistics for Imputed Data

(Non-Transformed and Transformed Distributions)

AFQT Score Distributions

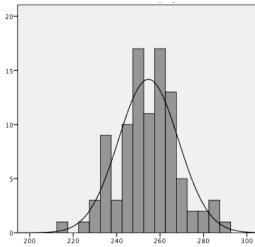
Non-Transformed

Arabic



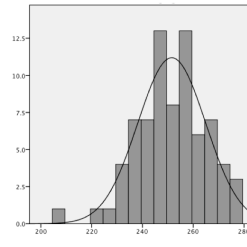
Mean = 252.81
SD = 15.23
N = 241
Skewness = -0.185
Kurtosis = -0.268

Chinese



Mean = 254.65
SD = 13.79
N = 98
Skewness = -0.052
Kurtosis = 0.133

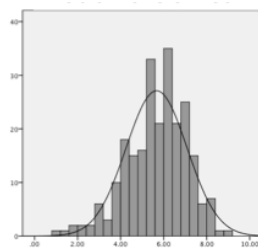
Korean



Mean = 251.53
SD = 13.73
N = 75
Skewness = -0.394
Kurtosis = 0.610

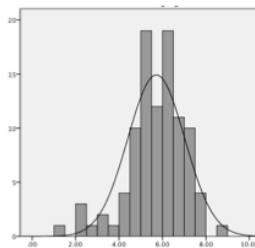
Transformed

Arabic



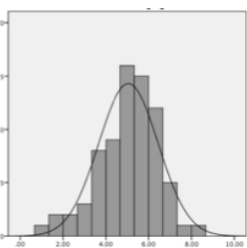
Mean = 5.67
SD = 1.41
N = 241
Skewness = -0.57
Kurtosis = 0.391

Chinese



Mean = 5.71
SD = 1.31
N = 98
Skewness = -0.878
Kurtosis = 1.573

Korean

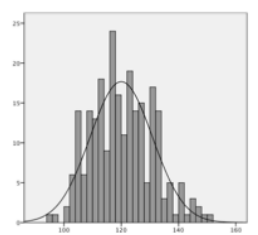


Mean = 5.05
SD = 1.40
N = 75
Skewness = -0.541
Kurtosis = 0.533

DLAB Score Distributions

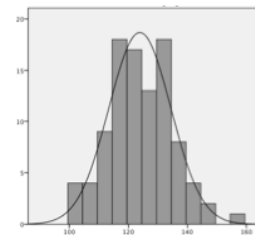
Non-Transformed

Arabic



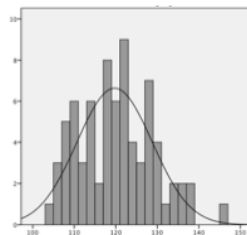
Mean = 120.00
SD = 10.86
N = 241
Skewness = 0.293
Kurtosis = -0.324

Chinese



Mean = 123.86
SD = 10.46
N = 98
Skewness = 0.182
Kurtosis = 0.256

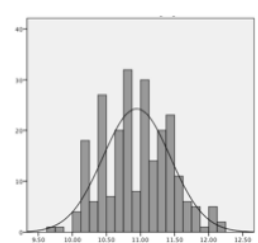
Korean



Mean = 119.72
SD = 9.025
N = 75
Skewness = 0.322
Kurtosis = -0.299

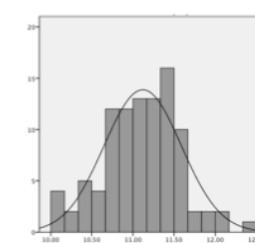
Transformed

Arabic



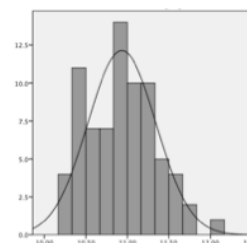
Mean = 10.94
SD = 0.49
N = 241
Skewness = 0.188
Kurtosis = -0.404

Chinese



Mean = 11.12
SD = 0.47
N = 98
Skewness = 0.049
Kurtosis = 0.143

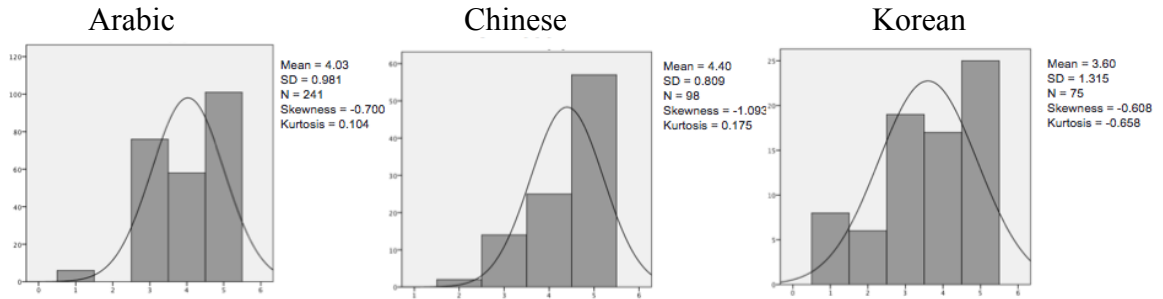
Korean



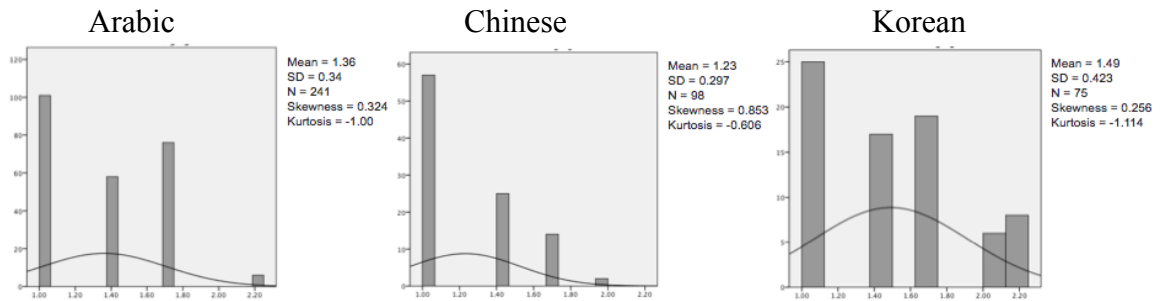
Mean = 10.93
SD = 0.41
N = 75
Skewness = 0.237
Kurtosis = -0.410

Language Preference Self-Assessment Score Distributions

Non-Transformed

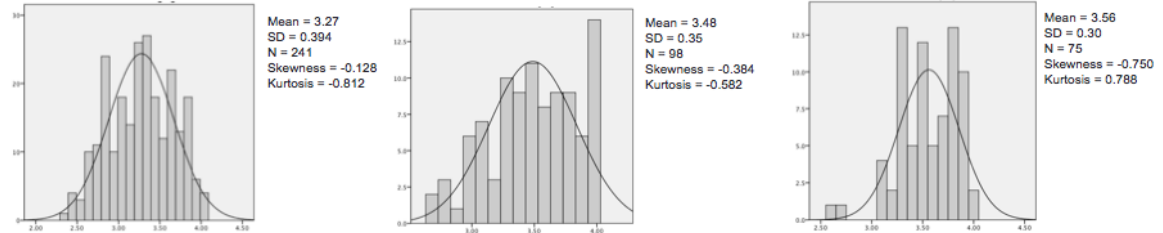


Transformed

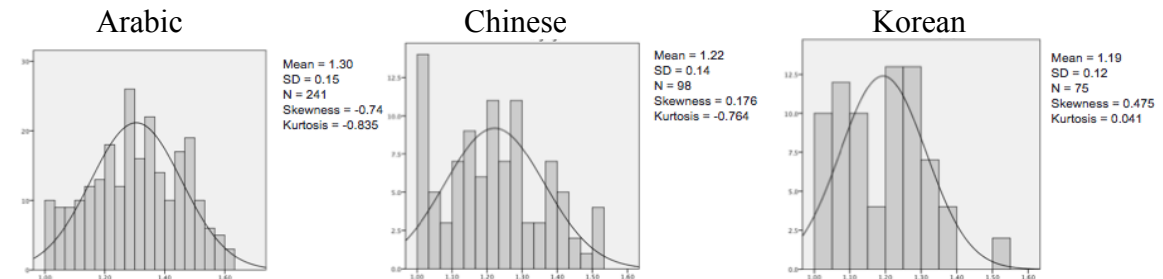


100-Level Average Coursework Distributions

Non-Transformed

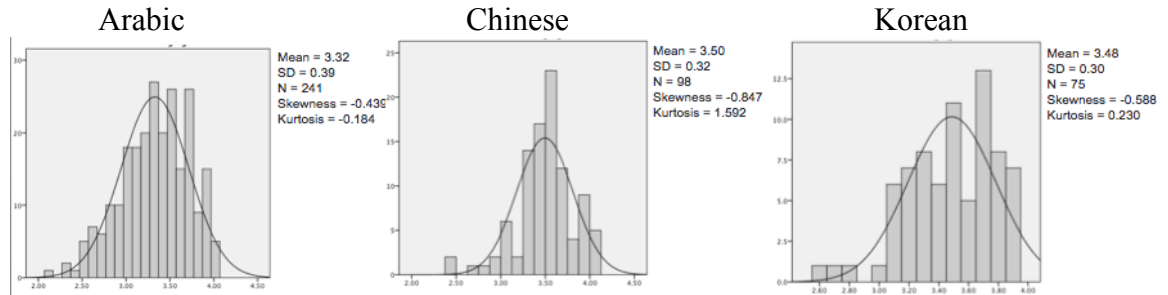


Transformed

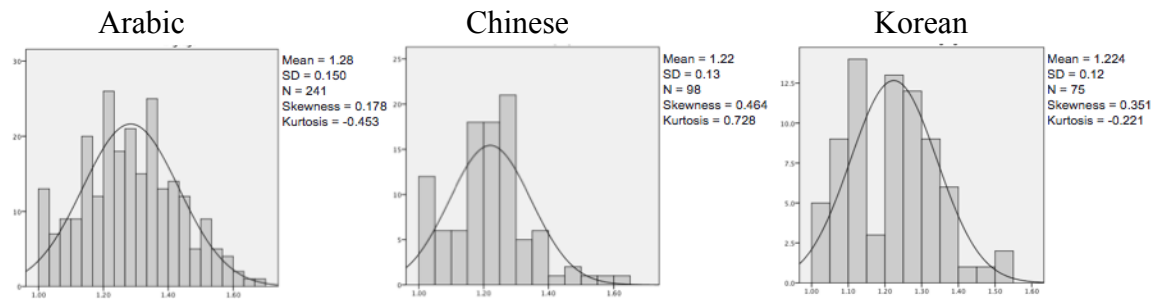


200-Level Average Coursework Distributions

Non-Transformed

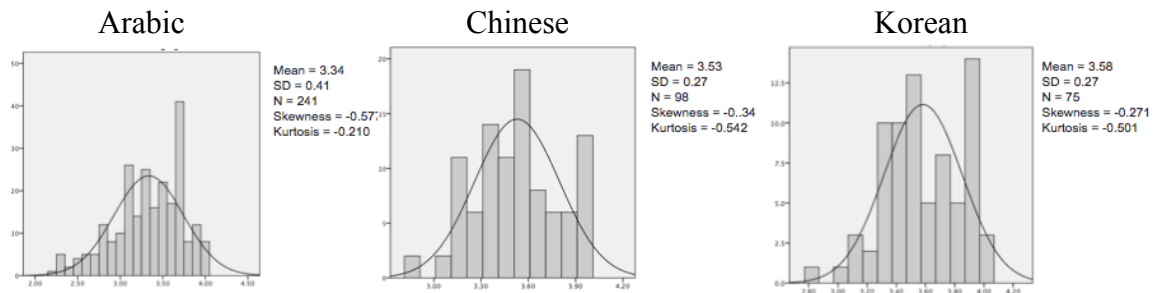


Transformed

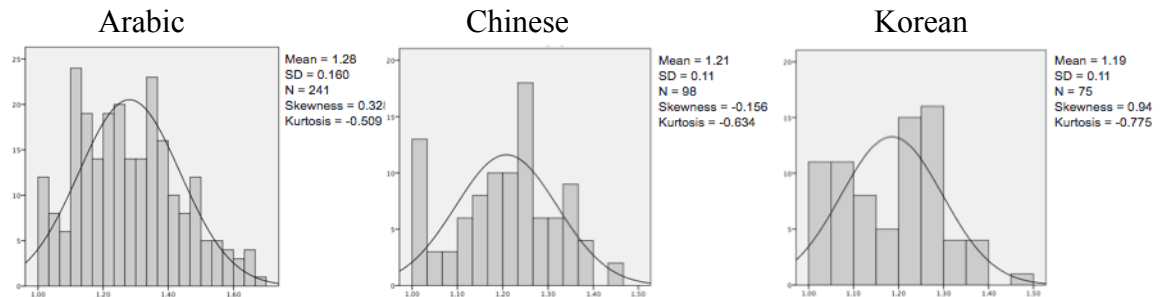


300-Level Average Coursework Distributions

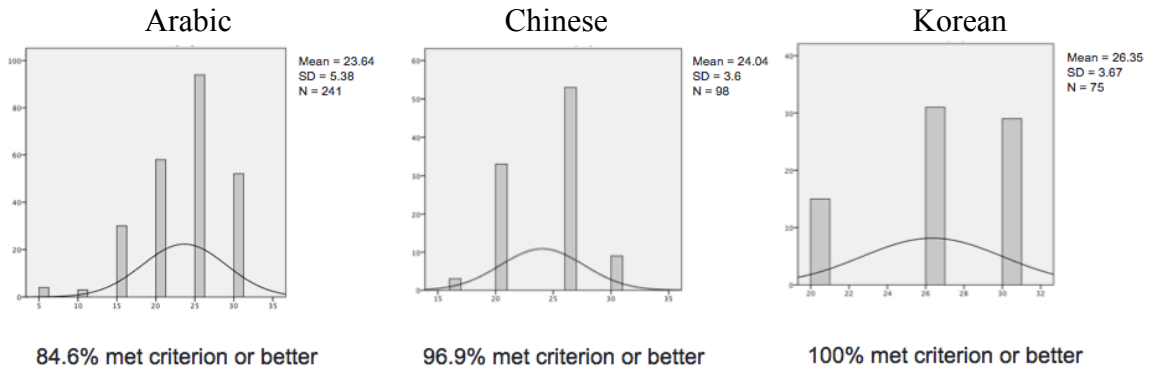
Non-Transformed



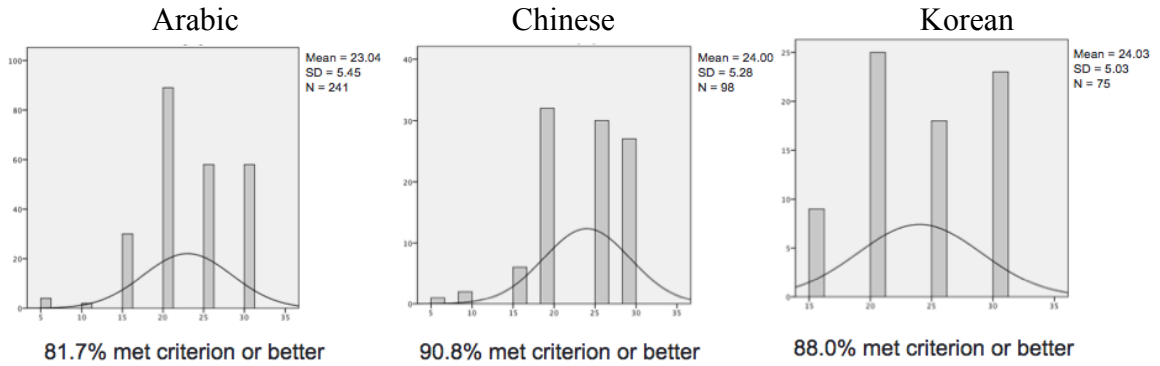
Transformed



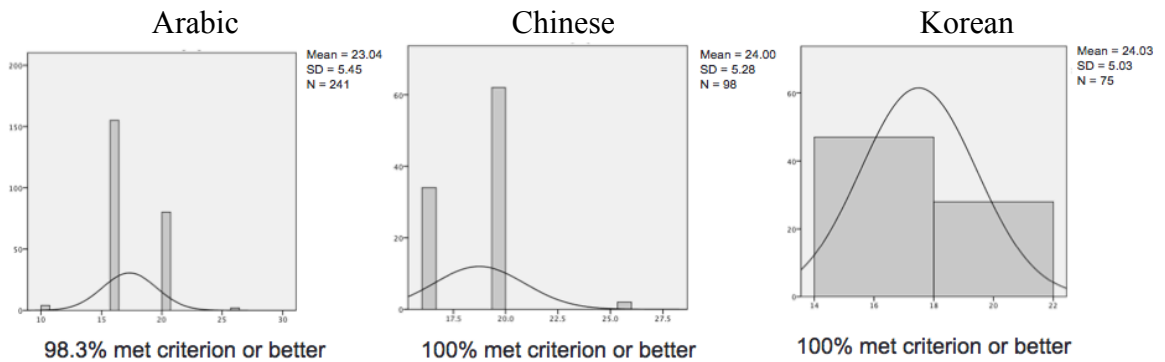
DLPT Reading Outcomes



DLPT Listening Outcomes



OPI Speaking Outcomes



Appendix F: Wave 2 and Wave 3 Observed and Imputed Mean-Level Comparisons

(All Languages)

Language	Variable	Observed Mean	Imputed Mean
Arabic	100-level average coursework outcomes	3.27	3.05
	200-level average coursework outcomes	3.33	3.04
	300-level average coursework outcomes	3.35	3.00
	DLPT Reading	23.04	19.35
	DLPT Listening	23.64	20.01
	OPI Speaking	17.31	15.32
Chinese	100-level average coursework outcomes	3.49	3.22
	200-level average coursework outcomes	3.49	3.31
	300-level average coursework outcomes	3.52	3.34
	DLPT Reading	24.00	22.73
	DLPT Listening	24.04	22.19
	OPI Speaking	18.73	17.52
Korean	100-level average coursework outcomes	3.57	3.47
	200-level average coursework outcomes	3.49	3.28
	300-level average coursework outcomes	3.58	3.39
	DLPT Reading	24.03	22.98
	DLPT Listening	26.35	25.36
	OPI Speaking	17.49	17.00

Appendix G: Minimum, Maximum, and Median RMSEA Values of Imputed Models

Arabic RMSEA Values (Imputed Data)

Model	Minimum RMSEA Value	Maximum RMSEA Value	Median RMSEA Value
Wave 1 to Wave 2 (100-level coursework outcomes)	0.105	0.204	0.162
Wave 1 to Wave 2 (200-level coursework outcomes)	0.106	0.143	0.124
Wave 1 to Wave 2 (300-level coursework outcomes)	0.087	0.168	0.111
DLPT Reading Outcomes	0.439	0.601	0.512
DLPT Listening Outcomes	0.417	0.636	0.554
OPI Speaking Outcomes	0.404	0.576	0.499

Chinese RMSEA Values (Imputed Data)

Model	Minimum RMSEA Value	Maximum RMSEA Value	Median RMSEA Value
Wave 1 to Wave 2 (100-level coursework outcomes)	0.056	0.125	0.082
Wave 1 to Wave 2 (200-level coursework outcomes)	0.022	0.078	0.051
Wave 1 to Wave 2 (300-level coursework outcomes)	0.043	0.103	0.066
DLPT Reading Outcomes	0.192	0.756	0.303
DLPT Listening Outcomes	0.054	0.528	0.183
OPI Speaking Outcomes	0.073	0.533	0.306

Korean RMSEA Values (Imputed Data)

Model	Minimum RMSEA Value	Maximum RMSEA Value	Median RMSEA Value
Wave 1 to Wave 2 (100-level coursework outcomes)	0.046	0.113	0.074
Wave 1 to Wave 2 (200-level coursework outcomes)	0.043	0.119	0.0605
Wave 1 to Wave 2 (300-level coursework outcomes)	0.092	0.195	0.124
DLPT Reading Outcomes	0.210	0.327	0.2455
DLPT Listening Outcomes	0.204	0.317	0.255
OPI Speaking Outcomes	0.200	0.276	0.230

Appendix F: Sample DLPT Listening and Reading Specimen

Example: Spanish Reading

Level 2

From a letter to the editor about urban safety in Paraguay

La semana pasada escribí sobre la agresividad creciente que se percibe en muchas esquinas de la ciudad a causa de la presencia de los "limpiavidrios", desvalidas personas que ofrecen sus servicios en las calles de la ciudad, estampando esponjas cargadas de agua sucia con jabón sobre el vidrio, a cambio de propinas.

Justamente ayer, a plena luz del día, en la esquina de Mariscal López y San Martín, un hombre no aceptó la propuesta de limpieza de su parabrisas por parte de una mujer, ya que acababa de lavarlo. Ésta, en un ataque de rabia por el rechazo, le golpeó el techo con el mango del escurridor y le rayó la pintura. El hombre se bajó para ver los daños y la mujer se le tiró encima. Al instante acudieron en su ayuda otras personas que le propinaron no sólo una paliza al conductor, sino que además un hombre sacó un cuchillo y lo apuñaló. Dejaron al conductor más muerto que vivo, tirado en la calle.

2. What recurring issue is the writer denouncing in the letter?

- (A) Road rage is becoming a serious problem in the city.
- (B) The large number of peddlers creates traffic problems.
- (C) Homeless people continue to break into cars.
- (D) Street people keep forcing their services on motorists.

The correct answer is **(D)**.

The passage is a complaint about people who approach motorists, wash their windshields, and expect to be paid. Therefore, (D) is the best answer. (A) is not the best answer because the topic of the passage is not about angry drivers. The passage does not say anything about traffic problems being caused by people selling things, which makes (B) incorrect. (C) is not the best answer because there is no mention in the passage of people breaking into cars.

Example: Spanish Listening

Level 2

This passage is from a news segment on Salvadoran radio.

Un custodio y 4 reos del Centro Real de Quetzaltepec en el departamento de la Libertad resultaron lesionados al protagonizarse disturbios en dichas instalaciones.

Los problemas comenzaron desde la madrugada de hoy martes cuando los reclusos se amotinaron y los custodios se vieron obligados a efectuar disparos para controlar la situación.

Desde tempranas horas desde martes las autoridades del centro penal de Quetzaltepec le pidieron el apoyo a los elementos de la unidad de mantenimiento del ordenomo para brindar mayor seguridad y efectuar una requisita para decomisar armas hechizas.

2. What happened at Quetzaltepec prison?

- (A) A prison guard was killed.
- (B) A prison guard was taken hostage.
- (C) The inmates rioted.
- (D) Inmates tried to escape.

The correct answer is (C).

The first paragraph says that there were disturbances at the prison and the second paragraph states that the prisoners mutinied, making (C) the best answer. The passage makes no mention of an attempted escape, eliminating (D) as a possible answer. The first paragraph states that a guard was wounded, not killed, and there is no mention of any other action taken against guards by the inmates, which eliminates (A) and (B) as possible answers.

3. According to the passage, what action did the authorities take?

- (A) They transferred the inmates to another prison.
- (B) They requested help from outside of the prison.
- (C) They sealed off the area surrounding the prison.
- (D) They isolated the inmates who were responsible.

The correct answer is (B).

References

- Alderson, J. C. (1992). Guidelines for the evaluation of language education. In J. C. Alderson & Beretta, A. (Eds.). *Evaluating second language education* (pp. 274-304). Cambridge, UK: Cambridge University Press.
- Bentler, P. M. & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16, 78-117.
- Beretta, A. (1986). A case for field-experimentation in program evaluation. *Language Learning*, 36 (3), 295-309.
- Beretta, A. (1992a). Evaluation of language education: An overview. In J. C. Alderson & Beretta, A. (Eds.). *Evaluating second language education* (pp. 15-24). Cambridge, UK: Cambridge University Press.
- Beretta, A. (1992b). What can be learned from the Bangalore evaluation? In J. C. Alderson & Beretta, A. (Eds.). *Evaluating second language education* (pp. 250-271). Cambridge, UK: Cambridge University Press.
- Bloomfield, A. N., Gynther, K., Masters, M. C., O'Connell, S. P., & Ross, S. J. (2012). *How does foreign language proficiency change over time? Results suggest foreign language reading and listening skills are stable over time, while speaking skills are not.* (TTO 82104). College Park, MD: University of Maryland Center for Advanced Study of Language.
- Bloomfield, A. N., Masters, M. C., Castle, S., Mackey, B., Ross, S. J., & Clark, M. (2013). *How language proficiency test scores change over time: Differences across language difficulty categories.* (DO 0039). College Park, MD. University of Maryland Center for Advanced Study of Language.
- Bloomfield, L. (1933). *Language*. New York: Reinhart and Winston.
- Brouselle, A. & Champagne, F. (2011). Program theory evaluation: Logic analysis. *Evaluation and Program Planning*, 34, 69-78.
- Bunting, M. F., Bowles, A. R., Campbell, S. G., Linck, J. A., Mislevy, M. A., Jackson, S. R., & Doughty, C. J. (2011). Reinventing DLAB – Potential new predictors of success at DLIFLC: Results from construct-validation field testing for DLAB2. College Park: University of Maryland Center for Advanced Study of Language.
- Burt, M. K., and Dulay, H. (1975). On TESOL '75: New directions in second language learning, teaching, and bilingual education. *TESOL*, Washington, DC.

- Byrne, B. M. (2001). *Structural Equation Modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Child, J. R. (1998). Language aptitude testing: Learners and applications. *Applied Language Learning*, 9 (1-2), 1-10.
- Callow-Heusser, C., Chapman, H., & Torres, R. (2005) *Evidence: An Essential Tool*. Prepared for National Science Foundation under grant HER-023382.
- Clark, M., Ross, O'Rourke, P., Jackson, S., Bloomfield, A., Aghajanian-Stewart, K., Kim, S. (2016a). *The development of the language difficulty categorization framework*. (DO 0088). College Park, MD: University of Maryland Center for Advanced Study of Language.
- Clark, M., Ross, S., Jackson, S., Kim, S., O'Rourke, P., Aghajanian-Stewart, K. (2016b). *Empirical investigation of language difficulty categories*. (DO 0088). College Park, MD: University of Maryland Center for Advanced Study of Language.
- Clark, M., Ross, S., Jackson, S., Kim, S., O'Rourke, P., Aghajanian-Stewart, K. (2016c). *Empirical investigation of language difficulty categorization: Examining within difficulty category variation*. (DO 0088). College Park, MD: University of Maryland Center for Advanced Study of Language.
- Connell, J., and Kubisch, A. (1998). Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects, and problems. In Anderson, K., Kubisch, A. & Connell (Eds.). *New approaches to evaluating community initiatives*. Washington, DC: Aspen Institute.
- Cysouw, M. (2013). Predicting language-learning difficulty. In Borin, L. & Saxena, A. (Eds.), *Approaches to Measuring Linguistic Differences* (57-82). Berlin/Boston: Walter de Gruyter.
- Enders, C. (2010). *Applied missing data analysis (Methodology in the Social Sciences)*. New York, New York: The Guilford Press.
- Ervin-Tripp, S. (1974). Is second language-learning like the first? *TESOL Quarterly*, (8).
- Frechtling, J. A. (2007). *Logic Modeling Methods in Program Evaluation*. San Francisco, CA: Jossey-Bass.
- Fries, C. (1945). *Teaching and learning English as a foreign language*. Ann Arbor, MI: University of Michigan Press.
- Genesee, F. (1983). Bilingual education of majority language children: The immersion experiments in review. *Applied Psycholinguistics*, 4, 1-46.

- Keating, R. F. (1963). A study of the effectiveness of language laboratories. New York: Institute of Administrative Research, Teachers College.
- Kiley, R. & Rea-Dickens, P. (2005). *Program evaluation in language education*. New York: Palgrave Macmillan.
- Hakuta, K. and Cancino, E. (1977). Trends in second language acquisition research. *Harvard Educational Review*, **47**: 294-316.
- Hancock, G. (2011). *Structural Equation Modeling*. Workshop given at the University of Maryland Center for Advanced Study of Language. June 1-3, 6-7.
- Hart-Gonzalez, L. and Lindemann, S. (1993). *Expected achievement in speaking proficiency*. School of Language Studies, Foreign Service Institute, Department of State: Washington, DC.
- Jackson, S. (2016). *Missing data and multiple imputation*. R-script tutorial generated at the University of Maryland Center for Advanced Study of Language (December 8): College Park, MD.
- Jackson, F. H. & Kaplan, M. A. (2001). Fifty years of theory and practice in government language teaching. In J. Alatis & A. Hui Tan (Eds.). *1999 Georgetown University Round Table on Languages and Linguistics* (pp. 71-87). Georgetown University Press: Washington, DC.
- Jackson, S. R., Hughes, M. M., Linck, J. A., Campbell, S. G., Tare, M., Silbert, N. H., Bowles, A. R., Bunting, M. F., Drasgow, F., Stark, S., & Chernyshenko, S. (in preparation). DLAB2 field testing results: Cognitive, perceptual, and non-cognitive predictors of language-learning success. Invited submission for special issue of *Military Psychology*.
- Jakobovits, L. A. (1969). SLL and transfer theory: A theoretical assessment. *Applied Language Learning*, **14** (1-2).
- Koch, A. and McCloy, R. Human Resources Research Organization. (2015). *Language difficulty categorization in the Defense Language Testing Program*. Alexandria, VA.
- Lado, R. (1957). *Linguistics across cultures*. University of Michigan Press, Ann Arbor, Michigan.
- Larsen-Freeman, D. and Long, M. (1991). *An introduction to second language acquisition*, pp. 52-80. Longman, New York, New York.

- Lett, J. A. (n.d.). [PDF document] DLIFLC student profiles. Retrieved from <https://www.utexas.edu/cola/centers/tlc/files/proficiencyconference/presentation/DLI/2.pdf> on February 23, 2015.
- Lett, J. A. & O'Mara, F. E. (1990). Predictors of success in an intensive foreign language-learning context: correlates of language-learning at the Defense Language Institute Foreign Language Center. In Parry, T. & Stansfield, C. (Eds.). *Language aptitude reconsidered*. Englewood Cliffs, NJ: Prentice Hall, 222-260.
- Little, R. J. A. and Rubin, D. (2002). *Statistical analysis with missing data*. New York, New York: John Wiley and Sons.
- Long, M. H. (1984). Process and Product in ESL Program Evaluation. *TESOL Quarterly*, 18 (3), 409-425.
- Lowe, P. Jr. (1998). Zero-based language aptitude test design: Where's the focus for the test? *Applied Language Learning*, 9 (1-2), 11-30.
- Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge, UK: Cambridge University Press.
- Mackey, B. (2014). *Aptitude as a predictor of individual differences in language proficiency growth* (Qualifying Paper 2 Proposal Defense). College Park, MD: School of Language, Literatures and Cultures.
- Masters, M. C. (2016). *Pathways to proficiency: An exploration of how cognitive and non-cognitive variables contribute to the development of achievement and proficiency outcomes* (Qualifying Paper 2 Defense). College Park, MD: School of Language, Literatures and Cultures.
- McLaughlin, J. & Jordan, G. (1999). Using logic models. In Wholey, J. Hatry, H. & Newcomer, K. (2004). *Handbook of practical program evaluations*. San Francisco: Jossey-Bass.
- Norris, J. M. (2006). The why (and how) of student learning outcomes assessment in college FL education. *Modern Language Journal*, 90 (4), 576-583.
- Norris, J. M. (2008). *Validity evaluation in language assessment*. New York, NY: Peter Lang.
- Norris, J. M. (Ed.). (2009). Understanding and improving language education through program evaluation [Special issue]. *Language Testing Research*, 13, 1-13.
- Peterson, C. R. & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement*, 36, 369-380.

- Rea-Dickins, P., & Germaine, K. P. (Eds.). (1998). The price of everything and value of nothing: Trends in language program evaluation. In Rea-Dickens, P. & Germaine, K. P. (Eds.). *Managing evaluation and innovation in language teaching: Building bridges*. London: Longman.
- Renger, R., & Titcomb, A. (2002). A three-step approach to teaching logic models. *American Journal of Evaluation*, 23(4), 493–503.
- Ross, S. J. (2003). A diachronic coherence model for language program evaluation. *Language Learning*, 53 (1), 1-33.
- Ross, S.J, Bloomfield, A., Masters, M., Nielson, K., Kramasz, D., O’Connell, S., & Gynther, K. (2011). *How does foreign language proficiency change over time?* (Tech. Rep. Objective 2, TTO 82104). College Park, MD: University of Maryland Center for Advanced Study of Language.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63 (3), 581-592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, New York: John Wiley and Sons.
- Scherer, G. A. C. & Wertheimer, M. (1964). *A psycholinguistic experiment in foreign language teaching*, New York: McGraw-Hill.
- Sinharay, S., Stern, H. S., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6 (4), 317-329.
- Skinner, B. (1957). *Verbal Behavior*. Appleton-Century-Crofts, New York.
- Smith, P. D. (1970). *A comparison of the cognitive and audio-lingual approaches to foreign language instruction: The Pennsylvania foreign language project*. Philadelphia: The Center for Curriculum Development.
- Stockwell, R., Bowen, J., and Martin, J. (1965). *The grammatical structures of English and Spanish*. University of Chicago Press, Chicago, Illinois.
- Swain, M. & Lapkin, S. (1982). *Evaluating bilingual education: A Canadian case study*. Clevedon, UK: Multilingual Matters.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.) Boston, MA: Allyn & Bacon.
- United States Army Pocket Recruiter Guide:
<http://www.usarec.army.mil/hq/apa/download/PRG13-14.pdf>

- VanBuuren, S. (2017). Multiple imputation. <http://www.stefvanbuuren.nl/mi/MI.html>
- Wagner, T. (2014). *Aptitude and achievement measures as predictors of growth in second language proficiency* (Qualifying Paper 2 Proposal Defense). College Park, MD: School of Language, Literatures and Cultures.
- Wardaugh, R. (1970). The contrastive analysis hypothesis. *TESOL Quarterly*, (4): 123-130.
- Watanabe, Y., Sylwester, B. & Norris, J.M. (2009). Foreign language program evaluation: An annotated bibliography of resources for foreign language educators: Honolulu: University of Hawai'i, National Foreign Language Research Center. www.nflrc.hawaii.edu/evaluation/biblio/index.cfm.
- Wayland, S., Saner, L., O'Connell, S., Linck, J., Kramasz, D., Gynther, K., Bloomfield, A., & Ralph, A. (2013). The long and the short of it: Passage length and information density in second language listening comprehension. College Park: University of Maryland Center for Advanced Study of Language.
- Weinreich, M. (1953). *Languages in Contact*. Linguistics Circle of New York.
- Weiss, C. H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review*, 21, 501–524.
- Weiss, C. H. (2000). Which links in which theories shall we evaluate? In Rogers, P. J., Hacsı, T. A., Petrosino, A., & Huebner, T. A. (Eds.), *Program theory in evaluation: Challenges and opportunities. New directions for evaluation*. Vol. 87 (pp.35–45). San Francisco: Jossey-Bass Publishers.
- Yang, B. (1992). A review of the contrastive analysis hypothesis. [*Journal article name written in Chinese*], (2): 133-149.
- Yong, A. G. & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79-94.