
Dissertations

Summer 8-10-2017

Online Formative Assessments as Valid Correlates of Foreign Language Proficiency Levels as Measured by ILR/DLPT5 Summative Tests

Alma Castro-Peet
cast5708@mail.brandman.edu

Follow this and additional works at: https://digitalcommons.umassglobal.edu/edd_dissertations



Part of the Bilingual, Multilingual, and Multicultural Education Commons, Curriculum and Instruction Commons, Educational Assessment, Evaluation, and Research Commons, Educational Psychology Commons, and the Language and Literacy Education Commons

Recommended Citation

Castro-Peet, Alma, "Online Formative Assessments as Valid Correlates of Foreign Language Proficiency Levels as Measured by ILR/DLPT5 Summative Tests" (2017). *Dissertations*. 130.
https://digitalcommons.umassglobal.edu/edd_dissertations/130

This Dissertation is brought to you for free and open access by UMass Global ScholarWorks. It has been accepted for inclusion in Dissertations by an authorized administrator of UMass Global ScholarWorks. For more information, please contact christine.bombaro@umassglobal.edu.

Online Formative Assessments as Valid Correlates of Foreign Language Proficiency
Levels as Measured by ILR/DLPT5 Summative Tests

A Dissertation by
Alma Sandra Castro-Peet

Brandman University
Irvine, California
School of Education

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Education in Organizational Leadership

August 2017

Committee in charge:

Keith Larick, EdD, Committee Chair

Carol Anderson-Woo, EdD

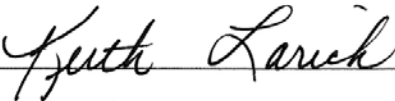
Carlos V. Guzman, PhD


BRANDMAN UNIVERSITY


Chapman University System

Doctor of Education in Organizational Leadership

The dissertation of Alma Sandra Castro-Peet is approved.


_____, Dissertation Chair
Keith Larick, EdD


_____, Committee Member
Carol Anderson-Woo, EdD


_____, Committee Member
Carlos V. Guzman, PhD


_____, Associate Dean

August 10, 2017

Online Formative Assessments as Valid Correlates of Foreign Language Proficiency

Levels as Measured by ILR/DLPT5 Summative Tests

Copyright © 2017

by Alma Sandra Castro-Peet

ACKNOWLEDGEMENTS

Through my job as a Spanish language instructor for the Defense Language Institute Foreign Language Center (DLIFLC) Distance Learning Division, I learned about DLIFLC's commitment to foreign language instruction through its deployment of instructors in any place of need worldwide. My assignments took me to distant and unusual places where I could work on my dissertation, which was any location that had Wi-Fi, including fast food places, libraries, a charming oyster restaurant next to the Alabama River, and the exquisite Joslyn Museum's Café Durham in Omaha, Nebraska.

The Online Diagnostic Assessment (ODA) developed by the DLIFLC is akin to the hidden gems I encountered during my teaching assignments. Not very well known in the United States, it is the only online formative assessment available that competes in scope, design, and complexity with its European counterpart DIALANG. While DIALANG has a plethora of research studies available, only a few researchers have studied the ODA. This project is a labor of love for assessment development and foreign language instruction. Having worked at CTB/McGraw-Hill for over 16 years, I witnessed excitement about the potential of online diagnostic assessments in the private sector. I also noticed the reduction of funds available to develop foreign language assessments. Perhaps one of the reasons for this reduction is a diminished interest in foreign language instruction in the U.S. educational system, as I discovered through this research. The DLIFLC surmounted the lack of guidelines and foreign language assessment tools and introduced itself into the history of foreign language assessment development with its contribution to the development of Interagency Language Roundtable (ILR) standards and the creation of foreign language summative assessments such as the Defense

Language Proficiency Test 5 (DLPT5) and foreign formative assessments like the ODA. In this context, I want to thank the provost at DLIFLC, Dr. Betty Lou Leaver, for reminding me of DLI's dedication to foreign language development and instruction and for inspiring me to keep this mission as the central part of my doctoral studies and dissertation. I want to thank my dissertation chair, Dr. Keith Larick, for having the insight to see the potential contribution of my dissertation, for his assistance in helping me build the best dissertation committee I could ask for, and for taking me under his wing when this project seemed to halt when I had to find my fourth dissertation chair. Fourth time's the charm! What a privilege to have the Brandman doctoral program chair guide my dissertation! Dr. Larick's availability despite his many other important priorities demonstrated to me the true meaning of leading by example. I also thank my awesome, dynamic, and highly gifted dissertation committee: Dr. Carol Anderson-Woo and Dr. Carlos V. Guzman. A doctoral candidate does not truly complete the rite of passage through the rigors of a doctoral thesis without a committee tearing its content apart. Thank you to the Director of Accountability Dr. Anderson-Woo for her critical eye. Her expertise in assessment and accountability was a tremendous asset to my dissertation.

My gratitude goes to the Dean of Distance Learning in the DLIFLC Directorate of Continuing Education, Michael Vezilich, who appreciated my passionate desire to understand the subtleties of the ODA by hand delivering the information I needed, despite his very busy schedule, one day in Alabama. Thank you. As a student in Organizational Leadership, I have seen in Distance Learning the highest expression of servant leadership and mentoring from a department that refers to its employees as family.

Speaking of family, I want also to thank my husband, Ronald Creighton Peet, my best friend, who never hesitated to support me during my dissertation process by putting everything he was doing aside, and I mean everything, to provide timely encouragement. Thank you! Thank you also to the source of my inspiration, my daughter, Victoria Brandon Peet, a very dedicated student at an Ivy League university who emboldens me to strive as strongly as she does and who reminds me that the American Dream to achieve the impossible can still be fulfilled by those who work hard and strive to do their best. I also extend thanks to my mother, Rosalina García Gomez, and father, José Gabriel Castro Cisneros, both teachers, who infused in me a love for learning. To my brother José Felipe Castro García and sister Lillián Carmina Castro García, thank you for being my daily inspiration in the way you handle life's challenges and blessings. Most important, thank you to my beloved God, my light and my source of strength, the source of my joy.

ABSTRACT

Online Formative Assessments as Valid Correlates of Foreign Language Proficiency

Levels as Measured by ILR/DLPT5 Summative Tests

by Alma Sandra Castro-Peet

Purpose: This study explored a technological contribution to education made by the Defense Language Institute Foreign Language Center (DLIFLC) in the formative assessment field. The purpose of this quantitative correlational study was to identify the relationship between online formative (Online Diagnostic Assessment; ODA) and summative (Defense Language Proficiency Test 5; DLPT5) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success in a basic course program for adult students at the DLIFLC.

Methodology: This nonexperimental correlational study included a standard regression model to determine correlations between ODA scores and DLPT5 final scores through a Pearson product–moment correlation.

Findings: Findings were as follows: (a) Category IV languages showed higher discrimination across levels than did a Category I language; (b) the ODA has a closer relationship to the DLPT5 for reading than for listening; (c) listening scores tend to consistently fall one to two levels lower than DLPT5 at Interagency Language Roundtable (ILR) Levels 3 and 2+; and (d) both reading and listening tend to have a consistent moderate relationship between the ODA and the DLPT5 at ILR Level 2.

Conclusion: Because the literature review revealed a disconnect between theory and practice when looking at formative and summative assessments, and because research

results showed that at least one ODA assessment demonstrated a higher degree of correlation (and score differentiation across ILR levels), the conclusion was that it is possible to devise assessments with dissimilar design constructs—formative and summative—but with common ILR requirements that, if designed appropriately, lead to comparable ILR results. Therefore, DLIFLC leaders are highly encouraged to devise similar ODA–DLPT5 correlations and benefit from the results of this research.

Recommendations: ODA developers and research experts need to study reasons for variance in correlation at upper ILR levels for listening as well as the differences between Category I and Category IV languages while considering (a) open-ended responses written in the English language, (b) the ODA semiadaptive features, (c) testing times, (d) differences between formative and summative assessments constructs, and (e) unique idiosyncrasies for assessing listening.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION.....	1
Background.....	3
Statement of the Research Problem.....	8
Purpose Statement.....	10
Research Questions.....	10
Significance of the Problem.....	11
Definitions.....	12
Delimitations.....	15
Organization of the Study.....	16
CHAPTER II: REVIEW OF THE LITERATURE.....	17
Review of the Literature.....	17
Assessment Theory.....	22
Summary.....	71
CHAPTER III: METHODOLOGY.....	74
Overview.....	74
Purpose Statement.....	74
Research Questions.....	74
Research Design.....	75
Population.....	80
Sample.....	81
Instrumentation.....	84
Data Collection.....	99
Data Analysis.....	101
Limitations.....	104
Summary.....	106
CHAPTER IV: RESEARCH, DATA COLLECTION, AND FINDINGS.....	107
Overview.....	107
Purpose Statement.....	108
Research Questions.....	108
Research Methods and Data Collection Procedures.....	109
Population.....	110
Sample.....	110
Demographic Data.....	115
Presentation and Analysis of Data.....	115
Summary.....	165
CHAPTER V: FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS.....	168
Purpose Statement.....	169
Research Questions.....	169
Research Methods and Data Collection Procedures.....	170
Population.....	170
Sample.....	171

Major Findings.....	171
Unexpected Findings	179
Conclusions.....	180
Implications for Action.....	186
Recommendations for Further Research.....	191
Concluding Remarks and Reflections.....	194
REFERENCES	197
APPENDICES	231

LIST OF TABLES

Table 1: DLPT5 and ODA Archived Scores Used for Study	80
Table 2: Excel Document Format for Data Delivery	82
Table 3: DLPT5 and ODA Archived Scores Used for Study	83
Table 4: Data Collection Instruments	84
Table 5: ODA Number of Testlets.....	86
Table 6: Theta Cut-Scores Based on the 70% Mastery Criterion.....	94
Table 7: Data Available from Second Data Pull.....	100
Table 8: Data Analysis.....	103
Table 9: DLPT5 and ODA Data Available.....	112
Table 10: DLPT5 and ODA Archived Scores Used for Study	114
Table 11: Data Available From Second Data Pull.....	116
Table 12: Student Sample	117
Table 13: DLPT5 and ODA Score Nomenclature	117
Table 14: DLPT5 and ODA Coding System	118
Table 15: Correlation Coefficient Values.....	118
Table 16: Correlation per Language for Listening	119
Table 17: Correlation per Language for Reading	120
Table 18: Organization of ODA Scores per ILR Level	121
Table 19: ODA Listening Relationship to the ILR Levels per DLPT5	127
Table 20: ODA Reading Relationship to the ILR Levels Per DLPT5.....	136
Table 21: Predominant ILR Listening Levels on the ODA per DLPT5	148
Table 22: Predominant ILR Listening Levels on the ODA per DLPT5	149
Table 23: Predominant ILR Reading Levels on the ODA per the DLPT5.....	157
Table 24: Predominant ILR Reading Levels on the ODA per the DLPT5.....	158

Table 25: Correlation Results for Listening.....	174
Table 26: Correlation Results for Reading	174
Table 27: ODA Listening Relationship to the ILR Levels per DLPT5	175
Table 28: ODA Reading Relationship to the ILR Levels per DLPT5	175
Table 29: Predominant ILR Listening Levels on the ODA per the DLPT5	178
Table 30: Predominant ILR Reading Levels on the ODA per the DLPT5	178

LIST OF FIGURES

Figure 1. The assessment triangle.....	23
Figure 2. The four spheres of work in educational assessment practice in a schema for appraising the current state of affairs.	34
Figure 3. ODA sessions by language by year.	63
Figure 4. Computer adaptive features of the ODA.	65
Figure 5. ODA subject area breakdown example.	67
Figure 6. ODA/DLPT5 data analysis.....	71
Figure 7. A schema for appraising the current state of affairs.....	76
Figure 8. DLPT5 item pools at ILR levels 1 and 1+.....	91
Figure 9. DLPT5 item pools at ILR levels 1, 1+, and 2.....	92
Figure 10. DLPT5 item pools at ILR levels 1, 1+, 2, and 2+.	92
Figure 11. DLPT5 item pools at ILR levels 1, 1+, 2, 2+, and 3.	93
Figure 12. The ODA and the DLPT5 scores of the same ID code representing a student.....	100
Figure 13. DLPT5 and ODA data pool score matches for listening.	112
Figure 14. DLPT5 and ODA data pool score matches for reading.....	113
Figure 15. Student sample per language.	113
Figure 16. Excel spreadsheet data columns.	116
Figure 17. Total Spanish ODA and DLPT5 score comparison—listening.....	122
Figure 18. Total Korean ODA and DLPT5 score comparison—listening.....	122
Figure 19. Total Chinese Mandarin ODA and DLPT5 score comparison—listening.....	123
Figure 20. Total Standard Arabic ODA and DLPT5 score comparison—listening.	123
Figure 21. Total Spanish ODA and DLPT5 score comparison—reading.	124
Figure 22. Total Korean ODA and DLPT5 score comparison—reading.	124
Figure 23. Total Chinese Mandarin ODA and DLPT5 score comparison—reading.....	125

Figure 24. Total Standard Arabic ODA and DLPT5 score comparison—reading	125
Figure 25. ODA relationship to the ILR—Listening Spanish.	128
Figure 26. ILR percentage distribution per DLPT5—Listening ODA Spanish.	129
Figure 27. ODA relationship to the ILR— Listening Korean.	130
Figure 28. ILR percentage distribution per DLPT5—Listening ODA Korean.	131
Figure 29. ODA relationship to the ILR—Listening Chinese Mandarin.....	132
Figure 30. ILR percentage distribution per DLPT5—Listening ODA Chinese Mandarin.	133
Figure 31. ODA relationship to the ILR—Listening Standard Arabic.	134
Figure 32. ILR percentage distribution per DLPT5—Listening ODA Standard Arabic.	135
Figure 33. ODA relationship to the ILR—Reading Spanish.	138
Figure 34. ILR percentage distribution per DLPT5—Reading ODA Spanish.	138
Figure 35. ODA relationship to the ILR—Reading Korean.	140
Figure 36. ILR percentage distribution per DLPT5—Reading ODA Korean.	140
Figure 37. ODA relationship to the ILR—Reading Chinese Mandarin.	142
Figure 38. ILR percentage distribution per DLPT5—Reading ODA Chinese Mandarin.	142
Figure 39. ODA relationship to the ILR—Reading Standard Arabic.....	144
Figure 40. ILR percentage distribution per DLPT5—Reading ODA Standard Arabic...	144
Figure 41. Predominant ILR listening levels on the ODA per DLPT5.	146
Figure 42. Total ODA and DLPT5 score comparison—Listening Spanish.	150
Figure 43. Relationship between the ODA and the DLPT5—Listening Spanish.	150
Figure 44. Total ODA and DLPT5 score comparison—Listening Korean.	151
Figure 45. Relationship between the ODA and the DLPT5—Listening Korean.	152
Figure 46. Total ODA and DLPT5 score comparison—Listening Chinese Mandarin....	153
Figure 47. Relationship between the ODA and the DLPT5—Listening Chinese Mandarin.	153
Figure 48. Total ODA and DLPT5 score comparison—Listening Standard Arabic.....	155

Figure 49. Relationship between the ODA and the DLPT5—Listening Standard Arabic.	155
Figure 50. Predominant ILR reading levels on the ODA per DLPT5.	156
Figure 51. Total ODA and DLPT5 score comparison—Reading Spanish.	159
Figure 52. Relationship between the ODA and the DLPT5—Reading Spanish.	160
Figure 53. Total ODA and DLPT5 score comparison—Reading Korean.	161
Figure 54. Relationship between the ODA and the DLPT5—Reading Korean.	161
Figure 55. Total ODA and DLPT5 score comparison—Reading Chinese Mandarin.	162
Figure 56. Relationship between the ODA and the DLPT5—Reading Chinese Mandarin.	163
Figure 57. Total ODA and DLPT5 score comparison—Reading Standard Arabic.....	164
Figure 58. Relationship between the ODA and the DLPT5—Reading Standard Arabic.	164
Figure 59. Predominant ILR listening levels on the ODA per DLPT5.	177
Figure 60. Predominant ILR reading levels on the ODA per DLPT5.	177

CHAPTER I: INTRODUCTION

In the 21st century, one of the constants of technology is change. The use of computer technology has reshaped many aspects of daily life (Thayer, 2013), as well as revolutionized teaching in classrooms and the type of accountability measures teachers employ for instruction (Taghizadeh, Alavi, & Rezaee, 2014). The military has been at the forefront in developing and implementing technological innovations that have become part of daily life, including the Internet, the computer, and the global positioning system (Singer, 2014).

The Defense Language Institute Foreign Language Center (DLIFLC) is a leader in foreign language education in the United States (Bergin, 2002; Panetta, 2011; Shin, 1999). The DLIFLC has contributed a variety of technology-based learning tools that are available for free to anyone interested in learning a second language, but their specific intent is to meet the needs of students preparing for deployment or students training to become linguists. The DLIFLC's technology-based products range from cultural awareness components to interactive learning tools that teach the basics of 32 target languages via a program called Headstart to computer-assisted language tools that independent learners can use without an instructor to improve their reading and listening skills in 40 languages through a program called the Global Language Online Support System (GLOSS).

The Online Diagnostic Assessment (ODA) is one of the technological contributions made by the DLIFLC. Learners can use the diagnostic-based formative assessment to evaluate their own learning progress to achieve their educational goals

based on established curriculum criteria (Andrade, Du, & Mycek, 2010; Radford, 2014; Taghizadeh et al., 2014).

Diagnostic assessments are instruments that identify students' strengths and areas of growth in learning to identify the adequate procedures for learning improvement (Alderson, 2005). Diagnostic assessments relate to the set of strategies devised to identify students' strengths and weaknesses (Alderson, 2005). The use of diagnostic assessments is common in such specialized areas as psychological research, mathematics, and physics. However, the pedagogical applications of diagnostic assessment for language instruction had not been studied until the 21st century (Ableeva, 2010; Antón, 2003, 2009; Croteau, 2014; Harding, Alderson, & Brunfaut, 2015; Lantolf & Poehner, 2004; Poehner, 2005), and research on listening has been sparse (Harding et al., 2015).

Researchers have reported findings on the effectiveness of online proficiency assessments in second language acquisition in Europe and the United States (Bachman & Clark, 1987; Berman, Whitt, & Salyer, 2008; Burwell, González-Lloret, & Nielson, 2009; Clark et al., 2014, Taghizadeh et al., 2014). Alderson and Huhta have reported that a true foreign language diagnostic test does not exist except for DIALANG (Alderson, 2005; Alderson & Huhta, 2005, 2011; Huhta, 2008). DIALANG is an online diagnostic assessment that tests students' reading, listening, writing, grammatical, and vocabulary skills in 14 European languages. This online diagnostic test was based on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). This online diagnostic assessment provides relatively limited diagnostic value because the basis of its design is the traditional concepts of listening, speaking, reading, and writing language from the CEFR (Alderson & Huhta, 2011). The focus of CEFR is

traditional taxonomies used in assessment, such as Bloom's taxonomy, rather than a theory of foreign language acquisition and use, which requires the identification of specific areas of strength and growth at a granular level that allows instructors to effectively implement customized learning instruction (Alderson & Huhta, 2011). According to Alderson and Huhta (2011), creating a true foreign language diagnostic assessment would require not only taxonomical measurements, but also phonological, morphological, syntactical, lexicological, and others in the context of second language acquisition.

The information resulting from DIALANG may not be relevant for learners studying a foreign language in the United States with a design based on the CEFR and not the American Council on the Teaching of Foreign Languages [ACTFL] guidelines (ACTFL, 2012; Clark, 2013). According to Alderson and Huhta (2011), CEFR has a greater focus on traditional taxonomies used in assessment, such as Bloom's taxonomy, rather than on a theory of foreign language acquisition that would require taxonomical measurements as well as phonological, morphological, syntactical, and lexicological criteria in the context of second-language-acquisition learning. This information makes it highly relevant to study an online diagnostic assessment developed in the United States such as the ODA, designed by the DLIFLC. This online diagnostic tool tests the foreign language skills of students in the United States and provides a yet-to-be-determined potential for new contributions to the field of formative assessments.

Background

Foreign language instruction has experienced a steep increase in the number of computer-based technologies designed to learn a second language, such as Duolingo,

Memrise, Pimsleur, LiveMocha, and Rosetta Stone. The number of free second language interactive learning tools available has also increased considerably. The Open Culture website alone has a collection of hundreds of free lessons in 48 languages. Therefore, there is a fair amount of literature regarding interactive learning tools and computer-assisted learning and its effect on language learning (Chen et al., 2004; Hubbard & Levy, 2006; Silye & Wiwczaroski, 2002; Son, 2008). According to McClanahan (2014), technology is particularly beneficial for second language acquisition because it delivers authentic materials in the format of videos, webpages, and audio recordings that support the acquisition of a second language in real-world contexts. New technologies provide automated ways to measure learning that help analyze the mastery of skills acquired, as well as the effectiveness of teaching (Alade & Buzzetto-More, 2006; Vendlinski & Stevens, 2002). According to Silye and Wiwczaroski (2002), new types of assessment instruments have surfaced on the Internet and have become more accessible to instructors and students. Assessments available on the Internet have many benefits. For example, the HTML format of the web permits the delivery of an entire test or a series of individual items. Test takers can answer test questions on their computers, send their responses back to the server through Internet browsers, and receive immediate feedback directly from the instructor or the organization overseeing the test administration. The feedback can be delivered with a predetermined script or an overall score available online after the test is complete (Silye & Wiwczaroski, 2002; Taghizadeh et al., 2014). In this context, information technologies provide high levels of flexibility in the design of assessment instruments and the delivery of results for traditional-item and passage formats or alternative assessments with open-ended questions, rubric scoring, pre- and posttesting,

and diagnostic testing (Alade & Buzzetto-More, 2006; Bennett, 2001, 2004). In the case of online courses without face-to-face interaction, some researchers recommend criterion-referenced performance-based language assessments, which can ensure accountability and a deep understanding of concepts and lead to reliable inferences on foreign language ability and instruction (Blake, 2009; Chapelle & Chung, 2010).

An increased awareness of the value of multiple measures of assessment created conditions for developing alternative classroom assessments, which included self-assessments, peer assessments, classroom observations, and student portfolios and interviews (Bachman, 2002; Bachman & Clark, 1987; Butler & Lee, 2010). New assessments such as formative assessments allow learners to judge their own learning progress and help them identify the best way to achieve their educational goals based on the established curriculum criteria (Andrade et al., 2010; Radford, 2013, Taghizadeh et al., 2014).

Myers (2008) described two central types of assessments administered in the classroom: formative and summative assessments. These two types of assessments have some clear differences, such as a goal to summarize what students know after instruction for summative assessments, while formative assessments provide diagnostic information through a school program to target instruction. Instructors use formative assessments to identify specific areas of improvement throughout a course to guide students and instruction and administer summative assessments at the end of a course (Sato & Atkin, 2006). Summative assessments tend to have nationwide implications and impact, whereas formative assessments have local, classroom, or individual outcome consequences (Gardner, Harlen, Hayward, Stobart, & Montgomery, 2010).

Formative assessments are designed to identify student progress (Bax, Branford-White, Heugh, & Jacoby, 2013) by gathering information about the strengths and weaknesses of a student during a course to devise strategies for customized instruction with the purpose of continuous learning (Atkin & Sato, 2005; Boston, 2002; Sadler, 1989). A summative assessment identifies a student's degree of learning by comparing how a student compares to other students and provides measureable assumptions regarding a student's knowledge of a subject learned to authenticate that the student has met the learning requirements (Atkin, Black, & Coffey, 2001; Pellegrino, 2014).

In 2004, Pellegrino identified four independent spheres that help describe the theories that have contributed to the types of assessments available. Pellegrino selected two categories, (a) theory and research and (b) educational practice, to differentiate the construct of classroom-based assessments (or formative assessments) and large-scale assessments (or summative tests). Pellegrino suggested that cognition theory and research influence formative, classroom-based assessments, and psychometric constructs influence summative, large-scale assessments. Pellegrino noted that cognition theory contributed to the progress made in developing formative classroom assessments in support of learning and asserted that formative classroom-based assessments and summative large-scale assessments do not contribute to each other's theories in the implementation of their respective assessment constructs because of fundamental differences between cognitive and psychometric theories with regard to large-scale assessment. Although psychometric theories are necessary in summative, large-scale testing to provide a quantitative measure of learning, assessments based on cognitive theories such as formative assessments tend

to require the output of individualized information at a more granular level (Black & Wiliam, 2004).

The relevancy and application of cognitive-based assessments have grown since the first publication of their utility by Black and Wiliam in 1998. Pellegrino (2012) and the Committee on Developing Assessments of Science Proficiency in K-12 for the National Research Council of the National Academies (2014) gave a glimpse into the possible future of assessments, with a unique application of cognitive-based approaches. Using science area studies as an example, this group recognized that the Next Generation Science Standards (NGSS) require instructors to change the way they teach science. As a result, curriculum, instruction, and assessment will need to be interconnected in every aspect of science education. What is meaningful about the challenges found on the NGSS is that the recommendations included multiple assessments or assessment tasks to identify students' mastery. In addition, any specific assessment task could assess more than one standard or performance expectation. The Next Generation Science recommendations in test design included (a) having multiple components that reflect the interconnectedness of different disciplines within science, (b) addressing the natural learning continuum of students, (c) providing information about the specific beginning and ending points of particular learning units, (d) having a system that allows for the interpretation of student responses at different levels of performance, and (e) providing information to assist educators in the next step of instruction at an individual level. Pellegrino and the National Research Council of the National Academies described a sophisticated version of a new generation of formative diagnostic assessments that emerged at the beginning of the 21st century. Anton (2009) described diagnostic

assessments as a complement to standardized assessments because of their unique conceptualization criteria based on Vygotsky's cognitive development theories.

According to Anton, one of the reasons cognitive development theories are at the core of diagnostic assessments is Vygotsky's theory of zone of proximal development (ZPD).

According to Vygotsky, ZPD is the point in which learning takes place. It is the gap between what a student is able to do independently and what a student is able to achieve with the assistance of an instructor (Vygotsky, 1978). Each gap or learning progression includes current stages and next stages of learning that are an inherent aspect of strongly designed formative assessments (Carpenter, Fennema, & Franke, 1996; Griffin & Case, 1997; Pellegrino, 2014). Therefore, the ideal focus of a diagnostic test is providing an evaluation of what a student is able to do and providing recommendations of the proximal skills that will allow the student to go to the next level of performance growth. A diagnostic assessment that does not include a specified diagnosis of the proximal skills to learn would not take into account the interaction with instructional measures to prepare the student for the next phase in the learning process, which is an essential component (Lidz, 1987).

Statement of the Research Problem

Identifying and building the foreign language expertise of military personnel has required U.S. Department of Defense (DoD) leaders to provide foreign language training, monetary incentives, and reliable standardized testing procedures to ensure the appropriate qualifications of military staff (Christensen, 2013). The DoD language-training program has also required increased linguistic proficiency requirements to graduate. In 2017, the graduation criteria at the DLIFLC were raised to the minimum

achievement score of 2+ in listening and 2+ in reading on the summative Defense Language Proficiency Test 5 (DLPT5; DLIFLC, 2015e). The efforts to meet the increased graduation standards require reliable assessment instruments such as the predictive Defense Language Aptitude Battery test (DLAB) and the summative DLPT5, which help in placement and estimate expected student outcomes at the end of a course program, respectively. These efforts also require the use of descriptive diagnostic measures to know if a student is acquiring sufficient language during the course and is ready to meet higher language requirements with the help of assessment tools such as the ODA. This formative assessment tool provides descriptive information about the next level of learning needed to cross the threshold to the subsequent skills required toward foreign language acquisition. In this context, the ODA is one of the essential components for DLIFLC students. Although researchers know about the DLAB and the DLPT5 through published research studies, little is known about the ODA, also developed by DLIFLC. Multiple regression studies and linking studies have been published for the DLAB to identify its role to predict student success (Anderson, 1997; Wong, 2004). There are also published research studies about the DLPT, which is a summative test that estimates proficiency level, along with full accreditation statements regarding its psychometric qualities (DoD, 2009). However, researchers have not fully studied the properties of the ODA as a formative diagnostic test through published correlation or validation studies. Without validating the ODA as a tool that identifies progress toward the next level of proficiency, a critical formative assessment that could identify if a student is acquiring sufficient language to meet higher requirements may not be used to its full potential. Although DLIFLC has made a tremendous effort to develop a

substantial online diagnostic assessment tool in multiple languages, verifying its validity through this research could lead to using the ODA to its full potential.

The lack of research on the ODA is understandable when looking at the history of assessment in the United States. Most online diagnostic assessment research studies are based on online diagnostic instruments not related to second language acquisition. Leaders, educators, and researchers in highly specialized areas such as psychological research, mathematics, and physics have widely implemented diagnostic assessments and assessed their benefits. However, the pedagogical applications of diagnostic assessment for language instruction had not been studied until recent years (Ableeva, 2010; Antón, 2003, 2009; Lantolf & Poehner, 2004; Poehner, 2005). Although new studies include findings regarding the effectiveness of second-language-acquisition online proficiency assessments, mostly in Europe (Berman et al., 2008; Burwell et al., 2009; Clark et al., 2014, Taghizadeh et al., 2014), the number of studies is still very small.

Purpose Statement

The purpose of this quantitative correlational study was to identify the relationship between online formative (ODA) and summative (DLPT) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success in a Basic Course program for adult students at the DLIFLC.

Research Questions

1. What is the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered at the end of the course and students' final summative DLPT5 scores?

2. What is the relationship between the ODA and the Interagency Language Roundtable (ILR) levels for Spanish, Korean, Chinese Mandarin, and Standard Arabic as measured by the DLPT5?
3. Are the relationships found between ODA and DLPT5 for Spanish, Korean, Chinese Mandarin, and Standard Arabic consistent across the levels or is there variance in the relationship depending on the level?

Significance of the Problem

At DLIFLC, one of the critical requirements of instructors and managers of linguists is to identify individualized remedial procedures for students with a wide variety of linguistic needs. With the recent increase of graduation requirements at DLIFLC to 2+ in reading and 2+ in listening, the appropriate use of the ODA could support DLIFLC in achieving these goals by leveraging the ODA diagnostic information available in 18 languages to customize instruction to meet individual learning requirements. The lack of published research available on the ODA has skewed the understanding of this tool and its impact in the United States, despite the fact that over 35,000 users, mostly from the military, take the ODA each year (DLIFLC, 2015d). The potential for new contributions by studying the ODA is considerable given the breadth and scope of the ODA because the ODA provides diagnostic assessments for listening in 17 languages and for reading in 13 languages specifically tailored to the needs of students learning a foreign language in the United States using the ACTFL criteria: the ILR standards. While research studies regarding an online diagnostic instrument based on the CEFR exist, there is a paucity of research on examining foreign language acquisition via online diagnostic assessments developed in the United States. Additionally, although online diagnostic assessments

provide information to determine different proficiency levels in current language skills and future language needs (Clark et al., 2014), the full use of the ODA may not have been tapped in DLIFLC language schools partially because there are not enough research studies published about this instrument. Because instructors' perceptions of an assessment play an important role in effectively implementing an assessment tool, the results from this study could contribute to the further validation of the ODA and help instructors verify its correlation to the DLPT5 to guide instruction and close the learning gap. Investigating whether a relationship exists between formative and summative assessments in foreign language through this research provides new knowledge. This research contributes to academic studies in the field of second language acquisition by looking at the relationship between foreign language instruction formative online diagnostic tests and summative assessments to determine the validity of foreign language diagnostic tools to estimate student success.

Definitions

Computer adaptive test (CAT): An assessment that uses computerized algorithms to modify test content to correspond to the abilities of the test taker. A CAT requires a large pool of items and passages to identify the specific level of abilities of the test taker (Data Recognition Corporation, 2013; "The Glossary of Education Reform," 2014).

Criterion-referenced test: A test that yields detailed data about the specific competencies of a student (Zhou, 2010). A criterion-referenced test is different from a norm-referenced test, in that the student score is compared to the clearly delineated standards rather than the scores of the rest of the population who took the test (Clark et al., 2013).

Defense Language Proficiency Test (DLPT): A summative assessment developed to measure the foreign language proficiency in reading and listening of students whose first language is English. The test identifies civilians and military language analysts who may be eligible for salary incentives or operational deployment for specific linguistic assignments or determines training decisions (DLIFLC, 2015b).

Diagnostic assessment: An assessment designed to obtain reliable data about the strengths and weaknesses of a learner on a specific skill (Zhou, 2010). The diagnostic feedback provided should emphasize specific strategies for future improvement rather than a mere summary of weaknesses (Harding et al., 2015). A strongly designed diagnostic assessment includes (a) comprehensive observations about strengths and areas of growth, (b) a construct design that allows for a series of evaluations in a continuum starting with the observations and tools available that include help resources, and (c) information that will help test takers succeed at the next level of diagnostic evaluation (Alderson et al., 2014).

Formative assessment: An evaluation tool that allows the gathering of information about the strengths and weaknesses of a student during a course to devise strategies for customized instruction with the purpose of continuous learning (Atkin & Sato, 2005; Boston, 2002; Sadler, 1989). Formative assessments might vary, but have a similar approach in that they are designed to identify student progress (Bax et al., 2013).

Interagency Language Roundtable Skill Level Descriptions (ILR): Provides criteria to measure language proficiency in reading, speaking, listening, writing, translation, interpretation, and intercultural communication. The descriptors specify

predictable capabilities that are common at different stages of the foreign language learning development process (ILR, 2015).

Norm-referenced test: A test designed to provide information about a group of students by comparing the test results of each student against the results of all the test takers. This process involves placing the results of all test takers in a scoring range that allows the identification of the abilities of each student relative to the scores of the population of students who took the test (Clark et al., 2013).

Online Diagnostic Assessment (ODA): A web-based assessment instrument that identifies the individual areas of strength and the areas of growth required for a specific learner to advance to the next level of proficiency. The ODA identifies existing language proficiency as well as future proficiency skills (Clark et al., 2013).

Proficiency: The level of mastery based on a set of specified standards usually measured through an evaluation system or assessment (“The Glossary of Education Reform,” 2014).

Test reliability: An essential aspect of the quality of a test associated with the consistency in results when an assessment is administered again to the same group of examinees (Setzer & GED Testing Service, American Council of Education, 2009). Test results should be able to provide meaningful information that permits a comparison of group scores and individual scores at different points in time (Clark et al., 2013).

Summative assessment: A summative assessment can be either norm referenced or criterion referenced. As a norm-referenced test, it can be used at the end of a course or a school program to evaluate if a student or a group of students has met course requirements. In this context, it identifies how a student compares to other students. As a

criterion-referenced test, results are reported based on how well students meet a set of standards and not on how students perform compared to a norm group. A summative assessment identifies the degree of learning and provides measureable assumptions regarding a student's knowledge of a subject learned to authenticate that a student has met the learning requirements (Atkin et al., 2001; Pellegrino, 2014).

Standardization: A set of strategies established to implement the same test-taking conditions for all test takers to ensure the reliability of the test results. By standardizing the development, administration, and testing conditions, the expectations about the test results can be more predictable (Mislevy, 1992).

Validity: The aspect that ensures a test conforms to the skills and abilities taught and expected (Takala, 1998). Validity is derived to some extent by the quality, design, and suitability of the assessment content; if an assessment instrument does not correspond to the criteria, difficulty, and predicted outcome, the test will not be valid (American Educational Research Association, American Psychological Association [APA], & National Council on Measurement in Education, 1999).

Delimitations

The population was delimited to students in the DLIFLC Spanish, Korean, Chinese Mandarin, and Standard Arabic Basic Course in 2014 and 2015. The archived data were the ODA results administered at the end of the program and the DLPT5 summative results administered at the end of the program as part of the graduation requirements.

Organization of the Study

This study consists of five chapters. Chapter II contains the review of literature and the current findings on online foreign language formative assessments and their specific role in foreign language instruction. A review of literature includes the theoretical concepts involved in the development of formative and summative assessments, information research on instructional technology, the history of assessment development in the United States, and a detailed description on the design and conceptualization of the ODA. Chapter III includes an explanation of the research approach and methodology, population, sample, instrumentation, and data analysis. Chapter III includes the rationale for the research design and the procedures for collecting archived data of the formative ODA and the summative DLPT5. Chapter IV presents the findings of the study, an analysis of the data regarding the correlation between formative and summative assessments in foreign language acquisition, and the impact of online formative assessment in providing meaningful information related to foreign language proficiency in reading and listening as measured by a summative test. Chapter V provides a summary of findings, conclusions, and recommendations for further research.

CHAPTER II: REVIEW OF THE LITERATURE

This research study involved examining the relationship between formative and summative assessments in foreign language by looking at the relationship between foreign-language-instruction formative online diagnostic tests and summative assessments to determine the validity of foreign language diagnostic tools to estimate student success. This study also addresses theories for linking assessment instruments, including a discussion on the advantages for validating a formative assessment through a summative test.

This chapter contains the review of literature and presents theoretical concepts involved in the development of formative and summative assessments, along with current findings on formative assessments and their specific role in foreign language instruction. The history of assessment development in the United States is discussed to identify the contribution of DLIFLC in the field of second language acquisition and assessment in the United States. A section is dedicated to the DLIFLC placement test DLAB, the summative test DLPT5, and the diagnostic test ODA, along with its European counterpart, the diagnostic test DIALANG.

Review of the Literature

The Defense Language Institute Foreign Language Center

Certified by the Council for Higher Education and the U.S. Department of Education through the Accrediting Commission for Community and Junior Colleges of the Western Association of Schools and Colleges, the DLIFLC is the DoD's main agency for foreign language training and provides basic, intermediate, and advanced foreign language instruction to every branch of the armed forces (DLIFLC, 2015a). Trained

resources help support the goals of the DoD and provide qualified personnel to meet the requirements of field commanders, embassies, and foreign institutions such as the North Atlantic Treaty Organization (Christensen, 2013).

Over 1,900 language instructors provide training to military students preparing to become linguists for the DoD. The length of instruction for the Basic Course program ranges from 36 to 64 weeks, depending on the language difficulty, for the 23 languages and dialects taught at the institution. Languages are organized into four language-difficulty categories determined by what a native English speaker can understand. In order of difficulty, French, Spanish, and Portuguese are considered Category I; German and Indonesian are Category II; Hebrew, Hindi, Persian Farsi, Russian, Serbian/Croatian, Tagalog, Turkish, and Urdu are Category III; and Standard Arabic, Arabic (Egyptian, Iraqi, Levantine, Sudanese), Chinese Mandarin, Japanese, Korean, and Pashto are Category IV (DLIFLC, 2015c).

DLIFLC started granting over 11,500 associate of arts in foreign language degrees in 2002 after it received federal authorization from the U.S. Congress in October 2001. To maintain its accreditation, DLIFLC must comply with over 120 standards of accreditation (DLIFLC, 2015a; DLIFLC, 2015c). Each calendar year, approximately 3,500 students attend the Basic Course programs available at the DLIFLC Presidio of Monterey (DLIFLC, 2015a). All military service branches (Air Force, Navy, Marines, Special Forces, and Coast Guard) take foreign linguistic training offered at DLIFLC (Hsueh, 2008; St. Pierre, 2008).

The DLIFLC is one of the key sources of foreign language proficiency training in the United States (DLIFLC, 2015). While educational institutions in the United States

historically developed assessment materials primarily using assessment organizations in the private sector (Alade & Buzzetto-More, 2006; Urciuoli, 2005), foreign language proficiency assessment measures in the United States had their origins in the government (Clark et al., 2014). With no foreign language standards available in the 1950s to measure the foreign language skills of people in the United States, the U.S. government through the Foreign Service Institute developed the ILR scale, which is a set of standardized descriptors of foreign language proficiency for listening, reading, speaking, and writing skills. These descriptors were developed to rate the language ability of government employees (Clark et al., 2014; Defense Intelligence Agency, 2015). According to Herzog (2015), due to the lack of a grading system in the United States to measure foreign language competence, the Foreign Service Institute worked with an interagency committee to create a single scale ranging from 1 to 6. This scale rated foreign language fluency under an overall language rating. In 1956, assessment instruments were introduced to measure language proficiency for all Foreign Service officers (Herzog, 2015). According to Herzog, the single scale was adjusted over time to represent different scales for each skill to include six levels ranging from 0 to 5. Zero represented no functional skill or ability, and 5 represented fluent native ability equivalent to that of a highly educated native speaker. In 1985, the ILR Scale was updated to include the + or plus levels of the 0 to 5 scale. These adjustments increased the objectivity and reliability of the ILR Scale (Clark et al., 2014). According to Herzog, the ACTFL validated the ILR scale by publishing proficiency guidelines for academic use based on the ILR criteria. According to Clark et al. (2014), the revisions and standardization strategies implemented

to improve the ILR Scale increased the reliability of the scale and contributed to its use in academia based on the adoption of the ILR scale by the ACTFL (Clark, 2013).

In February 2005, almost 4 years after the terrorist attacks on September 11, 2001, the DoD disseminated the *Defense Language Transformation Roadmap*, which highlighted the strategies required by the DoD to improve the language capability of regional languages and dialects. In 2010, an update to the *Defense Language Transformation Roadmap* had a unique title on the document to be presented to the House of Representatives: *Bearing the Burden of Today's Educational Shortcomings*.

Panetta (1999) stated that, unlike those in most countries, the educational system in the United States does not provide the foreign language training required to allow students to ease their way into the 21st century defined by its globalization. Since September 11, 2001, politicians, educators, and business leaders have recognized the inadequate supply of foreign language expertise in the United States. In this context, the DoD was required to continue to be the main supplier of foreign language resources capable of crossing the linguistic gap with other cultures and responding appropriately to unforeseen dangers in the face of an increasing demand of language capabilities and despite budgetary challenges (N. A. Brown, 2009). This included the need to develop foreign language standardized assessment instruments that appropriately measure the foreign language skills of its military staff.

The State of Foreign Language Acquisition in the United States

The United States and the DoD have been at a disadvantage when it comes to obtaining readily available language expertise to respond to the political challenges of the 21st century. As reported by the 2006 General Social Survey, only 25% of the

respondents declared they know a second language. The percentage of those who speak a second language with mastery is even lower. Although it is compulsory for a student to be fluent in more than one language in Europe, except for Ireland and Scotland, the United States does not have a national policy for foreign language learning (Devlin, 2015). It is important to recognize this discrepancy in foreign language learning priorities in the United States and Europe. A student in the United States learning a second language will probably be an adult student, whereas in Europe, foreign language is compulsory in elementary and middle school, and in some countries such as in Belgium, students learn a second language at age 3 (Devlin, 2015). Although the differences in the developmental age is one of the factors that distinguish foreign language learners in the United States and Europe, another distinction is the way adult students learn a secondary language. Adults already have a set of linguistic tools available from their first language frame of reference, which serves as a frame of reference as they learn a second language (K. McManus, 2015). Another distinction relates to the linguistic characteristics of the first language learned as adults, compared to the differences in the linguistic characteristics of the second language learned. Because of the variations in the lexical and grammatical constructs of a primary language, foreign language learners cannot assume that producing meaning in one language will automatically require similar strategies for producing meaning in a secondary language, particularly when the secondary language learned has grammatically and syntactically different characteristics (Roberts & Liszka, 2013). Consequently, second language acquisition is usually acquired in the context of the linguistic knowledge, cultural understanding, and frame of reference in which the primary language was acquired (Izquierdo & Collins, 2008; Oxford, 2017;

Salaberry, 2008; Skehan, 2014; Skehan & Foster, 1997; Sugaya & Shirai, 2007; Turner, 1993). In this context, while it is not the intent of this chapter to describe the language acquisition strategies and theories of learning a first and second language, it is important to recognize (a) the developmental differences in the second language learners of Europe and the United States; (b) the lack of emphasis on foreign language learning instruction in the United States; and (c) the primary lexical and grammatical differences in the primary language and specific foreign language learned make it pedagogically challenging in the United States to acquire foreign language assessment tools developed for the specific developmental needs, learning strategies, and standards of European foreign language learners.

Assessment Theory

Assessments are instruments developed to gather data that otherwise cannot be observed. These are developed with distinctive design constructs, depending on their intended use. Regardless of their purpose or design, all assessments share a common characteristic: to obtain information about an expected outcome. In this context, the purpose of assessments is to obtain valid and reliable information of what an individual understands and is able to do (Pellegrino, 2014). According to the National Institute for Learning Outcomes Assessment (2014), obtaining student data does not serve a functional purpose if these data do not provide information that could be used for relevant purposes. It is therefore essential that the information obtained from an assessment is meaningful and can be understood from a determined frame of reference (Pellegrino, 2014; Schum, 1978).

Three essential components are needed in the development of assessment instruments, whether formative or summative: (a) cognition, which is a theory that includes well-founded premises regarding the skills and abilities expected from a student; (b) observation, which is a group of tools or precepts that contributes to the evidence for the expected outcomes either through statistical models or through qualitative descriptions; and (c) interpretation, which is an analytical procedure that appropriately interprets the information obtained from the assessment instrument (Committee on the Foundations of Assessment, 2001; Pellegrino, 2014). These three elements (see Figure 1) that are an intrinsic part of any assessment cannot be isolated. The congruent connection of these three elements will determine the quality of an assessment (Pellegrino, 2014).

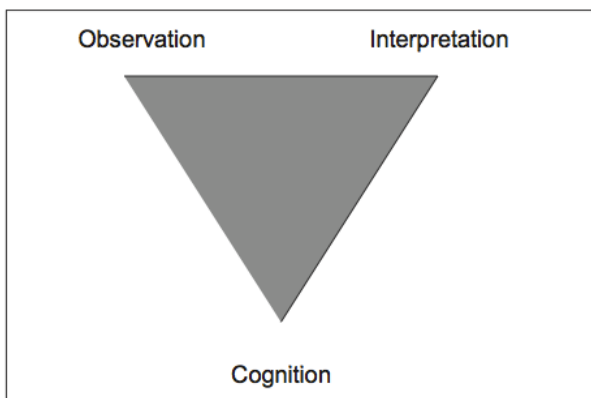


Figure 1. The assessment triangle. From *Knowing What Students Know: The Science and Design of Educational Assessment* (p. 44), by J. W. Pellegrino, N. Chudowsky, & R. Glaser, 2001. Copyright 2001 by the National Academy of Sciences. Reprinted with permission.

Assessments have different aims and designs depending on the official (high-stakes) or nonofficial (low-stakes) outcomes expected. Therefore, an assessment may have higher or lower test design flexibility depending on its purpose. The higher the number of expected outcomes, the more the validity of this assessment may be compromised (Pellegrino, 2014). Consequently, it is necessary to provide substantiated

data that show that each intended outcome for a given assessment is accomplished. Ruiz-Primo, Shavelson, Hamilton, and Klein, (2002) noted that different assessments may need to be interpreted based on their alignment to a goal and their location on a lining order that places them along specific points named: (a) immediate assessments, which include student observations; (b) close assessments, which include classroom quizzes and other informal assessments; (c) proximal assessments, which include specific evaluations with a formal quality related to the classroom curricula; (d) distal assessments, which include criterion-referenced tests and formative assessments such as the ODA; and (e) remote assessments, which include high-stakes assessments or norm-referenced assessments such as the DLPT. Positioning specific assignments on their proper location in this classification may help to understand accurately their specific purpose and their association with other assessments and may help to identify how congruent an assessment is to its specific design and constraints innate to its requirements (National Research Council, 2003; Pellegrino, 2014). Because it is impossible for one type of assessment instrument to fulfill the specific needs of different stakeholders and because there is a need in the education field to provide assessment information for a wide variety of reasons, a suite of reliable and well-crafted assessments designed to fulfill different functions is recommended to evaluate the effectiveness of learning and instruction (Bachman, 2013; Darling-Hammond & Pecheone, 2010; Pellegrino, 2006, 2014; Pellegrino, Chudowsky, & Glaser, 2001). There are three common reasons why assessment instruments are developed: (a) for student placement before a program starts, (b) for diagnostic purposes through the course program, and (c) for accountability at the end of a course program (Ronan, 2015).

Placement Tests

Placements tests can roughly fall into the first point in the lining order (Primo et al., 2002). These are immediate assessments given before or at the beginning of a course program and help identify students' abilities (H. D. Brown, 2004). These evaluations are administered to identify a student's strengths and to avoid student misplacement at a program that may not be appropriate to the level of the learner (Illinois, 2012). An appropriate placement test will help students and school programs ensure a student will have a higher chance of success after being suitably placed in a specific class program (Fulcher, 1997). Thus, it is important that a placement test is valid and reliable. An inappropriate student placement may compromise the opportunities for a student to succeed at a program (Al-Adawi & Al-Balushi, 2016). Validity in placement tests is critical for the success of a student and a school program. Validation studies of placement tests include preestablished metrics to evaluate a test, and the test administration results in large student populations (Scott-Clayton, 2012). According to Belfield and Crosta (2012), a placement test is validated by the criteria set for the school program and how these criteria are implemented in the placement test design, the congruent interpretation of test results for the intended placement purposes, and the pass/fail cutoff score. Lastly, the validity is based on the way the placement tests are used and how this use is consistent with the type, number, and continuum order of courses. A unique characteristic in the validation of a placement test is the student placement based on a cutoff score that applies equally to all students who scored one point or 20 points above a cutoff score (Belfield & Crosta, 2012).

Diagnostic Tests

Diagnostic tests are formative assessments usually administered at the individual, classroom, and local level to discover the strengths and weaknesses of students and to target instruction during a course program appropriately (Black & Wiliam, 2009; Rea-Dickinson & Gardener, 2000). Diagnostic formative instruments provide evidence of unique areas of strength and growth on a set of skills to personalize instruction to the specific needs of a student (Pellegrino, 2014; Popham, 2008). Due to its design as a tool to inform learning and instruction, some researchers describe formative diagnostic assessments as instructional tools rather than assessment tools (Heritage, 2008). In this context, these tools are sometimes considered “assessments for learning rather than assessments of learning” (Stiggins, Arter, Chappuis, & Chappuis, 2009). Because of their requirements to promote learning improvement, these tools may not always contribute to student scores but may have detailed feedback for current and next learning progressions, which is reflected in student reports. These reports may provide information about the next learning progression and consequently may lead to mastery of a skill (Clark et al., 2014). For this reason, these instruments are sometimes described as proximal formative assessments (Erikson, 2007) because of their origin in Vygotsky’s ZPD development theories and the proximal skills that would allow a student to perform at the next set of skills (Lidz, 1987). Learning progressions that include current stages and next stages of learning are an inherent aspect of strongly designed formative assessment instruments (Carpenter et al., 1996; Griffin & Case, 1997; Pellegrino, 2014). These progressions usually include information about the learner’s development toward established skills, the

learner's cognitive process for achieving these skills, and a description of cognitive fallacies that may have led to learning mistakes (Supovitz, 2012).

Formative assessments could vary in test design, test length, and test grouping, as well as in test administration frequency. However, one element that makes a diagnostic assessment formative in nature is that it helps students identify which specific skill modification is required in their cognitive process (Pellegrino, 2014). Because a sound formative assessment identifies the skills required at specific stages in learning, assessment research experts consider the quality and soundness of the framework used as part of the validation process of a formative assessment.

Formative instruments require a design that contributes to the dynamic review of performance feedback and lesson planning based on the continuous tracking of student progress (Bax et al., 2013). Because decisions to support instruction are based on empirical assessment data, it is desirable for the formative evaluation gathering to be a habitual process of assessing learning progressions. Therefore, educational organizations require training instructors to understand and effectively use formative instruments to ensure instruction is appropriately geared toward the specific areas of growth of a student (Pellegrino, 2014; Pellegrino, Baxter, & Glaser, 1999; Stiggins, 1997). The importance of the instructor's proactive initiative to implement instructional strategies per formative assessment feedback cannot be underestimated. The instructor's perception of an assessment plays an important role that may contribute to the impact of a formative instrument toward effectively closing the achievement gap. Although it is not the intent of this study to address how instructors' perceptions may affect the implementation and impact of an assessment, it is important to recognize instructors' essential contribution to

the success of an assessment tool based on their perception of its value and therefore the appropriate implementation of this tool (Fox, 2009; Jang, 2005, 2009). Sadler (1989) identified three components of a successful implementation of a formative assessment as (a) clearly determined instructional goals that are part of the instructional program, (b) assessment information about the strengths and weaknesses of a student, and (c) instructional strategies to ensure growth in the areas for improvement. In this context, effective instruction and student growth can only take place through an appropriate application of formative assessment results.

According to researchers at the Wisconsin Center for Education Research (2009), a strong formative tool should (a) be an ongoing element of instruction, (b) be consistent with the summative assessments of an organization by sharing the same standards and learning targets to provide a tridimensional representation of summative and formative data required on a student, (c) provide meaningful and reliable information to guide content and direction of instruction, and (d) be clearly formulated through obtainable instructional targets.

End-of-Course Assessments

While formative tests are usually administered on an ongoing or periodic basis, summative assessments are generally administered at the end of a course or after completion of a specified block of instruction. This study included only the summative assessments used for final course grade levels. Instructors or learning institutions use end of course summative assessments at the national or state level to evaluate if a student has met the course requirements and to identify how the student compares to other students. These instruments are also used for accountability and certification purposes (Harlen,

2005). Usually a score range is given to identify the mastery of the skills acquired. The essential goal of a summative test is to identify the degree of learning and provide measureable assumptions regarding the student's knowledge of the subject learned and to authenticate that the student has met the learning requirements (Atkin et al., 2001; Pellegrino, 2014). According to Pellegrino (2014), because of their design construct, summative assessments do not require knowing the reason why students may be having difficulty mastering a skill; these tools only need to reveal whether mastery was obtained to perform policy-making decisions. For this reason, stakeholders who are not classroom participants usually administer these assessments. Although these tools are not part of the instructional process, they serve an essential role in measuring the learning process on a large-scale level. The evaluation from summative assessments may not provide the level of granularity usually available in formative assessments. As a result, summative assessments are usually not used for customized instruction. The statistical analysis and norming procedures performed to validate summative assessments provide information that help weigh the test results of a student against a group of other students at a regional, state, or national level (Clark et al. 2014). Summative tests are usually validated through a strict set of psychometric validation procedures that include test specifications comprised of a blueprint with the description of the design construct, the purpose of the assessment, a description of standards addressed specific to the items developed (Leighton & Gierl, 2007), and a test design showing the specific item formats and their corresponding distribution in a set of validated standards (Gierl, 1997; Webb, 2006). These include a description of the item development process, procedures and item formats, strategies for minimizing item bias, a description of the item review process, the

administration process, the student population, student results, data obtained, a description of analysis of statistical data and results, scoring procedures, and a summary of validity evidence (Data Recognition Corporation, 2011-2012). Because of the high-stakes nature of summative assessments, these instruments need to be standardized and to go through strict norming procedures. For this reason, the higher the stakes that result from these tests, the more structured, conservative, and statistically based is the methodology (Rabinowitz, 2011).

Item response theory (IRT) is often used in summative tests that provide quantitative information mostly to assess academic skills in a primary language regarding how a student or groups of students respond to each test question (Yang & Kao, 2014), as well as other quantitative information that includes information about the difficulty of each item in relation to other items (Rasch, 1960). IRT also accounts for statistical information about each item that may be the result of chance (Creswell, 2008) to ensure accuracy of test results. According to Bock (1997), psychological and mathematical statistical estimation theory motivated the development of the IRT, first conceptualized by Louis Leon Thurstone in 1925 as a system to scale psychological and educational tests. This system included common IRT models such as the probability of a student responding correctly to each test item and the location of each item on a quantitative scale. By using this system, Thurstone was able to place items on a graded scale by age. Modern IRT models are one of the most commonly used instruments in testing and commonly rely on student samples to identify probabilities for responding to each test item instead of individual student responses (Bock, 1997). In their chapter on modern approaches to measurement, Sternberg and Grigorenko (2002) and Embertson and Reise

(2000) cited the IRT as one of the most important assessment development instruments, and Hambleton and Slater (1997) noted the practical and theoretical benefits of its implementation. Sternberg and Grigorenko classified the different IRT models available into unidimensional and multidimensional. The former model generally links item difficulty to the probability a test taker of a determined skill level responds to the item accurately and then places it onto a scale, whereas the latter model takes into consideration the diverse skills needed for responding to each individual item, including problem-solving strategies. According to Mislevy (1992), the IRT provides parameters that help estimate the difficulty of each test question, the probability to respond to a test question correctly, and student mastery on the subject. As a result, a well-built assessment following the IRT model will help place students in a quantitative location that will compare them with other students at a local, estate, and national level. This strategy will allow for the development of norm-referenced data. The accuracy of the IRT model and norm-based research will be ultimately based on how closely the questions represent student competencies. Mislevy emphasized that IRT models are estimates and future inferences should be taken with caution because groups of students, standards, pedagogical learning strategies, and motivation change over time at different programs. In this context, the true value of an assessment will be determined by how well it meets its intended function (Black & Dylan, 2003) and how dependable its information is for evaluating either an individual student or a whole language program (Clark et al., 2014). According to the function and purpose of an assessment, a different type of evidence or student outcome may be necessary (Mislevy, 1992).

For an assessment to measure what it intends to measure, assessment literacy is an essential skill required from people assigned to develop assessment materials, along with the correct application of protocols that ensure the validity, reliability, and fairness of an assessment instrument. Whether at a large scale to assess a whole program as in the case of a summative test, or at a classroom level as in the case of a formative instrument, assessment literacy helps developers to create valid instruments. It also helps instructors and organizations select the appropriate assessment instruments for their corresponding intended purposes (Taylor, 2009).

Regardless of whether the test is used for placement, diagnostic, or summative purposes, six traits define a well-crafted assessment instrument: (a) its ability to measure different types of basic and procedural skills and high-order thinking skills; (b) its ability to mirror skills and tasks as closely as possible to the way they will be applied in the real world; (c) its capacity to include content that represents the expected level based on nationally or internationally accepted standards; (d) its inclusion of high-quality items and activities that discriminate between different levels of student performance; (e) its ability to uphold valid, reliable, and fair item development criteria along with accurate and consistent results; and (f) its bias-free qualities that help elicit higher or lower scores from groups or individuals with similar skills and abilities (Pellegrino, 2014). As the focus of this research will be on diagnostic tests that are formative, the following sections will address the formative assessment theory.

Formative Assessment Theory

Formative assessment theory may have its origins with Scriven (1967), who formulated this term to provide evaluation strategies for program improvement (Guskey,

2010). Bloom developed a cognitive taxonomy and used Scriven's term to devise strategies to assess students as part of the instructional program instead of at the end of a course, with the goal of finding individual cognitive needs for instruction (Bloom, Hastings, & Madaus, 1971; Guskey, 2010). The concept of learning as an active process of building knowledge through cognitive strategies inspired the works of sociocultural constructivists such as Vygotsky, who identified the cognitive process as requiring social interactions between students actively learning in small teams and instructors in the role of mediators (Tharp & Gallimore, 1991; Vygotsky, 1978). In this context, Vygotsky suggested that culture and social interaction play a role in learning and that learning is heightened in a social environment (Ash & Levitt, 2003; Koschmann, 1999; Vygotsky, 1978). Vygotsky is most well-known for his ZPD concept, which is at the core of formative assessment development and online diagnostic assessments, as well as on second language acquisition pedagogy, to address perceived second language gaps through systematic forms of instruction (Lantolf & Thorne, 2007). The ZPD is the cognitive-level gap at which a learner can complete a task without support. At the point where a student is unable to complete this task on his or her own, an instructor could mediate the process toward closing the learning gap and assisting in identifying the next learning clusters (Ash & Levitt, 2003; Black & Wiliam, 2009; Lantolf & Thorne, 2007; Walqui & van Lier, 2010). The ZPD helped define what is known in formative assessments as learning progressions. These are descriptions mapped in a continuum to show the developmental learning of different domains over a period of time (Harris, Bauer, and Redman, 2008; Heritage, 2008; Sztajn, Confrey, Wilson, & Edgington, 2012). These learning progressions help to identify key moments in the learning process and

identify with examples, key concepts, and descriptions, the learning acquired at a specific stage of a learning domain and the learning required in order to move to the next area of developmental learning (Wilson & Bertenthal, 2006).

Although some researchers are unsure about how theories of measurement should be applied to formative assessments and to what extent (Bennett, 2011), others recognize the innate differences in construct between formative and summative assessments and the corresponding theoretical differences (Pellegrino, 2004). Pellegrino observed that the formative classroom-based assessments and the summative large-scale tests do not seem to contribute to each other's theories in the implementation of their respective assessment constructs because there are fundamental differences between the cognitive and the psychometric theories due to the different expected outcomes of classroom assessments and large-scale assessments. Figure 2 shows the four spheres of work in educational assessment practice as described by Pellegrino.

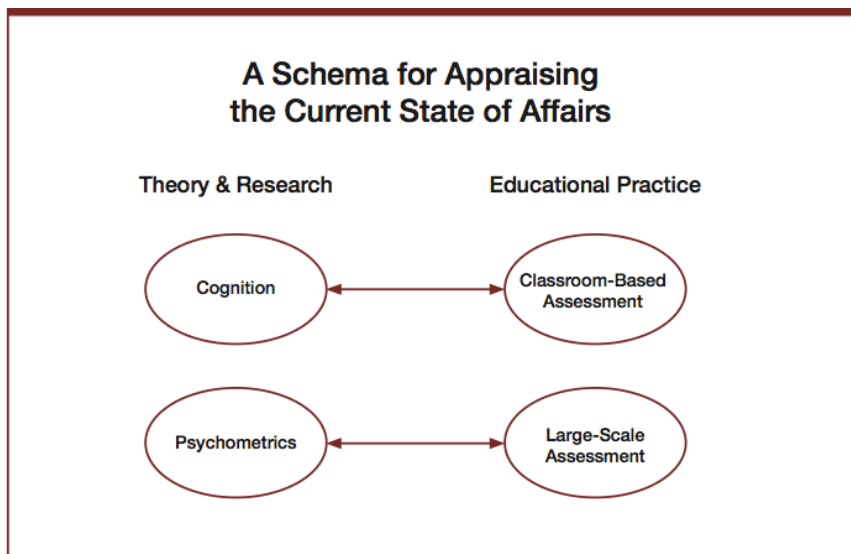


Figure 2. The four spheres of work in educational assessment practice in a schema for appraising the current state of affairs. From *The Evolution of Educational Assessment: Considering the Past and Imagining the Future* (p. 10), by J. W. Pellegrino, 1999, retrieved from <https://www.ets.org/Media/Research/pdf/PICANG6.pdf>. Copyright 1999 by J. W. Pellegrino. Reprinted with permission.

There are best practices for summative assessments, but disagreement exists among researchers about whether to consider these practices when selecting or developing formative tests. The practices include reliability measures that ensure the assessment results are (a) predictable and consistent when administered to students with the same skills and abilities, (b) valid so they measure what they intend to measure and their results lead to suitable instructional decisions, and (c) fair so that students' responses are predictable and consistent across all students (Haertel, 2006; Pellegrino et al., 2001; Trumbull & Lash, 2013). Formative assessments strengthen researchers' assessment constructs through the frequent evaluation of students (Durán, 2011) and therefore ensure their validity and reliability over time through the ongoing gathering of student data as done directly by instructors and the frequent updating of the assessment instruments based on input resulting from the data gathered (Shavelson, Black, Wiliam, & Coffey, 2007). It is therefore suggested that the effectiveness of a formative assessment depends on the successful implementation of the formative test results into relevant instruction and on the ongoing relationship of formative assessment tools with teaching and learning (Frohbeiter, Greenwald, Stecher, & Schwartz, 2011; S. McManus, 2008; Pellegrino, 2014). In this context, the three essential components of a formative assessment described by the Committee on the Foundations of Assessment: (a) a theory regarding the skills and abilities expected from a student, (b) a group of tools or precepts that contributes to the observed evidence for the expected outcomes, and (c) the interpretation of information obtained from the assessment instrument; need an additional component: the appropriate implementation of the information resulting from the assessment into specific and relevant instruction for the learner (Trumbull & Lash, 2013).

The evidence-centered design is a design approach recommended in developing formative tests to show evidence of the quality and validity of its construct (Mislevy, Steinberg, & Almond, 2003; Zhang et al., 2010). Formative assessments developed along a set of learning progressions require (a) the development of specifications for the type of student outcome expected to evaluate determined aspects of student learning; (b) the appropriate evaluation of the assessment activities developed to ensure these tools measure the specific knowledge, skills, and abilities they intend to measure and no other skills; and (c) the appropriate evaluation of the assessment activities developed to ensure they are free of bias (Trumbull & Lash, 2013).

Validating formative assessment constructs includes addressing the quality of their learning progressions, which is not a simple matter. Learning progressions for any domain seem to be the result of complex cognitive and nonlinear processes (Harris et al., 2008; Shavelson & Kurpius, 2012; Steedle & Shavelson, 2009) in which instruction plays an important role. However, the validation of learning progressions through empirical studies is sparse (Trumbull & Lash, 2013), and the studies of learning progressions on domains from different content areas of study seem to be inconsistent in their level of specificity and accuracy (Sztajn et al., 2012). In addition, learning progressions require an understanding of the complexities in the interaction between prior knowledge and new knowledge through appropriate instruction (Shavelson & Kurpius, 2012). Therefore, researchers have acknowledged that the strategies for validating learning progressions are limited. Although the research on learning progressions is at the emergent stage (Corcoran, Mosher, & Rogat, 2009; Shavelson & Kurpius, 2012), it is nevertheless relevant to address learning progressions as an essential component in understanding and

assessing the quality of formative assessments and providing meaningful feedback to students about their learning progress, as well as for devising instructional strategies for the areas of growth (Trumbull & Lash, 2013). While formative assessments require instructors to take steps toward devising lessons and strategies for closing the achievement gap found with the formative assessment, a diagnostic test provides information before instruction and after instruction to identify the size of the learning gap (Perie, Marion, & Gong, 2009). Therefore, formative assessments are usually part of a diagnostic assessment tool and are integrated into a classroom program.

Second Language Acquisition Formative Assessment Theory

Assessing foreign language proficiency at the lower levels of undergraduate school seemed for some time to have focused on student satisfaction and general university requirements (Anton, 2009; Chalhoub-Deville, 1999; Teschner, 1991), whereas oral proficiency tests, writing assessments, student portfolios, and exit oral exams following ACTFL guidelines appeared to be the most common practice at the undergraduate level (Anton, 2009; Glisan & Phillips, 1996). As the undergraduate foreign language course progresses into the third year, formative testing approaches through the application of diagnostic and dynamic testing are usually administered right before the selection of a major through the administration of grammar, vocabulary, listening, reading, writing, and oral interviews. Thus, the application of diagnostic and dynamic assessment techniques to assess second language learners appears to be one of the preferred techniques for identifying individualized foreign language needs, particularly in writing and speaking at the college level (Anton, 2009). Alderson and Huhta (2011) and Anton (2009) noted that theoretical concepts in foreign language formative assessment

are still at an early stage because most assumptions regarding reading learning progressions are based on the understanding of cognitive performance in a primary language. Therefore, conclusions arising from formative assessment instruments may be limited unless they take into account the differences between primary language and secondary language learning and establish specific formative assessment devices for learning a foreign language. Grigorenko (2009) noted that even though alternative forms of formative assessment are relatively young, theoretical literature on formative assessment to address diverse language learning needs has reached a mature level of development. However, the staggering volume of literature available for traditional forms of assessment tends to skew the perception of literature on alternative forms of formative assessments for diverse learners, thus leading to an incorrect conclusion that literature on this subject is emergent instead of reaching a mature theoretical ground (Grigorenko, 2009; Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002).

The different views regarding the stage of development of foreign language formative assessment theories may be due to the dissenting perspectives regarding the cognitive differences for learning a second language between children and adults (DeKeyser & Sokalski, 1996). Although some researchers have noted that adults have a tendency to replace grammatical skills with problem-solving skills (Bley-Vroman, 1988), others have indicated that the cognitive strategies used by children and adults are similar and problem-solving strategies are not a determining factor in learning a foreign language (Krashen, 1982, 1985, 1994). In the foreign language field, language researchers and instructors have debated the developmental and cognitive differences between adult and children learners, but also recognize the differences in skills, interests, and learning styles

that require individualized strategies for instruction (Ehrman & Leaver, 2003; Ehrman, Leaver, & Skekhtman, 2002; Ragini, 2016; Sternberg, Grigorenko, & Zhang, 2008). Consequently, some foreign language formative assessment experts and developers tend to be cautious not only of the factors that contribute to the process of learning a foreign language as a child and as an adult, but also of the possible limitations of selecting a one-size-fits-all formative assessment to measure learners with a variety of needs, styles, and skill-set differences (Bachman & Clark, 1987; Sternberg et al., 2008). Tied to these complexities is the measurement of listening skills, which requires nontraditional models of assessment as well as an acute understanding of the unique cognitive characteristics involved in the listening process, particularly for second language learners. Factors such as the speed or rate of listening stimuli, different types of foreign accents, hesitations, length of the recording, stimuli with inferred meaning, and cognitive skills involved in short- and long-term memory when listening to recording stimuli will affect the effectiveness of the assessment construct if they are not taken into careful consideration during the test design and development process (Buck, 2011).

Since the late 1980s, psychometricians have acknowledged the ramifications of applying traditional models of testing into second language acquisition proficiency assessments, particularly because normed studies in the past considered students with full linguistic abilities as part of their norming studies without taking into consideration the diverse levels of second language proficiency among students (Bachman & Clark, 1987). Bachman and Clark (1987), and later Bachman and Palmer (2010), formulated a framework for addressing the factors that affect language proficiency testing that includes (a) communicative language proficiency, which requires not only language abilities but

also the ability to apply these skills through strategic and psychophysiological abilities; (b) language competence, which considers the application of organizational and practical abilities for the use of grammatical and rhetorical conventions; (c) strategic competence, which requires the ability to identify relevant information to produce the highest possible meaning; and (d) psychophysiological skills, which require an ability to discern which of the abilities described above is more effectively executed into listening, speaking, reading, and writing. Because of these factors, Bachman, Clark, and Palmer suggested the clear discernment of the selection of a second language formative assessment along with the corroboration of data that validate the need for its administration, making the process of corroborating a test selection part of the validation process and a central component of their framework (Bachman 2013; Bachman & Clark, 1987; Bachman & Palmer, 2010).

Regarding the specific test design characteristics of second language formative assessments, Bachman and Clark (1987) and Bachman and Palmer (2010) seemed to prefer sizable performance-based assessments that lent themselves to a series of authentic tasks that are conducive to the authentic measurement of a learner's language abilities. These assessments should have the following characteristics: (a) evidence of the measurement of the communicative language proficiency, language competence, strategic competence, and self-monitoring skills; (b) the use of authentic materials and real-life scenarios; (c) evidence not only of test validity but also of a methodology that demonstrated the absence of negative factors during the test-taking process; (d) a sizable number of studies that determined the validity of the test, including correlation and validation studies; and (e) the practical use of the test, including its administration, scoring, and reporting information (Bachman 2013; Bachman & Clark, 1987; Bachman

& Palmer, 2010). While this model has many benefits because of its substantial performance-task-based characteristics and the extent of its validation procedures, critics have noted the considerable time required for its implementation. As a result, these assessments tend to have fewer sampling characteristics that limit the generalization of these types of instruments: “The art of assessment development is to balance the need for adequate sampling of the domain and consistency in scores across replications of the assessment with the need for tasks that are as authentic as possible” (Kane, 2011, p. 584).

Performance-based testing based on real-life tasks has been one of the preferred ways to measure formative foreign language testing since the late 1980s, but dynamic assessment has been of theoretical interest and practical implementation since the early 2000s (Lantolf & Poehner, 2004; Lantolf & Thorne, 2006). Elliott noted dynamic assessment is an “umbrella term used to describe a heterogeneous range of approaches” (as cited in Grigorenko, 2009, p. 16) implemented to address the dissimilarities in cultural and cognitive development environments (e.g., second language learners, new immigrants, underprivileged groups) to synthesize instruction into assessment (Grigorenko, 2009). Thus, there seems to be a natural synergy to use dynamic and diagnostic forms of assessment to measure second language proficiency. Traditional premises, with their traditional approaches toward continuous learning processes, did not seem to meet the needs of dissimilar classroom environments in second language acquisition classrooms. In contrast, dynamic testing considers not only current student knowledge and abilities, but also future learning indicators that take into account the possibility of peaks and valleys in learning, which suggests that learning is nonlinear and requires scaffolded testing to identify specific areas where skills and abilities have

reached a ceiling at an individual level (Grigorenko, 2009; Sternberg & Grigorenko, 2001, 2002). The concept of dynamic testing came from Vygotsky and Feurestein, who were trying to identify strategies to assess students in disadvantaged learning environments such as orphans and immigrants to place these students in mainstream classrooms (Grigorenko, 2009). Vygotsky theorized through the ZPD model that learners with diverse cognitive skills and needs could profit from early intervention, thus yielding a more accurate description of what each learner needs to know at his or her specific proximal level of learning (Minick, 1987, p. 120; Vygotsky, 1963, 1998).

The inordinate placement of immigrants and ethnic minorities in special education classes rather than second language acquisition courses led to a theoretical concept in formative assessment known as responsiveness, or response to intervention (RTI). RTI aided in identifying students with slow reading abilities through developing early remediation devices to discern whether there were learning differences based on learning ability or achievement, which led to developing proactive strategies for learning before student failure occurred (Morris et al., 1998; Stanovich & Siegel, 1994; Torgesen, Morgan, & Davis, 1992). According to Grigorenko (2009), although dynamic testing is a process that results in the assimilation of instruction into assessment, responsiveness or RTI is the process that results in the assimilation of assessment into instruction. Therefore, both processes are an essential component of diagnostic testing, instruction, and student learning.

Although dynamic testing appears to be one of the formative assessment modalities to assess foreign language learning, large-scale English second language proficiency tests such as the Test of English as a Foreign Language and the Michigan

English Language Assessment Battery (MELAB) use diagnostic assessment procedures to identify areas of strength and areas of growth in a set of learning abilities such as knowledge, skill, and learning strategies for which diagnostic models such as the fusion model are implemented (Kim, 2015). Determining the characteristics of these learning abilities according to empirical and theoretical indicators in a specific second language would help instructors and administrators identify specific treatment strategies for individual learners based on identified areas of strength and growth (Kim, 2015; Lee & Sawaki, 2009). When devising cognitive diagnostic assessments, these learning abilities are commonly denominated as cognitive attributes or cognitive procedures and comprise the “[cognitive] procedures, skills, or knowledge a student must possess in order to successfully complete the target task” (Birenbaum, Kelly, & Tatsuoka, 1993, p. 443).

Sternberg et al. (2008) recommend that instructors ensure students master analytical skills, including strategies for learning how to think. The mastery of analytical cognitive strategies ensures students can succeed when taking an assessment, regardless of the unique characteristics of the assessment construct.

Regarding the anatomy of formative assessments, Alderson and Huhta (2011) described the following attributes as representative of second language acquisition formative tests of a diagnostic nature:

1. provide higher level of specificity in the areas of growth;
2. provide comprehensive assessment results through individual performance level descriptors;
3. provide immediate feedback;
4. lead to positive testing conditions due to their low-stakes nature;

5. based on relevant instructional content and a well-founded language development theory;
6. based on research on second language learning or on a well-established linguistic theory;
7. include parceled tasks that are self-contained rather than thematic tasks that unite a subject matter across sections;
8. focus on the measurement of language and not necessarily on small language skills;
9. measure language skills at all levels except complex skills at the upper end of Bloom's taxonomy for higher order abilities due to the fact that these skills tend to combine several tasks;
10. use technologically based tools;
11. include information with strategies for areas of improvement; and
12. provide a high level of specificity in their diagnostic reports that lead to applied instruction.

Foreign language researchers have acknowledged that second language acquisition does not usually require oral mastery prior to reading mastery. For second language learners, low-level connections such as word recognition issues and syntax issues may appear more often during the completion of high-level tasks compared to first-language learners (Brunfaut, 2008; Shiotsu & Weir, 2007). In addition, in the case of adult students, the type of higher order thinking skills connections required in foreign language learning needs to be taken into account. Adult students may have differential prior knowledge based on their backgrounds and educational level (Alderson & Huhta,

2011). The analysis of second language formative assessments offers potential for researchers, including what makes a reading test item gradually more challenging from the perspective of what a second language learner knows and is able to do (Alderson & Huta, 2011).

Validity issues with formative assessments. Because of the intrinsic difference in the design and expected outcomes of formative and summative assessments, the implementation of summative quantitative strategies to measure student abilities does not seem to be appropriate for a formative test. Learning progressions measuring prior knowledge and new knowledge in any given domain (Shavelson & Kurpius, 2012) require qualitative descriptions of student knowledge rather than quantitative studies. Narrative descriptions of a student's cognitive patterns at the individual level on a set of specific domains are more relevant for a formative assessment to identify how to master the next learning progression domains (Trumbull & Lash, 2013). In addition, because of the scant research on learning progressions, and because of the intrinsic nature of instruction as part of the formative assessment process, the validation of a formative assessment instrument may be limited if there is lack of evidence of actual application of formative assessment results into informed instruction of specific areas of growth (Trumbull & Lash, 2013).

While considering the validation of a formative assessment construct defined by its process of interaction between the learner, the instructor, and the formative assessment instrument, it is also important to take into consideration the technological sophistication of formative diagnostic instruments that requires automated scoring. These instruments are most commonly used by English-as-a-second-language assessment agencies to

measure writing and speaking abilities and require the use of linguistic and statistical formulas to deconstruct the test taker's responses and translate these into meaningful scores (Chapelle & Chung, 2010). These linguistic and statistical formulae identify specific words, prepositional phrases, number of words, and word variations. Depending on their purpose, they could require multiple regression features to emulate the type of scoring they could have received from a specialized evaluator (Page, 2003; Valenti, Nitko, & Cucchiarelli, 2003). Although researchers have found that these types of assessments have shown to have less chance for human error (Keith, 2003), the threats to validity on an automated formative assessment are commonly the result of test takers who are able to understand and outsmart the automated scoring logic of the assessment instrument (Chapelle & Chung, 2010).

Theoretical criteria based on practices for secondary language testing are necessary for foreign language testing validity issues (Alderson & Huhta, 2011; Buck, 2011). Primary language assessments conceive reading comprehension as the result of low-level and high-level cognitive connections progressing in a continuum. However, for second language learners, low-level connections may occur in high-level tasks, as in the case of word recognition and syntax (Harding et al., 2015). Formulating an assessment for primary language learners assumes that students have already mastered the knowledge of certain words and syntaxes, while foreign language students may not have mastered these skills yet and may still be required to complete high-level tasks. In this context, the wealth of foreign language vocabulary knowledge (Brunfaut, 2008), as well as the mastery of syntax (Shiotsu & Weir, 2007), may need to be considered when validating a second language formative assessment instrument, particularly for reading.

In addition, cognitive learning strategies of adult second-language learners along with their wealth of prior knowledge from different educational backgrounds have an effect on their learning process (Galbraith, 2004). Phoener (2005) noted that different social and life experiences also have different cognitive learning ramifications in adults. These factors may need to be considered when selecting second-language formative assessment instruments.

Also important is recognizing the cognitive differences and distance between the grammatical concepts and language alphabet that the second language can have compared to the written and grammatical rules of the student's first language (Alderson & Huhta, 2011). Another factor of formative assessment design validity is the effect strong or limited literacy abilities in a primary language may have on the reading performance of students learning a second language (Alderson, 1984; Sparks & Ganschow, 1993; Sparks et al., 2006, 2008). Students may have demonstrated a ceiling level in foreign language production due to their first-language background knowledge, as well as from their high literacy in their mother tongue. As a result, formative assessment devices in foreign language acquisition may need to take into account several factors, including the cognitive processes related to the age of the learner, educational level, background knowledge, and the alphabetical and syntactical distance of the second language learned compared to the first language of the learner (Alderson & Huhta, 2011).

Durán (2011) suggested that traditional applications of validity and reliability measures may not be feasible with formative assessments. However, according to Durán, the application of formative strategies contributes to their validation because instructors have the option of measuring domains frequently. In this context, the possibility of

building a body of performance results from formative assessments administered on an ongoing basis increases the level of confidence in the type of assessment conclusions and strength of the formative assessment instrument (Shavelson et al., 2007).

In 2014, the Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999, 2014) introduced a revision to the standards to include recommendations that include criteria for diagnostic assessment strategies. In this context, the 2014 standards provide guidelines that increase the validity and reliability of formative assessments to ensure their appropriate application in educational programs and include considerations for innovative items formats, as well as other important issues that include automated scoring and general computer-based assessment considerations (Plake & Wise, 2014).

Validity of formative assessments through linking studies. Researchers have found a direct relationship between the use of formative assessments and student achievement (Black & Wiliam, 1998; Bower, 2005). Sly (1990) suggested that two factors that show a positive result in the outcome of summative assessments include the student's acquaintance with formative assessments and the specific observations resulting from the formative test that help learners to understand their specific cognitive errors.

S. T. Miller (2009) researched formative assessments in the form of computer-based assessments and found several studies that showed formative assessments having a positive impact on summative assessment instrument results (Henly & Reid, 2001; Pinckey, Mealy, Thomas, & MacWilliams, 2001; Pitt & Gunn, 2003). The positive impact of formative assessments seems to be demonstrated even on students who

underperform academically (Sambell, Sambell, & Sexton, 1999; Charman and Elmes, 1998).

However, it is sometimes unclear whether the positive effect of a formative assessment may be the result of an improvement in test taking skills (Sambell et al., 1999), higher motivation, the academic preparation of students taking the formative tests (Henly & Reid, 2001), or better testing conditions in the form of additional testing time (Pitt & Gunn, 2003). Additionally, the benefit of test taking practice and understanding of the classroom materials and goals may be also a factor in positive summative assessment results (Sambell et al., 1999). Formative results might not show positive outcomes on summative assessment instruments in some cases (Henly, 2003). In some of those cases, students seemed to have experienced boredom due to the low-stakes nature of the formative assessment or have used computerized formative assessments to retake the test to review their test results. As a result, the assessment instrument was used to provide the answers to their test, rather than as an evaluation of their true learned skills (Henly, 2003). Other studies have shown that in the case of independent learners, formative assessments have been able to help devise appropriate strategies for growth through learning progressions as long as learners were able to recognize the appropriate uses of formative assessments and their difference with summative assessment tools (Organisation for Economic Co-operation and Development, 2005). Because the demand for a suite of assessments designed for different purposes has increased, there is a need to ensure all assessment instruments by an organization align in goals, standards, and educational philosophy, above their specific differences in test design and conceptualization. There is also a need to ensure the success of these instruments by

providing the assessment information expected through their respective assessment designs (Herman, 2010; Pellegrino, 2006). Therefore, reliable strategies for aligning formative and summative assessments are necessary to verify the appropriateness of formative instruments in assisting in the learning process (Black & Wiliam, 2003; Lam, 2013).

The necessity to link two assessments may be the result of having to identify the outcome of one assessment as observed when identifying its correlation to the results in another assessment instrument (Deming, 1980). Mislevy (1992) suggested that the successful linking of two assessment instruments depends on the quality of the strategies used, as well as the commonality of assessment construct goals of these two assessments. In this context, a correlation of two assessment instruments that share the same content rationale, standard framework, and student population may have a higher chance for producing linking results that show meaningful correlation data than assessment instruments based on a disparate student population, content rationale, or standard criteria. Mislevy (1992) noted,

Two similar scores convey similar meanings to the extent that they summarize performances on suitably similar tasks, in suitably similar ways, for suitably similar students. We must be alert to patterns in individual students' data that cast doubt on using their test scores to compare them to other students, and we must be reluctant to infer educational implications without examining qualitatively different kinds of evidence. (p. 16)

Having common assessment design characteristics could then help researchers to identify patterns in variables that otherwise might not be easily identified. Test theories

such as equating and calibration could help researchers to measure and link assessment instruments, as long as they are able to corroborate comparability in assessment constructs.

History of Assessment in the United States.

Assessment instrument measurements in the United States started over 150 years ago. Strategies to give accreditation to higher education institutions surfaced in 1900 (Urciuoli, 2005). Since their inception, educational assessments have been guided by policies, cognitive theories, and technological capabilities (Pellegrino, 2004).

Accreditation institutions emerged in the United States in 1913. Instead of government agencies, private organizations provided the accreditation (Alade & Buzzetto-More, 2006; Urciuoli, 2005), which explains the number of private assessment development organizations in the United States, some of which were founded over 80 years ago, such as the California Testing Bureau (CTB), now part of Data Recognition Corporation.

According to Pellegrino (2004), three areas that defined the assessment design since 1957 include the theories of cognition, the curriculum requirements, and the sociopolitical pressures in education. Pellegrino considered 1957 a meaningful year for assessment development, because Cronbach proposed to the APA an innovative strategy that linked two areas of study: scientific psychology and correlational psychology. By doing this, Cronbach was able to unify theories on learning and instruction with the tradition of testing individual differences in cognitive capabilities (Cronbach, 1957; Pellegrino, 2004). With Cronbach's contributions, psychometric strategies and cognition strategies came together to validate and support curriculum and education. As cognitive theories evolved, the emphasis changed from intelligence and aptitude tests to the study

of instructional and learning domains. In 1964, governmental efforts to improve the quality of education through the Economic Opportunity Act contributed to the creation of the Head Start and to the Elementary and Secondary Education Act. These programs provide funding to local school districts through the federal funding known as Title 1 (Guskey, 2005). These governmental efforts required validation. From 1957 to 1990, educational assessments expanded their areas of study and used psychometric techniques to assess progress in academic instruction (Pellegrino, 2004). In the 1980s, statewide summative assessments became an essential tool to measure educational progress and to make school districts accountable (Klinger, DeLuca, & Miller, 2008). To this end, Standards for Educational and Psychological Testing were developed (American Educational Research Association et al., 1999, 2014) to establish standardized criteria to evaluate the quality of assessments and testing methods and provide guidelines for test development for assessment development organizations (Plake & Wise, 2014).

In the 1990s, theories of cognition inspired efforts to reconcile issues with curriculum and assessment. These theories were based on findings regarding specific stages of learning and their correlation with different types of skills, as well as differences in acquired knowledge and its corresponding variations in performance. These theories inspired the development of new assessments based on variable outcomes in stages of learning, skills, and performance (Pellegrino, 2014). These theories contributed to the redefinition of assessment to include traditional as well as emerging types of assessment instruments. Mislevy (1992), for example, broadened the term educational assessment to include not only standardized evaluation instruments but also other instruments such as dissertations and essays or components that may require a

specific evaluation over a period of time, such as portfolios. Assessments encompassed instruments that required a body of work on the part of the student as well as individual evaluations on the part of the instructor. Mislevy (1992) suggested that the goal intended for a given assessment determines the design and outcome of the evaluation as well as the type of validation procedures. Therefore, the elements that are essential to validate a given assessment may be irrelevant or unnecessary for another assessment instrument. Mislevy (1992) noted,

When the focus is on the individual, enough evidence must be gathered on each student to support inferences about him or her specifically. On the other hand, a bit of information about each student in a sample—too little to say much about any of them as an individual—can suffice in the aggregate to monitor the level of performance in a school or a state. (p. 4)

One of the most significant shifts in the 1990s was the attempt to use assessments as a channel to improve and affect change in learning and instruction. This approach assumed that, because of the faulty tendency to teach to the test, changes toward the creation of more complex assessments and test designs aimed at higher order thinking skills could promote changes in the instructional outcome. By providing a higher level of granularity to the expected outcome of a test, as in performance-based assessments, the expectation was that instruction will be driven to a higher level of thinking skills. This focus inspired many assessment organizations and institutions to identify and assess significant aspects of the learning process so that instructors can focus their attention on the instruction that resulted from the assessment outcome (Pellegrino, 2004).

Another significant contribution of the 1990s is the Improving America's Schools Act of 1994, Goals 2000, and America 2000 standards movement, which contributed to the efforts to ensure school and state accountability through measureable academic standards, as well as with assessment tools to demonstrate that those academic standards have been met (Cromey & Hanson, 2000; Schultz, 2012). By 2002, The No Child Left Behind (NCLB) initiative ensured a nationwide accountability system that resulted in the increase of assessments for high-stakes purposes (Tucker & Coddling, 2002). Schultz (2012) referred to this era and its emphasis on high-stakes summative assessments as the era of accountability. As a result, and at the individual level, instructors were seeking assessment instruments that were more organic to the learning and instruction process and satisfied multiple needs from finding information about the ongoing learning progress to identifying instructional strategies at the individual, classroom, and state level (Darling-Hammond & Pecheone, 2010; Pellegrino, 2006), but still had the capability to provide meaningful information consistent with the approved standards and educational objectives at the classroom and state level (Herman, 2010; Pellegrino, 2006).

An increased awareness of the value of multiple measures of assessment led to an environment for developing what were known as alternative classroom assessments, which included self-assessments, peer assessments, classroom observations, and student portfolios and interviews (Butler & Lee, 2010). New assessments such as formative assessments led learners to judge their own learning progress and to identify the best way to arrive at the place where they needed to achieve their educational goals based on the established curriculum criteria (Andrade et al., 2010; Assessment Reform Group, 2007; Radford, 2014, Taghizadeh et al., 2014). Alternative assessments were defined as those

that provided fast administration, high autonomy, and involvement on the part of the student in the assessment process and an increase in motivation (Blanche & Merino, 1989; J. D. Brown & Hudson, 1998; Taghizadeh et al., 2014). However, the complex tasks required by these types of assessments raised some concerns from psychometricians because of the absence of a construct design that could help to explain and measure the assessment outcomes (Messick, 1994; Pellegrino, 2004).

In 2004, Pellegrino identified two fields that had contributed to the types of assessments currently available: (a) theory and research and (b) educational practice. Pellegrino's goal was to recognize the differences in construct of two types of assessments: classroom-based assessments or formative assessments and large-scale assessments or summative tests. He suggested that formative, classroom-based assessments are influenced by cognition theory and research, and summative, large-scale assessments are influenced by psychometric constructs.

Almost every state had its own state standards and graduation criteria in the early 2000s with NCLB. However, at the national level, these standards did not have the same criteria. In this context, the Common Core Standards were developed in 2009 by looking at the best state standards available with the goal of having standards developed for the country by state departments of education, instructors, and experts representing all states. These core standards were classified by standards for college and career readiness and K-12 standards (Common Core State Standards Initiative, 2016). States use the Common Core Standards on a voluntary basis to ensure students have the tools necessary to succeed in college, including high skills needed in the workforce. New assessments have

been developed to align to Common Core Standards to make decisions about the future not only of students but also for instructors and school districts (Tucker, 2010).

By mid-2000, a new type of assessment was widely introduced: computer-based testing (Hogan, 2013). Computer-based assessments owe their quality and innovative edge to computer-based technologies, which contributed to the development of profuse item banks. Item banks became a new trend in assessment development, particularly in formative assessment. By 2005, the Educational Testing Service (ETS) announced the ETS Formative Assessment Item Bank, which at the time had more than 11,000 assessment items, which increased to more than 64,000 items aligned to the standards of all 50 states in the United States (ETS, 2011; Internet@Schools, 2005). With the development of copious item banks, the relevancy of formative assessments in the classroom was unequivocal, and formative assessment became an intrinsic part of the educational process, providing instructors the flexibility to customize assessments based on individual student needs as well as the adjustment of assessments to the specific standard requirements of different states (Olson, 2005). Computer-based technologies and robust item banks contributed to a new breed of formative and summative assessments known as computer adaptive tests (CATs) that adjust to the level of performance of the test taker (Linacre, 2000). According to Linacre (2000), CATs are able to identify a student's abilities through a series of algorithms, thus producing a specific test that becomes easier or more difficult according to the success in answering specific test questions provided. CATs were originally used with caution in summative assessments due to concerns that the grade-range clumps of CATs could lead to inaccurate grade-level classification for grade-specific testing (Horn, 2003). Kingsbury

and Hauser (2004) suggested that CATs could be effective in high-stakes testing and could support initiatives such as the NCLB by providing accurate student reporting information as well as a reduced level of student frustration, which may contribute to the increase in the accuracy of the CAT assessment results for accountability purposes. Yatzkanic (2015) asserted that computer adaptive assessments include some challenges with regard to test fairness due to group differences in test results. According to Yatzkanic, computer adaptive assessments such as the STAR Reading and Classroom tools are at an early stage and more research is necessary to demonstrate appropriate student skill interpretations.

The Common Core Standards are the first initiative at the national level by 42 states, the District of Columbia, and four territories to introduce standards nationwide to be incorporated by all states at a voluntary level (Standards in Your State, 2016). The goal with Common Core was for U.S. students to have the skills required to succeed in a global economy (Schultz, 2012). According to the National Assessment Governing Board (2012), this initiative started in 2010, when Common Core granted two testing consortia, the Partnering for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC), to develop assessments for English language arts and math to be fully operational by 2014-2015. PARCC is a consortium of 24 states working in partnership to develop Common Core assessments for Grades K-12 that will ensure students have the appropriate foundation for work and college and allow instructors to have enough information to guide instruction (Nellhaus, 2012). SBAC includes 22 states, with five states in the role of advisory members (Willhoft, 2012). While PARCC received a grant to develop Common Core assessments for K-12, SBAC

also received a grant to develop a Common Core assessment in 2010 for the development of computer adaptive tests, particularly for low and high performers, students with disabilities, and English language learning students.

Pellegrino (2012) and the Committee on Developing Assessments of Science Proficiency in K-12 for the National Research Council of the National Academies (2014) contributed to this vision of the future of assessments. The National Research Council of the National Academies recognized that the Next Generation Science Standards (NGSS) requires instructors to change the way science is taught considerably. As a result, curriculum, instruction, and assessment must be interconnected in every aspect of science education. What is meaningful about the challenges found in the NGSS standards and the contribution of SBAC and PARCC is that the recommendations for new assessments include multiple assessments or assessment tasks to identify students' mastery. Any specific assessment task could assess more than one standard, described here as a performance expectation. Recommendations include the development of test questions that are unique in that they are linked or related to each other. In the case of the Next Generation Science, recommendations in test design included (a) having multiple components that reflect the interconnectedness of different disciplines within science; (b) addressing the natural learning continuum of students, (c) providing information about the specific beginning and ending points of particular learning units; (d) having a system that allowed for the interpretation of the student responses at their different levels of performance and not of less importance; and (e) providing information to assist educators in the next step of instruction at an individual level. Pellegrino, the National Research Council of the National Academies, PARCC, and SBAC were envisioning a sophisticated

version of a new generation of formative diagnostic assessments that emerged a decade ago. In its basic definition, diagnostic assessments relate to the set of strategies devised for identifying a student's strengths and weaknesses (Alderson, 2005).

U.S. Assessments of Foreign Language at DLIFLC

Building foreign language expertise demanded that the DoD provide foreign language training, monetary incentives, and reliable standardized testing procedures to ensure the appropriate foreign language qualifications of military staff. The Foreign Language Proficiency Bonus (FLPB) implemented since the 1980s has helped the DoD shape the linguistic expertise needed among its own ranks while distributing incentives toward specific languages that serve the overall DoD mission (U.S. Department of the Army, 2016; DoD, 2013). To obtain a FLPB, it is necessary to submit to an annual assessment of reading and listening abilities through the DLPT5. Monthly bonus incentives range from \$100 to \$500, depending on the service member's score on the DLPT5 from Levels 1 to 4 on the ILR scale. Additional factors that affect the FLPB incentive rate include the category of the language. Category I and II Languages may be paid a lower rate than Category III and IV languages¹ (U.S. Department of the Army, 2016).

Placement test: DLAB. Military students take the DLAB at their accessing stations. The results on the DLAB, combined with the military branch language mission requirements, contribute in part to the foreign language program taken at DLIFLC. Low

¹ The language categories were established by the Foreign Service Institute (FSI) in 1973. The languages currently taught at DLIFLC include Category I & II: French, Spanish, and Indonesian; Category III: Hebrew, Persian Farsi, Russian, Tagalog and Urdu. Category IV: Modern Standard Arabic, Arabic Egyptian, Arabic Iraqi, Arabic Levantine, Arabic Sudanese, Chinese Mandarin, Japanese, Korea, and Pashto.

or high scores on the DLAB influence the language of assignment; those with higher scores are typically assigned a more difficult (Category III or IV) language (Anderson, 1997; Wong, 2004). The DLAB was developed to measure the aptitude of students in learning a foreign language (CASL, 2017; Peterson & Al-Haik, 1976). Multiple regression studies and other validation studies performed on DLAB for Categories I and II languages have indicated that the DLAB provides score information that could help guide the selection of a category of language and has the potential of predicting the success of learning a language at DLIFLC (Anderson, 1997; Peterson & Al-Haik, 1976; Wong, 2004). Although the validation study did not include sample data to measure the success for Category III and IV languages, a cut score of 100 along with the needs of the specific military service units has been helpful in classifying students at more difficult languages while having relatively low attrition (U.S. Department of the Army, 1994a, 1994b). On September 21, 2015, DLIFLC announced the collection of data for a new aptitude test, the DLAB 2, developed in collaboration with the Maryland Center for the Advanced Study of Language, which is expected to replace the current DLAB, although the operational date has not been specified (CASL, 2017; DLIFLC Midterm Report, 2015).

Summative test: DLPT5. Students study a full language program comprised of 6-7 hours of classroom instruction per day plus independent time for homework assignments. At the end of the 36- to 64-week program of language instruction, students take the DLPT5 to determine their proficiency levels. As of 2017, graduation criteria for DLI were raised to the minimum achievement of 2+ in listening, 2+ in reading, and 2 in speaking on the DLPT5 and Oral Proficiency Interviews (DLIFLC, 2017). The DLPT5 is

a high-stakes summative test developed by DLIFLC. The DLPT5 is the newest version. The DLPT5 is a computer-based assessment instrument that measures the foreign language proficiency in reading and listening of English native speakers (DLIFLC, 2015b). The ILR scale was used in a more systematic way with the development of the DLPT5 to ensure greater validity and calibration methods, which included the configuration of standard-setting panels for setting DLPT5 cut scores. As part of the DLPT5 validation, new processes were introduced with ILR experts from different languages during the item development process. Each passage and item went through an independent review by the Proficiency Standards division to ensure a consistent interpretation of the ILR performance-level descriptors across languages during the test development phase. After the test development was completed and verified by the Proficiency Standards Division, a pre-standard-setting discussion with ILR experts from different languages was introduced to the validation process to interpret the ILR performance-level descriptors in the context of DLPT5 measurement. The pre-standard-setting panel was an important strategy set to ensure the ILR was used in a more systematic way. Lastly, the standard-setting phase, as a crucial step in the validation process, applied standardized procedures that used the ILR performance-level descriptor statements in a clearly organized and categorized process across languages, resulting in explicit standard setting that ensured greater validity for the different DLPT5 language instruments (M. Hoffman, personal communication, June 28, 2017). The ILR Scale determines the scores for the DLPT5. An average of the reading and listening score is created to provide an ILR score. The scores range from 0+ to -4. The DLPT5 is available in two difficulty ranges: the Lower-Range test (for levels 0+ to -3) and the Upper-Range

test for students who received an ILR score of 3 in the Lower-Range test (DLIFLC, 2015b). The DLPT5 has two assessment instruments: reading and listening. At DLIFLC, speaking is assessed on a one-on-one basis with certified oral proficiency interview testers. For the Lower-Range test, each instrument includes approximately 60 test items, including 30 stimuli for reading and 40 stimuli for listening. Passages and listening stimuli range in length but do not exceed 500 words or over 2.5 minutes per listening stimulus. Stimuli contain excerpts of authentic target language reading and listening materials, which may include newspaper articles, radio or television advertisements or broadcasts, or website information with content representative of the culture and language measured and relevant to the military student. Each stimulus has at least four multiple-choice items for reading and two multiple-choice items for listening. Test takers have the opportunity to listen to a given stimulus twice. This is a timed test completed in 3 hours for each content area, with a 15-minute break in between each content area (DLIFLC, 2015b). For the Upper-Range test, each reading and listening assessment instrument includes about 36 test items. There are about 14 stimuli for reading and 14 for listening. Each reading stimulus contains five multiple-choice items, and each listening stimulus contains three multiple-choice items. Test takers have the opportunity to listen to stimuli twice. While the assessment stimuli are delivered in the target language, the assessment questions are administered in English (DLIFLC, 2015b).

Formative test: ODA. The ODA is a web-based, semiadaptive diagnostic assessment instrument that measures the foreign language skills of learners for Levels 1 to 3 on the ILR Scale (DLIFLC, 2015d). The ODA helps to identify the specific areas of strength and the areas of growth that would allow a foreign language learner to grow to

the next level of language proficiency (DLIFLC, 2015d). The federal government uses the ODA formative assessment for foreign language training and maintenance curriculum (U.S. Department of the Army, 2015) to identify existent language proficiency at the time of the assessment, as well as future language skills required by providing a report of specific linguistic areas to work on to achieve the next proficiency level (Clark et al., 2014). The ILR Scale determines scores for the ODA. A separate score is provided for listening and reading. The scores range from 1 to 3.

The first two ODA assessments were available in 2007 for Standard Arabic and Korean for the reading content area. Over time, additional languages were developed, along with listening diagnostic components. ODA delivers formative diagnostic assessments for 18 languages (DLIFLC, 2015d). The use of the ODA has increased over time. In 2015, over 35,000 sessions of the ODA were administered for all languages available. Figure 3 shows ODA sessions by language per year from 2008 to 2015.

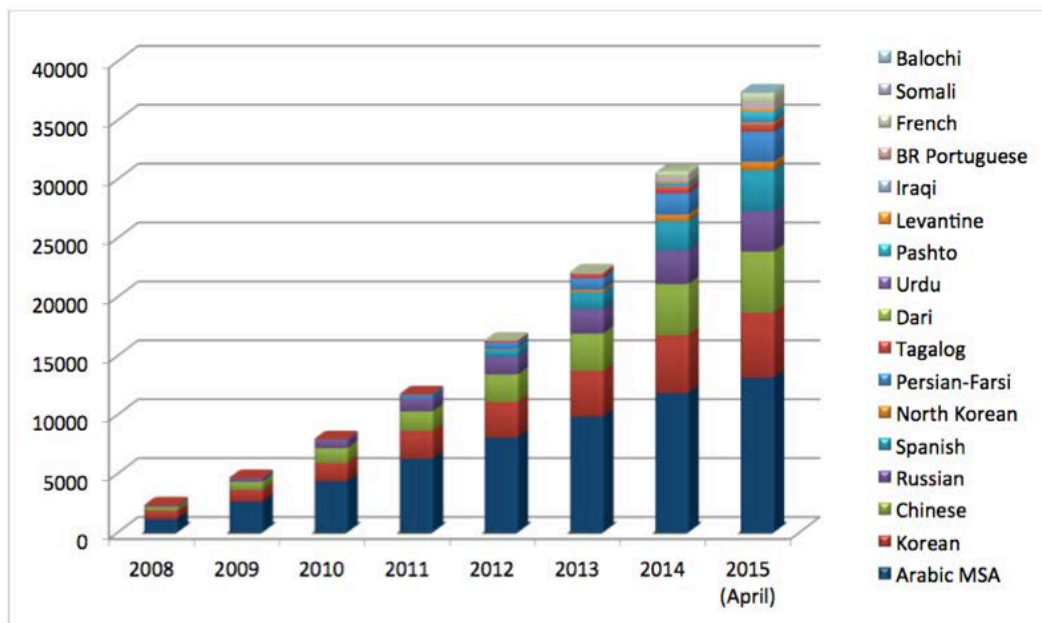


Figure 3. ODA sessions by language by year. From *Online Diagnostic Assessment Team Program Review* (p. 78), Defense Language Institute Foreign Language Center, 2015d, Monterey, CA: Author. DLIFLC.

Originally developed to address the language maintenance and enhancement needs of military staff who had already graduated from the Monterey Basic Course program (nonresident linguists), the ODA has grown to support the formative diagnostic requirements of DLIFLC resident students as well as nonresident students at the basic, intermediate, and advanced levels. It provides an individualized evaluation at a high level of granularity for two tests, a listening and reading test. Both tests assess the student comprehension of either an audio or a reading stimulus and are followed by vocabulary, sentence structure, and text structure (DLIFLC, 2015d).

One of the critical requirements of instructors and managers of linguists is to identify individualized remedial procedures for students with a wide variety of linguistic needs, even though they might have comparable proficiency test scores (U.S. Department of the Army, 2015). Specific strengths and weaknesses can be identified through the ODA to customize instruction to meet individual learning requirements. According to the U.S. Department of the Army (2015),

ODA (1) offers language assessment that adapts to the learner's performance; (2) determines and verifies floor and ceiling levels of proficiency; (3) collects diagnostic data; (4) generates diagnostic profiles and; (5) provides the learner with individualized feedback. Sampling of learner abilities is systematic across a variety of levels, topics, tasks, and specific linguistic features. (para. 2)

The ODA contains reading stimuli, audio, and multiple-choice and open-ended questions called constructed response type questions (CRTs). The CRTs require an English response. The ODA is semiadaptive, so the multiple-choice and CRT items are automatically scored through an algorithm. By collecting diagnostic information from the

learner's responses, the algorithm generates a new set of items and passages for the next performance level, whether it is a higher or a lower level. The system continues to adjust the level of the test taker's performance to a higher or a lower testlet until the highest performance ceiling is identified. To ensure accuracy of results, test takers receive two sets of items at the ceiling of their performance level. Once the assessment is completed, an ODA diagnostic profile is generated (DLIFLC, 2011, 2015d). Figure 4 shows a visual representation of the ODA computer adaptive features.

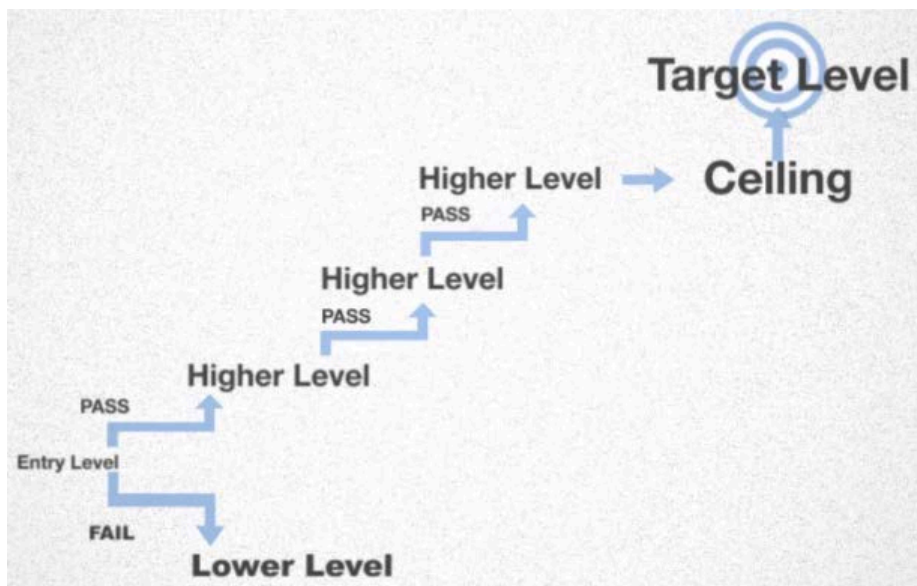


Figure 4. Computer adaptive features of the ODA. From *Online Diagnostic Assessment*, by Defense Language Institute Foreign Language Center, 2011, retrieved from <https://vimeo.com/16633421>. Figure is in the public domain.

The automated features of the reading and listening ODA produce an ODA diagnostic profile immediately upon completion of the test. The ODA diagnostic profile identifies the individual strengths and areas of growth of a student based on the ILR criteria for Levels 1 to 3. The diagnostic profile contains two evaluations. One evaluation describes the current level the student was able to achieve at the time the ODA was taken. The second evaluation describes the target level that the student failed to achieve. The

individualized student feedback available on the ODA diagnostic profile includes an estimate of the ILR level per content area, a graphic showing the student's performance at a glance, and two reports: a descriptive report with the successfully performed skills referred to as current level and a descriptive report with the skills to achieve performance growth referred as target level (DLIFLC, 2015d). The two reports are similar regarding the organization and feedback categories, but they differ on the breakdown of specific information given based on either the current skills or target skills (DLIFLC, 2014).

The two ODA evaluations provide score information based on the ILR Scale for Levels 1 to 3, along with a description of current skills and targeted skills that may require additional instruction based on individualized score results (DLIFLC, 2011, 2015d). Because of the level of granularity of these two ODA diagnostic profile reports, which includes a subject area breakdown with specific information on what the test taker needs to work on the most, the ODA could be used by independent learners as well by instructional programs (DLIFLC, 2011, 2015d). Figure 5 shows a portion of the diagnostic profile report's subject area breakdown information.

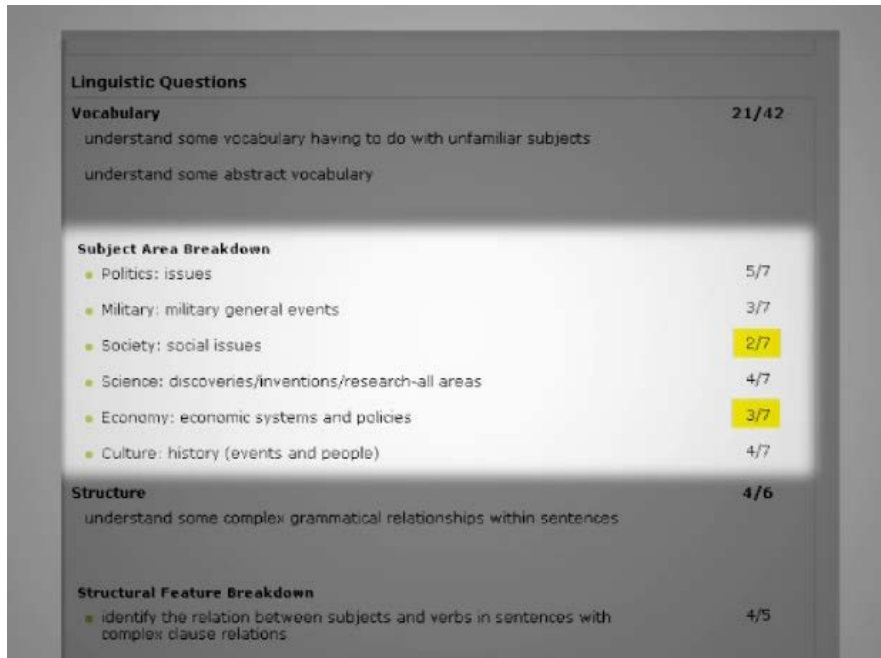


Figure 5. ODA subject area breakdown example. From *Online Diagnostic Assessment*, by Defense Language Institute Foreign Language Center, 2011, retrieved from <https://vimeo.com/16633421>. Figure is in the public domain.

The ODA is not a timed test, but requires about 1 to 2 hours for each listening or reading assessment. Assessment stimuli and questions are administered through the web-based semiadaptive features (DLIFLC, 2011). Included with the ODA diagnostic profile, a link to reading and listening learning activities from the DLIFLC GLOSS is generated for learners to work toward mastering the targeted areas (DLIFLC, 2011, 2015d). The delivery of learning activities specifically designed to meet the learner's requirements for the next level of foreign language proficiency is one of the recommended features of online diagnostic assessments that follow best formative testing practices.

European Diagnostic Assessment: DIALANG

The DIALANG is a low-stakes test used for diagnostic purposes rather than certification purposes. It was developed to identify the areas of proficiency and growth of adult foreign language learners in Europe (Council of Europe, 2001). DIALANG

includes a suite of self-assessments and vocabulary placement tests, along with a web-based diagnostic assessment tool (Alderson, 2005). It is designed to measure a student's foreign language skills in reading, writing, listening, grammar, and vocabulary of European foreign language learners. It is available for specific European languages such as Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish, Gaelic, Italian, Norwegian, Portuguese, Spanish, and Swedish, using CEFR. Like the ODA, DIALANG provides immediate score information with areas of strength as well as information about the areas of improvement (Alvarez & Rice, 2006). According to Clark et al. (2014), unlike the ODA, DIALANG gives the option of providing assessment information after each item has been completed and a final score at the end of the test for each skill set. DIALANG allows for an understanding of the student's foreign language level and provides strategies for learning improvement, which helps instructors to plan customized assignments (Alvarez & Rice, 2006). It also has the capability to store data, which provides pre- and posttest data on student progress. Available in the score report is a comparison of the self-assessment against the final diagnostic evaluation and descriptive information regarding the levels already mastered below their skill level, as well as narrative descriptions of the skills level immediately above their proficiency level (Clark et al., 2014). This free assessment has extensive student test data for certain languages that contribute to the validation of its diagnostic tool (Alderson, 2005). While the ODA was developed using the ILR Scale, recognized in the United States as the established framework of measurement for foreign language learning (Clark et al., 2014), DIALANG uses CEFR, which is the widely accepted scale for teaching and measuring foreign language in Europe. Alvarez and Rice (2006) noted that DIALANG's inability to

measure open-ended questions in the form of measuring writing and listening with extended responses and full written responses limits its capability to provide a full diagnostic measure. Haar and Hansen (2006) noted further work may be necessary for DIALANG to provide more comprehensive diagnostic criteria and expand its design to include more complex item formats to benefit fully from the capabilities of computerized assessments. Alderson and Huhta (2011) suggested that the diagnostic information available in the DIALANG reports may not provide a full spectrum of the second language learning blocks students have encountered and may not be detailed enough to provide a full understanding of each student's differentiated needs. A further limitation of DIALANG is the languages available with this tool, which mostly reflect the needs of European populations and the languages spoken in the European Union (Alderson & Huhta, 2005).

Specific ODA Studies

Only two studies were found on the ODA: a study from the University of Maryland Center for Advanced Study of Language (Clark et al., 2014.) and an unpublished Action Research study developed by McCartney and Perchaud (2014) for the DLIFLC Basic School Program. The first study provides an overview of the test design, content approach, online format, as well as diagnostic and semi-adaptive characteristics of the ODA in the context of addressing online diagnostic instruments available for second language learners, along with the assessment challenges in foreign language online instruction. The authors noted that, at this time, ODA generalizations are not feasible due to the limited materials available on the correlation of ODA raw scores to the ACTFL or ILR calibrations thus requiring administrators to read through all the

specifics of individual profiles in order to make their own generalizations. The second study, an action research correlational study of 14 DLIFLC student scores from the French Basic Course program, identified whether the French ODA for listening was an accurate diagnostic measure per the DLPT5. Data from this study indicated that there was predominant variance between the ODA and the DLPT5 of at least an ILR level higher on the DLPT5 when the ODA was administered within the same week of DLPT5 administration. Action research from McCartney and Perchaud also found that, for 43% of students, the ODA did not report a continuum increase in ODA ILR scores between two ODA administrations that had a period of instruction of 4 months between the two administrations. Additionally, action research results found that only three out of 14 students had comparable ODA/DLPT5 scores. Furthermore, only 21% of scores showed a correlation between the ODA and DLPT5 for listening. McCartney and Perchaud acknowledged that additional studies might be needed and noted that the study was performed during the validation process of the French ODA for listening. Therefore, discrepancies in ODA/DLPT5 scoring were expected to be adjusted over time after ODA validation was completed. Figure 6 shows the action research results from McCartney and Perchaud.

Students' code #	Final GPA	DLPT 5 Very Low Range	Online Diagnostic Assessment	DLPT5 Score	The difference between the ODA score and the DLPT5 score.
		Taken on 02/14	Taken on 06/05 06/06	Taken on 6/12	
001	3.5	1	1+	1+	No difference
002	3.5	1	1+	2+	Up 1 level
003	3.3	1	2	2	No difference
004	3.9	1+	2	3	Up 1 level
005	3.4	1	1+	3	Up 1.5 level
006	4.0	1+	2+	3	Up .5 level
007	3.9	1+	1+	2+	Up 1 level
008	3.2	No result	1	1+	Up .5 level
009	3.7	1	1	2	Up 1 level
010	3.8	1+	2	3	Up 1 level
011	3.5	1	1	2	Up 1 level
012	3.9	1+	No result	3	Don't know
013	3.3	1	Below 1	2	Up 1 level
014	3.4	1+	1+	1+	No difference
015	3.2	1+	1	2+	Up 1.5 level

Figure 6. ODA/DLPT5 data analysis. From *ODA Action Research Project* (p. 4), by E. McCartney & S. Perchaud, 2014, unpublished manuscript. Reprinted with permission from the authors.

Summary

The literature showed that the U.S. government has played a key role in the development of standards and accreditation measures for second language acquisition in the United States. One cause of this important role is the historical gap in the U.S. educational system when it comes to adequate foreign language instruction. It is impossible for one type of assessment instrument to fulfill the specific needs of different stakeholders, and there is a need in the education field to provide assessment information for a wide variety of reasons. Therefore, a suite of reliable and well-crafted assessments

designed to fulfill different functions is recommended to evaluate the effectiveness of learning and instruction (Darling-Hammond & Pecheone, 2010; Pellegrino, 2006, 2014; Pellegrino et al., 2001). Although Alderson and Huhta suggested that a true foreign language diagnostic test does not exist except for DIALANG (Alderson, 2005; Alderson & Huhta, 2005, 2011; Huhta, 2008), this online diagnostic assessment provides relatively limited diagnostic value because it was designed based on traditional concepts of language use rather than on a theory of foreign language acquisition and use. In this context, the ODA is more appropriate to use in the United States because (a) it takes into consideration developmental differences in the second language learners in the United States, (b) it is designed based on ACTFL instruction criteria in the United States, and (c) it is designed for students whose primary language is English. The federal government uses the ODA formative assessment for foreign language training and maintenance curriculum (U.S. Department of the Army, 2015) to identify language proficiency at the time the assessment is taken, as well as future language skills required by providing a report of specific linguistic areas to work on to achieve the next proficiency level (Clark et al., 2014). Literature review of the content development and validation process of the ODA (Chapter II, Appendix B) suggest that this online diagnostic tool provides substantiated documentation regarding the ODA standardized procedures for the development of items and stimuli, as well as for the quality control and validation procedures. Literature review also showed evidence that the ODA, generates diagnostic profiles, and provides individualized diagnostic information that helps to identify the specific areas of strength and growth that would allow a second language learner to acquire the skills at the next level of language proficiency (Appendix B and C). This

information makes it highly relevant to study the ODA. The ODA has the capability to inform teaching, give immediate feedback, and allow for remediation, and efforts to correlate this instrument to a summative assessment to identify its ability to predict student success could reassure instructors and language schools on the advantages of fully incorporating this instrument into their programs.

CHAPTER III: METHODOLOGY

Overview

A review of the literature indicated a disconnect exists between theory and practice when looking at formative and summative assessments in a more integrated manner, and limited research addressed the correlation between formative assessments and summative assessments (Crooks, 2011; Croteau, 2014; Knight, 2000; Taras, 2005). This study involved exploring the correlation between an online formative test and a summative assessment in second language acquisition. This chapter includes a description of the methodology undertaken in this study. It also includes the research questions, design, population, sample, and data collection and data analysis procedures.

Purpose Statement

The purpose of this nonexperimental correlational study was to identify the relationship between online formative (ODA) and summative (DLPT5) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success in a Basic Course program for adult students at the DLIFLC.

Research Questions

1. What is the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered at the end of the course and students' final summative DLPT5 scores?
2. What is the relationship between the ODA and the ILR levels for Spanish, Korean, Chinese Mandarin, and Standard Arabic as measured by the DLPT5?

3. Are the relationships found between ODA and DLPT5 for Spanish, Korean, Chinese Mandarin, and Standard Arabic consistent across the levels or is there variance in the relationship depending on the level?

Research Design

A nonexperimental design requires the observation of relationships without controlling or changing the phenomena or the subjects. A nonexperimental design typically includes a descriptive, comparative, survey, or correlational design (McMillan & Schumacher, 2006). The appropriate method for this research study was nonexperimental correlational research through a standard regression model. A quantitative correlational method requires data analysis to determine the relationships between selected factors (Cohen, Manion, & Morrison, 2003). A standard multiple regression model was suitable for identifying if a specific result existed and the amplitude or extent of this result.

As identified by Pellegrino (2004), four independent areas help identify the theories that have contributed to the types of assessments currently available: (a) cognition theory and research, (b) classroom-based assessments, (c) psychometrics theory and research, and (d) large-scale assessments. Formative, classroom-based assessments are influenced by cognition theory and research, and summative, large-scale assessments are influenced by psychometric constructs. Because the literature review revealed a disconnect between theory and practice when looking at formative and summative assessments in an integrated manner, this study involved exploring the correlation between an online formative test and a summative assessment. Figure 7 shows the four spheres of work in educational assessment practice as described by Pellegrino.

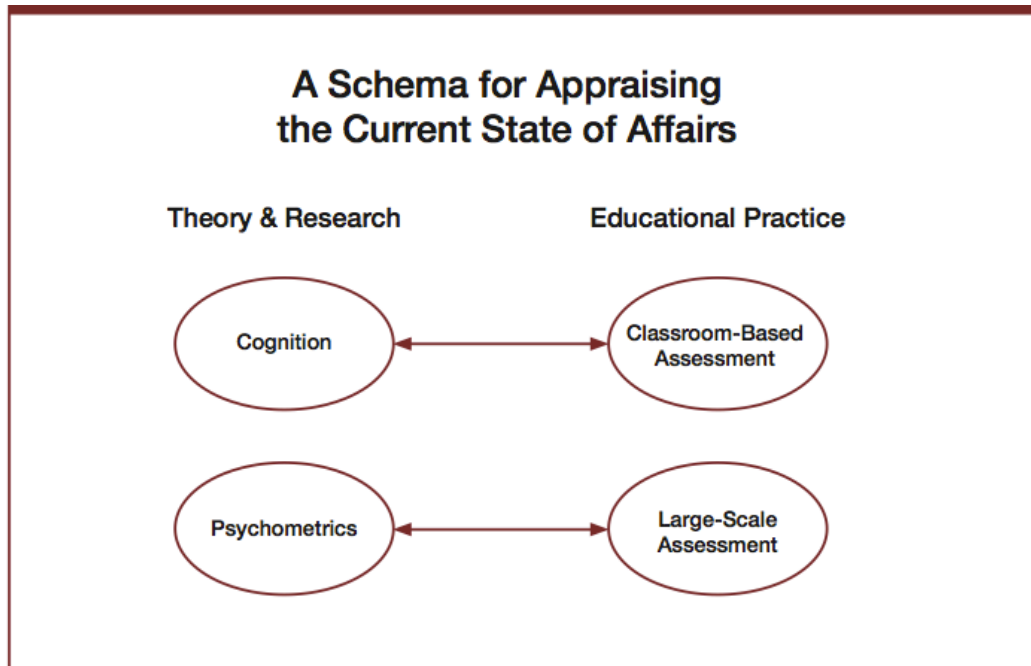


Figure 7. A schema for appraising the current state of affairs. From *The Evolution of Educational Assessment: Considering the Past and Imagining the Future* (p. 10), by J. W. Pellegrino, November 17, 1999, retrieved from <https://www.ets.org/Media/Research/pdf/PICANG6.pdf>. Copyright 1999 by J. W. Pellegrino. Reprinted with permission.

The research design included the correlation between two variables: (a) end-of-course ODA scores and (b) DLPT5 final scores. In general, the ODA is used during the semester program to inform instruction. It is then used at the end of a course program to measure student progress and to predict DLPT5 scores. The following archival scores were used for this nonexperimental correlational design:

- Archival scores for listening and reading from students who took the formative ODA at the end of the 36-week course in Spanish and archival scores of the same students who took the DLPT5 at the end of this program.
- Archival scores for listening and reading from students who participated in a formative ODA at the end of the 64-week course in Korean, Chinese, and

Standard Arabic and archival scores of the same students who took the summative DLPT5 at the end of this program.

The formative assessment (ODA) identifies the strengths and areas of improvement in Spanish, Korean, Chinese Mandarin, and Standard Arabic for listening and reading and provides individualized feedback according to the ILR guidelines. The DLPT5 is a summative assessment that measures the final foreign language proficiency in listening and reading, also based on the ILR. This study involved analyzing archived data from ODA and DLPT5 Spanish, Korean, Chinese Mandarin, and Standard Arabic for listening and reading in a nonexperimental correlational study using a multiple regression model. Maturation issues were avoided to the extent possible by not selecting extended courses that were beyond the standardized length of the Spanish, Korean, Chinese Mandarin, and Standard Arabic programs. Issues regarding internal validity were considered, particularly given that the data were the result of archival information from a period of 2 years.

A linking study requires a clear understanding of the type of evidence resulting from the relationship between two assessment instruments. This information could help formulate the appropriate correlation study and the type of quantitative instruments required (Deming, 1980). The success of this correlation depends on the quality of the strategies used, as well as the commonality of assessment construct goals of these two assessments (Mislevy, 1992).

A factor that may affect the quality of a correlation between two assessment instruments is the number of testing samples, procedure for selecting these samples, type of quantitative formulas used to estimate their margin of error, differences between test

administrators, scheduled times of administration, test-taking conditions, test instruments selected, and differences between data-gathering methods. In these cases, the selection of the appropriate statistical tools helps to discern and understand these data (Deming, 1980).

Researchers need to address two main issues carefully when performing a linking study of an assessment instrument: (a) understanding the type of evidence resulting from the relationship between two assessment instruments and (b) formulating an effective correlation study with the appropriate quantitative instrument (Deming, 1980). In this context, a careful understanding of the appropriate quantitative procedure for comparing two assessments is necessary to develop an adequate correlation. Two different assessments can be correlated through equating correlation, projection, or moderation studies to identify the relationship between the scores from these two assessment instruments. The decision to select a specific type of study depends on a clear understanding of the purposes of the study as well as an accurate understanding of the similarities and differences between the assessment instruments the researcher is trying to correlate. First, an equating correlation study assumes a close correspondence between the blueprint of two assessments so a one-on-one equating of items can be performed. Second, a calibrating correlation study assumes some differences in the length and type of tasks of the assessments so an adjustment of the scale is necessary to account for the differences between two assessment constructs. In this case, a one-to-one correspondence table between two assessments is not feasible. Third, a projection correlation study is appropriate for assessment instruments that have varied tasks, testing conditions, or purposes or are conducive to a different level of student motivation. These instruments

usually require a probability distribution estimate. Fourth, a moderation correlation study is necessary for studies where two assessments may be different and thus require administration to different groups of students; for example, a study to identify the type of comparability between a French and a Portuguese test. Unlike the previously described correlation studies that required sampling two tests with the same student population, a moderation correlation study requires two groups of test takers: the students who took a French test and those who took a Portuguese test. This type of study would require statistical moderation studies with score distribution studies known as scaling (Angoff, 1984; Mislevy, 1992). The selection of the projection correlation analysis was the most appropriate statistical tool. A projection correlation study is usually applied to correlate assessment instruments with different tasks, testing conditions, or purposes (Deming, 1980). The DLPT5 and the ODA meet these characteristics. Although these two assessments have assessment construct goals in common, they have different tasks, testing conditions, and differences in outcomes because of their respective summative and formative characteristics.

Another factor that may affect efforts to gather reasonable evidence of a correlation is the number of variables that need measuring. The more variables there are, the less confidence there is in the assumptions (Deming, 1980; Mislevy, 1992). To reduce the number of variables, this study ensured that only classrooms that had the ODA administered at the end of the course were selected. Other factors may increase the number of variables in a correlation among two assessment instruments, including the number of testing samples, the procedure for selecting these samples, the type of quantitative formulas used to estimate their margin of error, the differences between test

administrators, the scheduled times of administration, and the test-taking conditions. The selection of the projection correlation was the most appropriate statistical tool to take these factors into account (Deming, 1980).

Population

Research populations usually include of a number of individuals, cases, or elements that meet the requirements for a scientific study for which researchers want to make some generalizations. Because researchers may not be able to make generalizations of a whole population, they may rely on a specific sample or target population, known as the survey population or sampling frame (McMillan & Schumacher, 2010).

Each calendar year, approximately 3,500 students attend the Basic Course programs available at the DLIFLC (DLIFLC, 2015c). For the languages studied, the total population in 2016 and 2015 at the Basic School Program consisted of 342 students for Spanish, 426 students for Korean, 571 students for Chinese Mandarin, and 912 students for Standard Arabic. DLPT5 archived data from previous years were also obtained. The breakdown of the population of this study appears in Table 1.

Table 1

DLPT5 and ODA Archived Scores Used for Study

DLPT5 and ODA score matches available	Spanish	Korean	Chinese	Standard Arabic
1 week to 3 months of DLPT5 administration	116L/118R	35L/39R	65L/66R	53L/47R
Breakdown by school				
Number of students	118	39	66	53
Population per school/year (2016)	184	211	313	419
Population per school/year (2015)	158	215	258	433
Total population (2015 + 2016)	342	426	571	912

Note. R = reading. L = listening.

The DLIFLC Basic Course population of approximately 3,500 military students each year consists of students from all military branches in the United States. The student population has a variety of academic backgrounds and comes from all parts of the country. These students are assigned to a specific language school based on their score results on the DLAB placement test and the military's needs.

The population selected for this study consisted of 2,251 adult military students taking the 36-week Basic Course Spanish program or the 64-week Korean, Chinese Mandarin, or Standard Arabic Basic Course program in 2015 and 2016 in a government setting in Monterey, California. Students took the Spanish, Korean, Chinese Mandarin, or Standard Arabic ODA formative assessment a few days to three months before the end of the program. At the end of the course, the same students took a summative test, the DLPT5, as part of their graduation requirements.

Sample

The individuals from a group or population about whom studies or assumptions are being made are usually described as a sample, which can be the whole population or a smaller group selected from a population (McMillan & Schumacher, 2010). When the whole population cannot be studied, a target population is usually selected. The specific number of people from whom information can be obtained comprises the target population, also known as the sampling frame. The specific individuals selected from the sampling frame or target population are the sample. The larger a sample is, the higher the confidence of a close approximation to the results that can be obtained from a sampling frame or target population (Creswell, 2012).

For a quantitative study, a standard number generally recommended as the minimum sample size is 30. However, Onwuegbuzie (2003) cautioned that sample sizes of 30 might not provide strong information in correlation studies. Therefore, the research objective is also a factor that contributes to the sample size estimate. Onwuegbuzie recommended using statistical power analysis to determine the sample size in correlation studies. Per statistical power analysis, the recommended sample size for correlation studies is 64 participants for one-tailed studies and 82 participants for two-tailed hypotheses (Onwuegbuzie & Collins, 2007).

With the approval from the Office of the Commandant, delivery of data to the researcher was granted. With approval of the DLIFLC provost, archived DLPT5 score information was delivered to the Office of the Deputy Chief of Staff for Information Technology (DCSIT) by the Directorate of Academic Affairs. Student information was replaced with an identification (ID) code. DCSIT matched DLPT5 scores to ODA scores and delivered an Excel document via a secure site containing the cells shown in Table 2.

Table 2

Excel Document Format for Data Delivery

	DLPT5	DLPT5	DLPT5	DLPT5	ODA L	ODA L	ODA R	ODA R	
Language	ID	L score	L testing	R score	R testing	score	testing	score	testing
code	level	date	level	date	level	date	level	date	
Spanish									
Korean									
Chinese									
Mandarin									
Standard									
Arabic									

Note. R = reading. L = listening.

Out of the 800 DLPT5 scores that consisted of 200 scores per language for listening and reading, it was estimated that a minimum of 100 DLPT5 and ODA score

matches per language could be obtained for Spanish, Korean, Chinese Mandarin, and Standard Arabic. The original assumption was that schools might administer the ODA consistently at the beginning and at the end of the school program. According to the archived data, the actual scores available were fewer than the estimated minimum of 100 per language, as shown in Table 3.

Table 3

DLPT5 and ODA Archived Scores Used for Study

DLPT5 and ODA score matches available	Spanish	Korean	Chinese	Standard Arabic
1 week to 3 months of DLPT5 administration	116L/118R	35L/39R	65L/66R	53L/47R
Breakdown by school				
Number of students	118	39	66	53
Population per school/year (2016)	184	211	313	419
Population per school/year (2015)	158	215	258	433
Population (2016/2015)	342	426	571	912

Note. L = listening. R= reading.

Except for the Spanish sample, Onwuegbuzie’s (2003) formula for sample sizing could not be implemented due to the actual sample size of the archived data. Therefore, Krejcie and Morgan’s (1970) formula for determining research sample sizes was used instead. This formula helped identify the level of confidence and margin of error of the study based on the population and sample size for each language:

$$n = \frac{\chi^2 \times N \times P \times (1 - P)}{(ME^2 \times (N - 1) + (\chi^2 \times P \times (1 - P)))}$$

where n = sample size, χ^2 = chi square for the specified confidence level at 1 degree of freedom, N = population size, P = population proportion, and ME = desired margin of error.

According to Krejcie and Morgan’s formula, the following was determined:

1. Spanish sample size of 118 students = 82% level of confidence with a .05 margin of error.
2. Korean sample size of 39 students = 49% level of confidence with a .05 margin of error.
3. Chinese Mandarin sample size of 66 students = 61% level of confidence with a .05 margin of error.
4. Standard Arabic sample size of 53 students = 54% level of confidence with a .05 margin of error.

Only archived data that showed the ODA was administered at the end of the course were selected to ensure reliable test administration results and homogeneous population samples. The archived data selected were representative of available sampling strategies (McMillan & Schumacher, 2010). Researchers use available sampling in cases of limited data accessibility.

Instrumentation

The data collection instruments used in this research study consisted of archived data from eight formative ODA assessments and eight summative DLPT5 assessments developed by DLIFLC, as noted in Table 4.

Table 4

Data Collection Instruments

Spanish	Chinese Mandarin	Korean	Standard Arabic
Reading ODA	Reading ODA	Reading ODA	Reading ODA
Listening ODA	Listening ODA	Listening ODA	Listening ODA
Reading DLPT5	Reading DLPT5	Reading DLPT5	Reading DLPT5
Listening DLPT5	Listening DLPT5	Listening DLPT5	Listening DLPT5

Each of the ODA reading and listening assessments for Spanish, Korean, Chinese Mandarin, and Standard Arabic consisted of a set of four to six items following the configuration below:

Testlet for Level 1

Section 1: Content-based items

Main idea type question

Supporting idea type question

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structure item

Testlet Level 1+

Section 1: Content-based items

Main idea type question

Supporting idea type question

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structural item

Discourse item

Testlet for Levels 2, 2+, and 3

Section 1: Content-based items

Main idea type question

Supporting idea type questions (two items)

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structural item

Discourse item

(DLIFLC, 2015a, 2015b, 2015d).

After items are approved and placed into testlets, they go through a cycle known as testlet iteration. This process requires a minimum of three testlets per level for Levels 1 to 3 to fulfill the computer adaptive requirements for upward or downward performance-level mobility. After all testlets are developed and accurately reviewed to ensure that items within each testlet measure the specific levels targeted, the adaptive features can also be tested. Sets of three testlets are necessary for upward and downward mobility to verify the accurate proficiency level of test takers.

Therefore, an ODA iteration requires a minimum of three testlets for each level and a total of six testlets for Levels 1 to 3. This procedure ensures the adaptive requirements of the ODA are met, as well as the quality standards specific to formative assessments such as the ODA. The minimum number of testlets needed to meet ODA computer adaptive requirements appears in Table 5.

Table 5

ODA Number of Testlets

Level	Number of testlets
1	6
1+	12
2	9
2+	6
3	6
Total	39

Note. From Online Diagnostic Assessment Team Program Review (p. 13), Defense Language Institute Foreign Language Center, 2015d, Monterey, CA: Author. Public domain.

The ILR Scale determines the scores for the ODA. A separate score is provided for listening and reading. The scores range from 1 to 3. Each of the DLPT5 computer-based summative tests for Spanish, Korean, Chinese, and Standard Arabic has a Lower-Range test and an Upper-Range test. For the Lower-Range test, each instrument includes approximately 60 test items, including more than 30 stimuli for reading and 40 stimuli for listening. Each stimulus has about four multiple-choice items for reading and two multiple-choice items for listening. For the Upper-Range test, each reading and listening assessment instrument includes about 36 test items. There are approximately 14 stimuli for reading and 14 for listening. Each reading stimulus contains five multiple-choice items, and each listening stimulus contains three multiple-choice items.

The scores for the DLPT5 are determined by the ILR Scale. An average of the reading and listening score is created to provide an overall ILR score. The scores range from 0+ to -4. The DLPT5 is available in two difficulty ranges: the Lower-Range test (for levels 0+ to -3) and the Upper-Range test for students who received an ILR score of 3 in the Lower-Range test (DLIFLC, 2015b).

Validity and Reliability of the DLPT5

The DLPT5 is the only approved summative assessment instrument used by the DoD for the certification of foreign language proficiency for military personnel. It was approved by the under secretary of defense for personnel and readiness (DoD, 2013). The DLPT5 is also the only summative assessment approved to identify the qualifications in foreign language proficiency to grant foreign language proficiency bonuses to military personnel. As part of the validity process, the deputy under secretary of defense for program integration oversees the DLPT5 in terms of the research analysis, quality

control, and development and test administration; provides quarterly reports on different activities that include research analysis and possible irregularities; and works with DLIFLC to ensure and sustain the established psychometric criteria (DoD, 2009).

According to Petersen and Cartier (1975), DLPT5 validity is ensured in terms of (a) criterion-related validity, which is not the same as criterion-referenced tests, through the comparison of test scores—and the indicators resulting through these scores—against an external criterion or variable; (b) content validity, which is represented by the accuracy in which the content of the test reflects the subject matter of instruction; and (c) construct validity, which addresses the degree to which the DLPT5 measures what it intends to measure. Petersen and Cartier noted that the construct validity of foreign language tests of variable language complexity requires extensive consideration. According to Petersen and Cartier, and due to the complexity of some of the languages assessed, “since construct validation presents enormous theoretical and practical problems, the most reasonable intermediate approach to establishing the validity of the DLPTs appears to be through content validation” (p. 115). According to Petersen and Cartier, the larger weight on content validity lends itself to a heavier reliance on linguistics over statistics; however, they noted that an important factor that contributes to the validity (criterion-related validity and construct validity) of the DLPT5 is the availability of plentiful data and large sample sizes within DLIFLC.

According to the Test Development Division, Evaluation and Standardization, DLIFLC (2007), two parallel forms were developed for each foreign language assessment to ensure test validity. Items were administered prior to item selection by choosing a sample of test takers from DLIFLC, military bases, and universities, with a higher

number of students from DLIFLC and military bases. The items were administered to more than 100 students (Keesling, 2007) after development and quality control cycles using item development industry standards and ILR criteria (Test Development Division, Evaluation and Standardization, DLIFLC, 2007). According to Dr. Mika Hoffman, former Dean of the Test Development Division at DLIFLC, the ILR Scale was used in a more systematic way with the development of the DLPT5 to ensure greater validity and calibration methods, which included the configuration of standard-setting panels for setting DLPT5 cut scores. As part of the DLPT5 validation, new processes were introduced with ILR experts from different languages during the item development process. Each passage and item went through an independent review by the Proficiency Standards division to ensure a consistent interpretation of the ILR performance-level descriptors across languages during the test development phase. After the Proficiency Standards Division completed and verified test development, a pre-standard-setting discussion with ILR experts from different languages was introduced to the validation process to interpret the ILR performance-level descriptors in the context of DLPT5 measurement. The pre-standard-setting panel was an important strategy set to ensure the ILR was used in a more systematic way. Lastly, the standard-setting phase, as a crucial step in the validation process, involved applying standardized procedures that used the ILR performance-level descriptor statements in a clearly organized and categorized process across languages, which resulted in an explicit standard setting that ensured greater validity for the different DLPT5 language instruments. According to M. Hoffman,

The standard-setting itself was a crucial step in operationalizing the ILR PLDs [performance level descriptors] for the DLPT5, since we were explicitly

determining expected performance at the item level, and using IRT psychometric analysis to use that information in determining cut scores, rather than using an arbitrary standard of percent correct, which did not take into account variations in difficulty. (personal communication, June 28, 2017)

Item results were analyzed using classical item analysis statistics along with other types of item analysis, including three-parameter logistic analysis. Items statistics were used to identify items with a negative point biserial, that is, items not answered correctly by students with high test results. Items with a negative point biserial and very easy items characterized by a very low discrimination or positive point biserial were not selected as part of the test calibration (Keesling, 2007). Although it was important to select items that accurately represented the ILR levels of proficiency during the DLPT5 phase, items that showed a high or difficult level of performance due to a poor item design rather than a high item discrimination along with content and ILR criteria were eliminated (Test Development Division, Evaluation and Standardization, DLIFLC, 2007). As part of the DLPT5 validation process, after problematic items were eliminated, qualified items representing DLIFLC's content requirements, ILR requirements, and item parameter and three-parameter estimate requirements were included in the item calibration process using a program known as BILOG-MG. Through this process, item parameters that identify the probability in which each item answered correctly relates to proficiency were identified. Although it is understood that the more proficient a student is, the more likely he or she is to answer each item correctly, the probability for less proficient students to respond correctly to an item was expected to decrease based on student proficiency.

According to Keesling (2007), construct validity is demonstrated through the required statistical procedure appears in Figures 8-11. This procedure is used to select items expected to measure each ILR level. As each set of items representing each ILR level is selected, items are expected to fall to the right of the prior ILR curve.

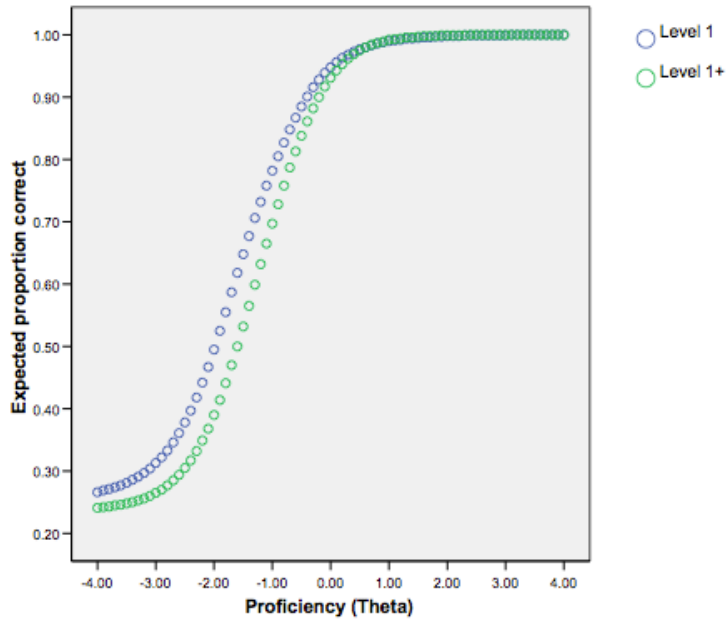


Figure 8. DLPT5 item pools at ILR levels 1 and 1+. From *Validity and Reliability of DLPT5 Multiple-Choice Tests* (p. 101), by J. W. Keesling, 2007, retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf. Defense Language Institute Foreign Language Center, 2009. Monterey, CA: Author. Public domain.

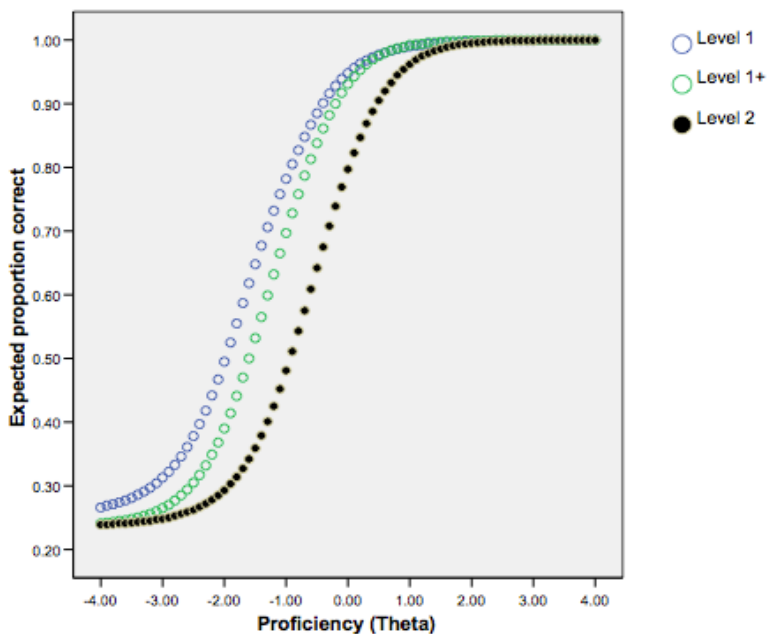


Figure 9. DLPT5 item pools at ILR levels 1, 1+, and 2. From *Validity and Reliability of DLPT5 Multiple-Choice Tests* (p. 101), by J. W. Keesling, 2007, retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf. Defense Language Institute Foreign Language Center, 2009. Monterey, CA: Author. Public domain.

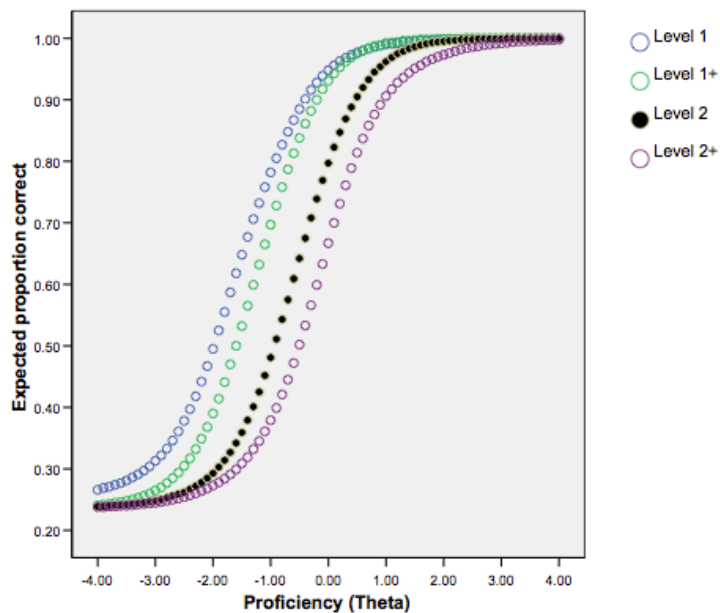


Figure 10. DLPT5 item pools at ILR levels 1, 1+, 2, and 2+. From *Validity and Reliability of DLPT5 Multiple-Choice Tests* (p. 102), by J. W. Keesling, 2007, retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf. Defense Language Institute Foreign Language Center, 2009. Monterey, CA: Author. Public domain.

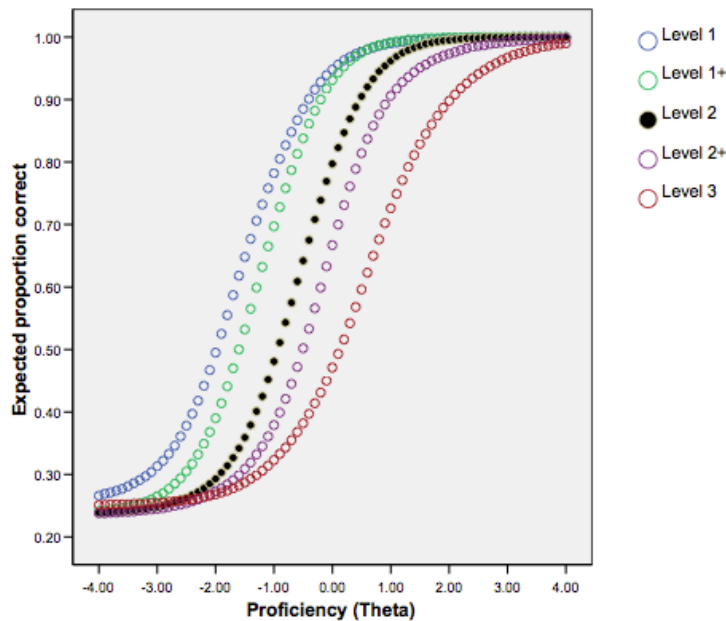


Figure 11. DLPT5 item pools at ILR levels 1, 1+, 2, 2+, and 3. From *Validity and Reliability of DLPT5 Multiple-Choice Tests* (p. 102), by J. W. Keesling, 2007, retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf. Defense Language Institute Foreign Language Center, 2009. Monterey, CA: Author. Public domain.

The DLPT5 also included other methods to ensure construct validity, such as the theta cut-score. This method was used to identify the progressive approximation of proficiency levels after items at the appropriate ILR level were selected. After this procedure was complete, the next step toward validation of the DLPT5 items included the final selection of items for two operational forms for listening and reading (Keesling, 2007). This item selection required the two forms to be parallel in length and item distribution as well as in ILR difficulty range. Other criteria included the selection of items with good discrimination, which means items contain plausible but incorrect responses as described by the statistical information in their point-biserial correlation. After forms have been selected, the calibration process was started to identify cut-scores that differentiated among the ILR levels. At this stage, the raw scores from each

operational form that corresponded to the preestablished theta cut-scores were identified. By using the theta score corresponding to each ILR level, along with the raw score from the operational items for each form, a probability score was produced. This probability score established the probability for answering each item correctly. According to Keesling (2007), this process validated the item selection by showing the probability for an easy or difficult item to be answered. While the probability for an easy item was expected to be high, the probability for a difficult item to be answered was expected to be low. When those probabilities were added, the raw score for each proficiency level was identified.

Keesling (2007) included an example of a number of correct cut-scores for two operational forms. This information shows that careful psychometric criteria were used for the validation of the DLPT5, although an extended document showing all cut-scores for all forms selected per language may further enhance the thorough criteria for validation presented in this document. Table 6 shows an example of the DLPT5 theta cut-scores provided by Keesling.

Table 6

Theta Cut-Scores Based on the 70% Mastery Criterion

	Theta	Number correct	
		Form A	Form B
Cut-score between Levels 0+ and 1	-1.320	17.808	17.058
Cut-score between Levels 1 and 1+	-0.992	20.833	40.457
Cut-score between Levels 1+ and 2	-0.325	29.645	29.893
Cut-score between Levels 2 and 2+	0.101	36.158	36.385
Cut-score between Levels 2+ and 3	0.894	45.661	45.266

Note. From *Validity and Reliability of DLPT5 Multiple-Choice Tests* (p. 104), by J. W. Keesling, 2007, retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf. Defense Language Institute Foreign Language Center, 2009. Monterey, CA: Author. Table is in the public domain.

Researchers at the Test Development Division, Evaluation and Standardization, DLIFLC (2007) ensured the reliability of the DLPT5 by processing the aggregated data from the raw responses through a calibration of the operational forms. A statistical tool known as WINSTEPS was used to compute an estimate of the measure of internal reliability consistency for each pair of operational forms. This procedure is known as Cronbach's alpha or KR-20. To show parallel forms reliability, both forms need to show agreement in the score production and demonstrate consistency in the production of lower and higher scores. The reliability should also show consistency in the production of graduation scores. In 2007, the criteria for graduation scores required a minimum score of 2. As the graduation criteria were recently raised to Level 2+, an update of this information to demonstrate reliability for parallel forms at Level 2+ was recommended. This might be particularly meaningful, as the cases where parallel forms showed some differences were at the plus levels (Keesling, 2007). The DLPT5 describes procedures to demonstrate reliability across forms and across levels, which include the Pearson product-moment correlation, Spearman correlation, Kappa correlation, and intraclass correlation coefficient and describes the assets and limitations of these correlations and their preferred procedure of using the intraclass correlation coefficient to show the most accurate estimates for the requirements of the DLPT5 forms.

Validity and Reliability of the ODA

Two types of validity are considered for assessment instruments: face validity and content validity (Lynn, 1986). Researchers use face validity to address issues that relate to how an assessment reflects what it intends to measure at face value based on its external appearance. Based on the validation process shown on the ODA Validation

Process for the ODA Test Design, Stimulus Selection, Item Distribution, and Examples of Item Formats (see Appendix B), the ODA shows face validity. Concerns about content validity were addressed by the quality control procedures described as part of the ODA development process, including testlet iteration, ODA workflow, and ODA server database system (see Appendix B). Also of relevance for the content validity is the review and validation process of the ODA, which requires different ODA stakeholders to participate in the field-testing process. Reviewers include in-house developers, students, language schools, military bases, and DLIFLC language training detachments. The field-testing process includes checking the performance of the site as well as the item testlets (see Appendix B). According to DLIFLC (2015d), the validation cycles after the ODA testlets are complete are as follows:

1. Item testlets are made available for testing through an Internet testing site.
2. Through the test-taking process, the system is debugged to ensure the interface works as expected.
3. Developers and reviewers take the test in its preoperational form through the Internet site.
4. Revisions are made based on input from developers and reviewers.
5. Testing is performed with native speakers to review appropriateness of test at the higher levels, particularly Level 3.
6. Revisions are made.
7. Items are validated through the administration of the testlets to groups of students with different language ability levels and at different stages in the school semester to verify testlet levels and item discrimination.

8. Revisions are made.
9. Items are made operational through the ODA official Internet site.
10. Items are monitored to verify that they measure the target level and are able to produce discriminating output between levels according to ILR criteria.
 - a. Items are verified to ensure they lead to the targeted student performance outputs.
 - b. Testlets are verified to ensure they produce the expected floor and ceiling output per testlet ILR level design.
 - c. Level testlets are validated to make sure that, for example, a Level 1+ student performs as expected on a Level 1 testlet but has difficulty at a Level 2 testlet, while a Level 2 student performs as expected on a Level 2 testlet, but has difficulty with a Level 3 testlet.
11. Items are also monitored to ensure they lead to the expected open-ended item responses, the answers have the expected complexity and completeness, and all possible correct responses are taken into account (DLIFLC, 2015d).

Of equal importance to the content validity of the ODA is the incremental integration of testlets over time, as well as its technical capability to monitor the ODA results to make timely updates to the ODA assessment instrument after it is fully functional. This monitoring and updating of the ODA helps developers remove unexpected outliers, unforeseen discrepancies, or unidentified content issues found by users and include a user's survey. For an online test, the test taker's responses could further strengthen the quality of its diagnostic assessment and diagnostic profile. In this context, overseeing and reviewing the ODA's assessment performance results once the

ODA has become operational is an important step that is critical for the validation process of the ODA and is unique and relevant to well-designed formative diagnostic assessments.

Lastly, and an essential aspect of the content validity for well-designed online diagnostic tests after ODA items and testlets become operational, the ODA system is monitored through a database. The database includes an automated feature labeled “item-user correlation” that helps identify the level of discrimination between items and testlets across levels as well as the validation of all possible correct answers for open-ended items. Through this monitoring process, some items may be replaced or updated because content may have become outdated, societal and cultural exposure to certain content may elicit prior knowledge responses over time, items may not provide the expected outcomes, or a need arises to develop new content on an area or skill where gaps exist (DLIFLC, 2015b).

Best practices exist for summative assessments, but researchers disagree about whether these practices should be considered when selecting or developing formative tests. These include reliability measures that ensure the assessment results are (a) predictable and consistent when administered to students with the same skills and abilities, (b) valid so they measure what they intend to measure and their results lead to suitable instructional decisions, and (c) fair so that students’ responses are predictable and consistent across all students (Haertel, 2006; Pellegrino et al., 2001; Trumbull & Lash, 2013).

Durán (2011) noted that traditional applications of validity and reliability measures may not be feasible with formative assessments. However, according to Durán,

the application of formative strategies contribute to their validation because instructors have the option of measuring domains frequently. In this context, the possibility of building a body of performance results from formative assessments administered on an ongoing basis increases the level of confidence in the type of assessment conclusions and strength of the formative assessment instrument (Shavelson et al., 2007). In this context, formative assessments strengthen their constructs through the frequent evaluation of students (Durán, 2011) and therefore ensure their validity and reliability over time through the ongoing gathering of student data directly by instructors and the frequent updating of the assessment instruments per input resulting from the data gathered (Shavelson et al., 2007). Evidence of reliability on the ODA is shown by the frequent updates of foreign language assessments resulting from ongoing student data gathered during the ODA process.

Data Collection

Data collection began after the study received Institutional Review Board (IRB) approval from both the Brandman and the DLIFLC IRB committees. On March 14, 2016, the research protocol was approved by the DLIFLC Scientific Review Board and submitted to the Office of the Commandant, Philip J. Deppert, for consideration. On March 15, the Office of the Commandant expressed its willingness to grant permission to use the requested archival data in the study upon submission of IRB approval from the Brandman IRB committee. The researcher obtained archived scores of the ODA and the DLPT5 from the DLIFLC's administrative review and IRB review and after final approval from the Office of the Commandant. Obtained data included ODA scores and

DLPT5 test scores from DLIFLC archives from students who took both tests. Archived scores obtained included the information presented in Table 7.

Table 7

Data Available from Second Data Pull

Language	DLPT5 minimum number of archived scores requested	ODA score matches found (second data pull)	DLPT5 and ODA score matches meeting correlation requirements
Spanish	200	166	116L/118R
Korean	200	174	35L/39R
Chinese Mandarin	200	179	65L/66R
Standard Arabic	200	179	53/47R

ODA and DLPT5 data were matched by DCSIT and provided to the researcher with an ID code to ensure student confidentiality. Information was provided showing the ODA and the DLPT5 scores of the same ID code representing a student, as shown in Figure 12.

Language KOREAN	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date
Language CHINESE	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date
Language STANDARD ARABIC	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date
Language SPANISH	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date

Figure 12. The ODA and the DLPT5 scores of the same ID code representing a student.

Information collected was transferred into an Excel database and to analytical software known as SPSS. Data were screened to remove the records of subjects whose ODA scores were outside of the testing window requirements for this research. The

collected data consisted of an ID number, DLPT5 scores for listening and reading, and ODA scores for listening and reading. The archived data identified the time of the year when the ODA was administered.

Data Analysis

Quantitative correlation studies involve a relationship between two variables. If the variables are simple, a simple correlation is needed. If a researcher needs to determine how a score from an independent variable predicts a score for a dependent variable, then a correlation study known as bivariate regression is more appropriate. Because of the need to address several independent variables in this study, a multiple regression was necessary. Multiple regression provides the flexibility needed in correlation studies with different types of variables, whether ordinal or nominal. Researchers also commonly use multiple regression in testing to understand why a group of test takers may have different scores when correlated to a dependent variable. Regression studies are highly recommended for monitoring specific variables to identify a group of independent variables and a dependent variable (McMillan & Schumacher, 2010).

A multiple regression analysis was performed with the ordinal variables end-of-course ODA scores in listening and reading and final DLPT5 scores in listening and reading to determine if a relationship existed between an online formative assessment, the ODA, and the summative assessment DLPT5. Researchers can correlate two different assessments through equating correlation, projection, or moderation studies to identify the relationship between the scores from the two assessment instruments. The decision to select a specific type of study depends on a clear understanding of the purposes of the study, as well as an accurate understanding of the similarities and differences between the

assessment instruments being correlated. A projection correlation study is appropriate for assessment instruments that have varied tasks, testing conditions, or purposes or are conducive to a different level of student motivation. These instruments usually require a probability distribution estimate (Deming, 1980). The projection correlation study is usually applied to correlate assessment instruments with different tasks, testing conditions, or purposes, as in the case of the DLPT5 and the ODA. Although these two assessments have a commonality of assessment construct goals, they have different tasks and testing conditions and differences in outcomes because of their respective summative and formative characteristics. A projection correlation study was not performed for this research due to the limited archived data available, along with the sparse projection correlation models available that could be applied to this specific study.

Multiple regression analyses were performed using SPSS software with two dependent variables, ODA reading and listening scores, and two independent variables, DLPT5 reading and listening scores, to identify (a) the measurable gains in Spanish, Korean, Chinese Mandarin, and Standard Arabic reading and listening proficiency obtained by using the formative ODA, as measured by the summative test DLPT5; (b) the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered during the fourth quarter of the course and students' final summative DLPT5 scores; and (c) the impact of Spanish, Korean, Chinese Mandarin, and Standard Arabic online formative assessments as a valid measure of foreign language proficiency in terms of ILR levels as measured by the summative DLPT5.

This study followed rigorous criteria and all requirements for the application of multiple regression models to ensure systematic and scientific results emerged from a correlation study. The data performed for the research questions were analyzed using descriptive statistics and progression correlation techniques. The results were used to identify the relationship between online formative (ODA) and summative (DLPT5) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success. A summary of the data analysis for the study appears in Table 8.

Table 8

Data Analysis

Question	Data used	Analysis
1. What is the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered at the end of the course and students' final summative DLPT5 scores?	ODA reading and listening posttest scores, final DLPT5 scores	Pearson product-moment correlation study of the ODA reading and listening score results to the DLPT5 reading and listening score results
2. What is the relationship between the ODA and the ILR levels for Spanish, Korean, Chinese Mandarin, and Standard Arabic as measured by the DLPT5?	ODA reading and listening posttest scores, final DLPT5 scores	Pearson product-moment correlation study of the ODA reading and listening score results to the DLPT5 reading and listening score results along with Excel spreadsheet distribution of ODA scores by ILR level per DLPT5.
3. Are the relationships found between ODA and DLPT5 for Spanish, Korean, Chinese Mandarin, and Standard Arabic consistent across the levels or is there variance in the relationship depending on the level?	ODA reading and listening posttest scores, final DLPT5 scores	Pearson product-moment correlation study of the ODA reading and listening score results to the DLPT5 reading and listening score results along with Excel spreadsheet distribution of ODA scores by ILR level per DLPT5.

Limitations

According to McMillan and Schumacher (2009), validity is “the degree to which scientific explanations of the phenomena match reality” (p. 104). Therefore, validity helps to identify if the data reflect the observed phenomena. Reliability is the degree to which an assessment tool produces consistent results (Phelan & Wren, n.d.).

One issue that may affect the correlation of two assessment instruments is the environment in which the assessment took place. For example, students may show what they truly know to a higher or lower extent based on their level of motivation as well as the testing conditions to which they were exposed. In this context, testing conditions of a classroom or school program, as well as the level of motivation toward taking an assessment instrument, may affect the accuracy of correlation assumptions (Deming, 1980; Mislevy, 1992).

Another factor that may affect an attempt to gather reasonable evidence of a correlation is the number of variables that need measuring. The more variables there are, the less confidence there is in the assumptions. Doing a correlation study where the formative assessment may be administered halfway through a semester course while the summative assessment is administered at the end of the semester course may introduce too many variables that could affect attempts to formulate clear inferences (Deming, 1980; Mislevy, 1992). Therefore, this study only used data from students who took the ODA at the end of the course.

Two main issues need to be carefully addressed when performing a linking study of an assessment instrument: (a) understanding the type of evidence resulting from the relationship between two assessment instruments and (b) formulating an effective

correlation study with the appropriate quantitative instrument (Deming, 1980). In this context, a careful understanding of the appropriate quantitative procedure for comparing two assessments is necessary to develop an adequate correlation. A projection correlation study is usually applied to correlate assessment instruments with different tasks, testing conditions, or purposes, as in the case of the DLPT5 and the ODA. Although these two assessments have a commonality of assessment construct goals, they have different tasks, testing conditions, and differences in outcomes because of their respective summative and formative characteristics. The posttest ODA and DLPT5 data already existed in archived data. A projection correlation study was not performed due to the limited archived data available, as well as the sparse projection correlation models available.

Internal validity regarding history was a concern. To avoid maturation issues as much as possible, the 6-month Basic Course program was selected instead of the 9-month Basic Course program for the Spanish course. For the Korean, Chinese Mandarin, and Standard Arabic courses, only 64-week Basic Course programs were selected. Because the administration of ODA at the end of the program was close to the final DLPT5 administration, there was not a concern that the ODA test results may be the result of lack of instruction. However, there was a concern that the DLPT5 administration could have occurred within a few days of completing the Basic Course to a few weeks or near to a date that was not consistently set within the same time frame for all test takers. Another issue of concern was that ODA data as well as DLPT5 data were archived data already available without the students knowing they were participants. Regarding ethical considerations, APA ethical guidelines indicate some research projects such as those that include anonymous surveys or questionnaires do not need informed consent from

participants. Secondary data such as student scores fit into this category, as long as the data are free from any identifying student information. To address ethical considerations, the data obtained by the researcher did not have any names associated with them. Ethical risk was minimized by making sure that the data were released to the researcher without any names or any other personal identifying information.

Summary

This chapter included a discussion on the methodology selected for this study. The population selected consisted of adult students taking the 36-week Spanish course or the 64-week Korean, Chinese Mandarin, or Standard Arabic Basic Course in a government setting. Archived data consisted of the ODA administered to students at the end of the program and their respective summative results of the DLPT5 administered at the end of the course. It was estimated that a minimum of 100 DLPT5 and ODA score matches per language could be obtained per language for Spanish, Korean, Chinese Mandarin, and Standard Arabic. According to the archived data, with the exception of Spanish, the actual scores available were fewer than the estimated minimum of 100 per language, as shown Table 3. Multiple regression analyses were performed using SPSS software. To address ethical considerations, the data obtained by the researcher did not have any names associated with them. Analysis of data and study results appear in Chapter IV.

CHAPTER IV: RESEARCH, DATA COLLECTION, AND FINDINGS

Overview

Identifying and building the foreign language expertise of military personnel has required DoD leaders to provide foreign language training, monetary incentives, and reliable standardized testing procedures to ensure the appropriate qualifications of military staff (Christensen, 2013). The DoD language-training program has also required raising the linguistic proficiency requirements for graduation. In 2017, the graduation criteria at the DLIFLC increased to the minimum achievement score of 2+ in listening and 2+ in reading on the summative DLPT5 (DLIFLC, 2015e, 2017). The efforts to meet the increased graduation standards require reliable assessment instruments such as the predictive DLAB and the summative DLPT5, which help in placement and estimating expected student outcomes at the end of a course program, respectively. These efforts also require descriptive diagnostic measures to know if a student is acquiring sufficient language skills during the course and is ready to meet higher language requirements with the help of assessment tools such as the ODA. Although researchers know about the DLAB and the DLPT5 through published research studies, they know little about the properties of the ODA as a formative diagnostic test through published correlation or validation studies. A review of the literature indicated a disconnect exists between theory and practice when looking at formative and summative assessments in a more integrated manner, and limited research addressed the correlation between formative assessments and summative assessments (Crooks, 2011; Croteau, 2014; Knight, 2000; Taras, 2005).

Chapter IV includes a detailed report of the findings of a multiple regression study to identify if a relationship exists between online formative (ODA) and summative

(DLPT5) assessments by examining the archived data obtained from DLIFLC for Spanish, Korean, Chinese Mandarin, and Standard Arabic.

Purpose Statement

The purpose of this nonexperimental correlational study was to identify the relationship between online formative (ODA) and summative (DLPT5) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success in a Basic Course program for adult students at the DLIFLC.

Research Questions

1. What is the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered at the end of the course and students' final summative DLPT5 scores?
2. What is the relationship between the ODA and the ILR levels for Spanish, Korean, Chinese Mandarin, and Standard Arabic as measured by the DLPT5?
3. Are the relationships found between ODA and DLPT5 for Spanish, Korean, Chinese Mandarin, and Standard Arabic consistent across the levels or is there variance in the relationship depending on the level?

The research questions were suitable for studying the predictability between the ODA scores and the DLPT5 scores with the goal to find out whether performance on the ODA correlated to the DLPT5 when the ODA is administered within 1 week to the last 3 months before the DLPT5 test administration. Because of the limited archived data, 1 week to 3 months served as the testing window for this study. This time frame was suitable because (a) although the Spanish course program is shorter (36 weeks), Spanish

archived data showed ODA test administrations predominantly closer to the DLPT5 administration (5 to 8 weeks) and (b) the Category IV languages studied had longer courses (64 weeks) for which fewer variables were expected to result from additional instruction. To ensure additional unforeseen variables, all ODA scores immediately after DLPT5 administration were considered invalid. This analysis was performed primarily through a Pearson product–moment correlation.

Research Methods and Data Collection Procedures

A nonexperimental design was chosen through a standard regression model to determine the relationships between two variables: (a) end-of-course ODA scores and (b) DLPT5 final scores. Several statistical analysis tests helped to identify correlations between ODA scores and DLPT5 final scores using multiple regression analysis. The data collection instruments used in this research study consisted of archived data from eight online diagnostic formative assessments (ODA) and eight summative DLPT5 assessments developed by DLIFLC:

- Archival scores for listening and reading from students who took the formative ODA at the end of the 36-week course in Spanish and archival scores of the same students who took the DLPT5 at the end of this program.
- Archival scores for listening and reading from students who participated in a formative ODA at the end of the 64-week course in Korean, Chinese Mandarin, and Standard Arabic and archival scores of the same students who took the summative DLPT5 at the end of this program.

Population

Research populations usually include of a number of individuals, cases, or elements that meet the requirements for a scientific study for which researchers want to make some generalizations. Because researchers may not be able to make generalizations of a whole population, they may rely on a specific sample or target population, known as the survey population or sampling frame (McMillan & Schumacher, 2010). Each calendar year, approximately 3,500 students attend the Basic Course programs available at the DLIFLC for 17 languages (DLIFLC, 2015c). For the languages studied, the total population in 2015 and 2016 at the Basic Course program consisted of 342 students for Spanish, 426 students for Korean, 571 students for Chinese Mandarin, and 912 students for Standard Arabic.

Sample

The individuals from a group or population about whom studies or assumptions are being made are usually described as a sample, which can be the whole population or a smaller group selected from this population (McMillan & Schumacher, 2010). When researchers cannot study the whole population, a target population is usually selected. The specific number of people from whom information can be obtained comprises the target population, also known as the sampling frame. The specific individuals selected from the sampling frame or target population comprise the sample. The larger the sample is, the higher is the confidence of a close approximation to the results that the researcher can obtain from a sampling frame or target population (Creswell, 2012). Onwuegbuzie (2003) recommended using statistical power analysis to determine the sample size in quantitative correlation studies with a minimum of 64 participants for one-tailed

hypotheses and 82 participants for two-tailed hypotheses. For this reason, the researcher sought a minimum sample size of 100 participants per language for a total of 400 participants from archival DLPT5 and ODA scores.

With approval from the Office of the Commandant and the DLIFLC provost, DLPT5 archived score information from 800 students was matched to the corresponding ODA scores from each student by DCSIT. Student information was replaced with an ID code. DCSIT delivered an Excel document via a secure site with DLPT5 score information from 200 test takers per language (800 students total) and any possible ODA score matches.

Out of the 800 DLPT5 scores consisting of 200 scores per language for listening and reading, it was estimated that a minimum of 100 DLPT5 and ODA score matches would be obtained per language for Spanish, Korean, Chinese Mandarin, and Standard Arabic. The original assumption was that instructors from language schools would administer the ODA consistently at the beginning and at the end of the school program. Through the data collection, the researcher found that although instructors from all schools administered the ODA, not all of the ODA is administered consistently at the end of the school program. Therefore, not all data available fit the requirements for this research. As seen in Table 10, with the exception of the Spanish school, a considerable portion of the data available for Korean, Chinese Mandarin, and Standard Arabic schools could not be used because in 2015 and 2016, these schools administered the ODA at time frames that were outside of the window of 1 week to 3 months from the DLPT5 test administration. Table 9 shows the DLPT5 and ODA data available for this study, out of

which only the data between 1 week and 3 months from the DLPT5 test administration were used.

Table 9

DLPT5 and ODA Data Available

DLPT5 and ODA score matches available	Spanish	Korean	Chinese	Standard Arabic
1 week to 3 months of DLPT5 administration	116L/118R	35L/39R	65L/66R	53L/47R
1 week to 4 months of DLPT5 administration	119L/122R	59L/62R	91L/86R	59L/57R
1 week to 5 months of DLPT5 administration	121L/122R	70L/70R	97L/91R	96L/84R
Total score matches available (~1 week to after 5 months of DLPT5 administration)	172	161	152	169

Note. L = listening. R= reading.

Figures 13 and 14 show the DLPT5 and ODA archived data provided. The 3-month spread bar represents the data that met the requirements for this study. All other data did not meet the criteria for this study because the ODA was administered outside of the testing window of 1 week to 3 months from the DLPT5 administration.

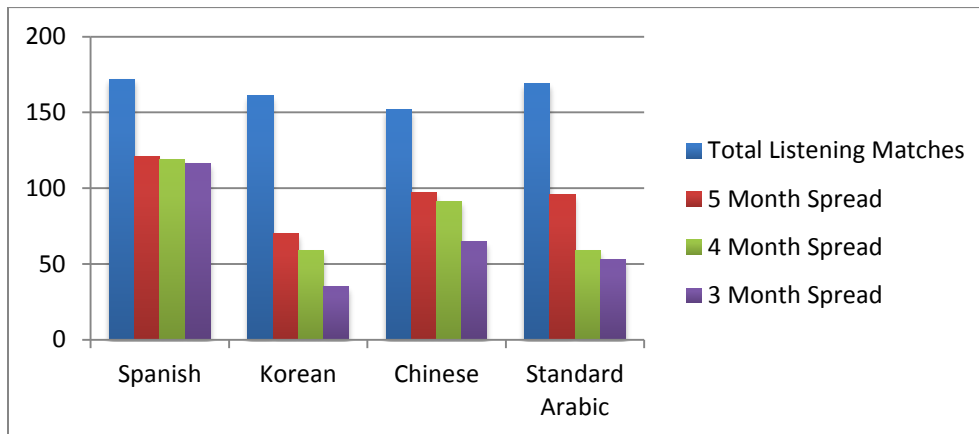


Figure 13. DLPT5 and ODA data pool score matches for listening.

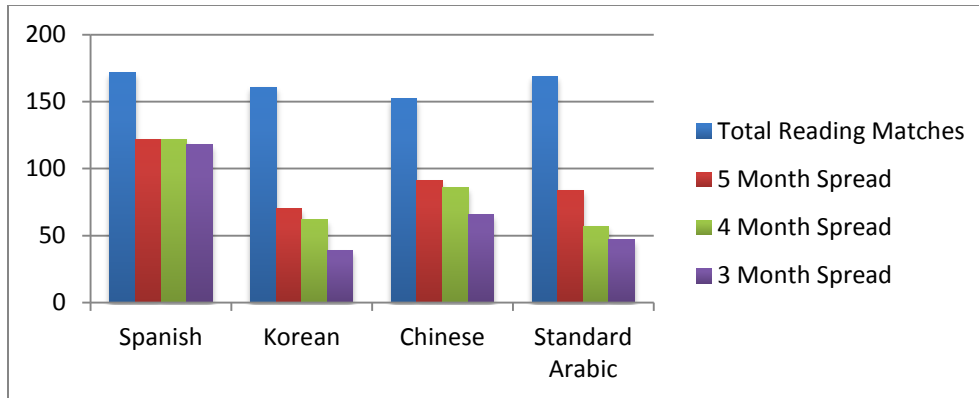


Figure 14. DLPT5 and ODA data pool score matches for reading.

Based on these selection criteria, 269 listening archived scores and 270 reading archived scores from 276 students for four languages were selected, which represented 7.7% of the total population in 1 year. They also represented 35% of the total Spanish school population, 8% of the total Korean school population, 12% of the total Chinese Mandarin school population, and 6% of the total Standard Arabic school population in 2015 and 2016. Figure 15 shows the percent of the sample population for each language studied, and Table 10 shows the archived scores available for this study compared to the total population per language school in 2016 and 2015.

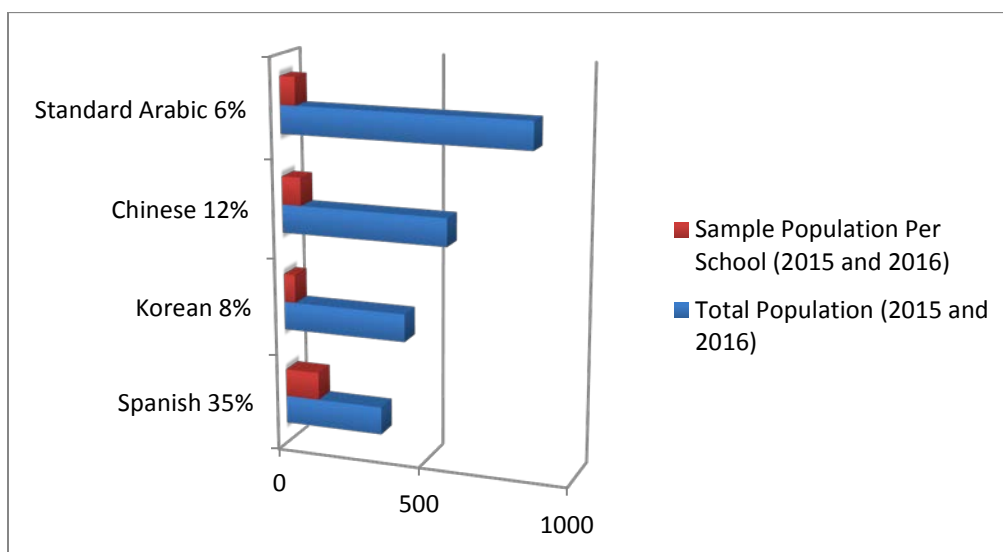


Figure 15. Student sample per language.

Table 10

DLPT5 and ODA Archived Scores Used for Study

DLPT5 and ODA score matches available	Spanish	Korean	Chinese	Standard Arabic
1 week to 3 months of DLPT5 administration	116L/118R	35L/39R	65L/66R	53L/47R
Breakdown by school				
Number of students	118	39	66	53
Population per school/year (2016)	184	211	313	419
Population per school/year (2015)	158	215	258	433
Population (2015 + 2016)	342	426	571	912

Note. L = listening. R= reading.

Except for the Spanish sample, Onwuegbuzie’s (2003) formula for sample sizing could not be implemented due to the actual sample size per archived data obtained.

Therefore, Krejcie and Morgan’s (1970) formula for determining research sample sizes was used instead. This formula helped identify the level of confidence and margin of error of the study based on the total population and sample size for each language:

$$n = \frac{\chi^2 \times N \times P \times (1 - P)}{(ME^2 \times (N - 1) + (\chi^2 \times P \times (1 - P)))}$$

where n = sample size, χ^2 = chi square for the specified confidence level at 1 degree of freedom, N = population size, P = population proportion, and ME = desired margin of error.

According to Krejcie and Morgan’s formula, the following was determined:

1. Spanish sample size of 118 students = 82% level of confidence with a .05 margin of error.
2. Korean sample size of 39 students = 49% level of confidence with a .05 margin of error.
3. Chinese Mandarin sample size of 66 students = 61% level of confidence with

- a .05 margin of error.
4. Standard Arabic sample size of 53 students = 54% level of confidence with a .05 margin of error.

Demographic Data

Specific demographic data were not available from the archived data provided. The general demographic population consisted of students from all military branches in the United States. The student population had a variety of academic backgrounds and came from all parts of the United States. These students were assigned to a specific language school based on their score results on the DLAB placement test and the military's needs. Most students started as nonnative speakers. Although some students started the assigned language program with some second language acquisition, archived data did not have this information available. Gender and ethnic background information was also unavailable.

Presentation and Analysis of Data

On March 15, 2016, the researcher received Scientific Review Board approval from DLIFLC, along with a letter from the Office of the Commandant expressing willingness to grant permission to use archived data for dissertation research, contingent upon Brandman University's IRB review and DLIFLC's administrative review. On August 18, 2016, IRB approval was received from Brandman University, Chapman University System. On March 10, 2017, the researcher received an official, securely delivered set of Excel files from DCSIT with the ODA and DLPT5 student matches for this research. Data available for this research are indicated in Table 11.

Table 11

Data Available From Second Data Pull

Language	DLPT5 minimum number of archived scores requested	ODA score matches found (second data pull)	DLPT5 and ODA score matches meeting correlation requirements
Spanish	200	166	116L/118R
Korean	200	174	35L/39R
Chinese Mandarin	200	179	65L/66R
Standard Arabic	200	179	53/47R

Note. L = listening. R= reading.

ODA and DLPT5 data from the same students were provided to the researcher in an Excel spreadsheet with an ID code replacing student names to ensure student confidentiality. Information was provided showing the ODA and the DLPT5 scores of the same ID code representing a student as shown in Figure 16. Additional columns were included in the Excel spreadsheet showing additional ODA scores resulting from additional test administrations along with the specific date of ODA test administration.

Language KOREAN	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date
Language CHINESE	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date
Language STANDARD ARABIC	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date
Language STANDARD ARABIC	Student ID Code	DLPT5 L Score Level	DLPT5 L Testing Date	DLPT5 R Score Level	DLPT5 R Testing Date	ODA L Score Level	ODA L Testing Date	ODA R Score Level	ODA R Testing Date

Figure 16. Excel spreadsheet data columns.

All students who took the ODA at the end of the course were included in the study. Data were analyzed to remove the records of subjects whose ODA scores were outside of the testing window of 1 week to 3 months of administration before the DLPT5;

thus, only scores between 7 days to a minimum of 3 months from the test administration were included, as indicated in Table 12.

Table 12

Student Sample

DLPT5 and ODA score matches available	Spanish	Korean	Chinese	Standard Arabic
1 week to 3 months of DLPT5 administration	116L/118R	35L/39R	65L/66R	53L/47R
Sample breakdown by school				
Number of students	118	39	66	53
Population per school/year (2016)	184	211	313	419
Population per school/year (2015)	158	215	258	433
Population (2015 + 2016)	342	426	571	912

Note. L = listening. R= reading.

The process for analyzing data began after the Excel files were cleaned and spreadsheets were separated by language, school, and content area. ODA test administrations that met the testing window requirements of the correlation study were identified. The archived data provided contained a different score information nomenclature for the DLPT5 and the ODA score information, as shown in Table 13.

Table 13

DLPT5 and ODA Score Nomenclature

DLPT5 score	ODA score
6	-1
10	1
16	1+
20	2
26	2+
30	3

Score information for the ODA was reclassified to match the same score nomenclature shown in the DLPT5 for the purposes of cleaning the data. After this

process, an ODA coding system was created to convert the DLPT5 and ODA scores into a code that will eventually be transferred into an Excel database and analytical software for SPSS data. Table 14 shows the DLPT5 coding system used for this correlation study.

Table 14

DLPT5 and ODA Coding System

Current scores	Corresponding codes
6 (-1)	0
10 (1)	1
16 (1+)	2
20 (2)	3
26 (2+)	4
30 (3)	5

Excel spreadsheets were employed to convert score information into the coding system. After the coding system was completed, data were imported into an SPSS database that analyzed data using a multiple regression analysis. ODA scores were the dependent variable. DLPT5 listening and reading scores represented the independent variable.

To determine the correlation between ODA scores and DLPT5 scores, a Pearson product–moment correlation (r) was calculated between the average of the ODA scores and the DLPT5 scores separately for each language and content area (listening and reading). After data were analyzed, the correlation coefficient for each language and content area were identified using the Pearson’s correlation standard values in Table 15.

Table 15

Correlation Coefficient Values

Correlation coefficient	Strength of the relationship
± .70 to 1.00	Strong
± .32 to .69	Moderate
± .00 to .29	None (.00) to weak

Detailed Analyses: Results for Research Questions

Research Question 1. Research Question 1 was as follows: What is the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered at the end of the course and students' final summative DLPT5 scores? To determine the correlation between ODA scores and DLPT5 scores, a Pearson product–moment correlation (r) was calculated between the average of the ODA scores and the DLPT5 scores separately for each language and content area (listening and reading). For the listening content area, an r value of .32 for Spanish, .40 for Korean, and of .56 for Standard Arabic indicated a moderate correlation of the ODA listening tests to the DLPT5 for these languages. The Standard Arabic ODA listening test indicated the highest level of correlation to the DLPT5 with an r value of .56. The Chinese Mandarin ODA listening test had an r value of .20, which indicated the weakest correlation to the DLPT5 from the four languages studied. In the case of the reading content area, the Chinese Mandarin ODA had an r value of .34, and the Standard Arabic ODA indicated an r value of .30, which indicated a moderate correlation to the DLPT5. The Korean reading ODA had an r value of .23, which indicated a weak correlation. The Spanish ODA for reading indicated the weakest correlation with an r value of .14. Tables 16 and 17 show the correlation results per language for listening and for reading.

Table 16

Correlation per Language for Listening

Listening	Correlation	Strength of the relationship
Spanish	.32	Moderate
Korean	.40	Moderate
Chinese Mandarin	.20	Weak
Standard Arabic	.56	Moderate

Table 17

Correlation per Language for Reading

Reading	Correlation	Strength of the relationship
Spanish	.14	Weak
Korean	.23	Weak
Chinese Mandarin	.34	Moderate
Standard Arabic	.30	Moderate

Research Question 2. Research Question 2 was as follows: What is the relationship between the ODA and the ILR levels for Spanish, Korean, Chinese Mandarin, and Standard Arabic as measured by the DLPT5? To answer this question, all the ODA scores selected for this study were classified by their ILR student scores for each content area (listening and reading) for each of the four languages. Excel spreadsheets were organized by the ILR levels per the DLPT5 with a spreadsheet for each ILR level. Next to the ILR level per the DLPT5, an additional column contained the score that the same student scored on the ODA. After these spreadsheets were created, additional Excel spreadsheets were generated to tally the data to identify the number of ODA scores for each ILR level per the DLPT5. Columns in this Excel spreadsheet included the number of students scoring at each ILR level per the DLPT5, along with the total number of scores at the same level per the ODA. To tally the students who scored at other ILR levels, additional columns were added, as shown in Table 18.

Table 18

Organization of ODA Scores per ILR Level

Language/ content area	DLPT5 scores at ILR target level per DLPT5	ODA scores at ILR target level per DLPT5	ODA scores three levels higher than ILR level per DLPT5	ODA scores two levels higher than ILR level per DLPT5	ODA scores one level higher than ILR level per DLPT5	ODA scores one level lower than ILR level per DLPT5	ODA scores two levels lower than ILR level per DLPT5	ODA scores three levels lower than ILR level per DLPT5
ILR Level 3								
ILR Level 2+								
ILR Level 2								
ILR Level 1+								
ILR Level 1								
ILR Level 0+								

After data were classified, the total number of ODA scores for each ILR level was compared against the total DLPT5 scores for each DLPT5 level for the only purpose of identifying general trends. These data provided information about the relation between ODA and DLPT5 scores at a global level, which helped identify general score distribution trends. The results of this general score distribution comparison appear in Figures 17-24.

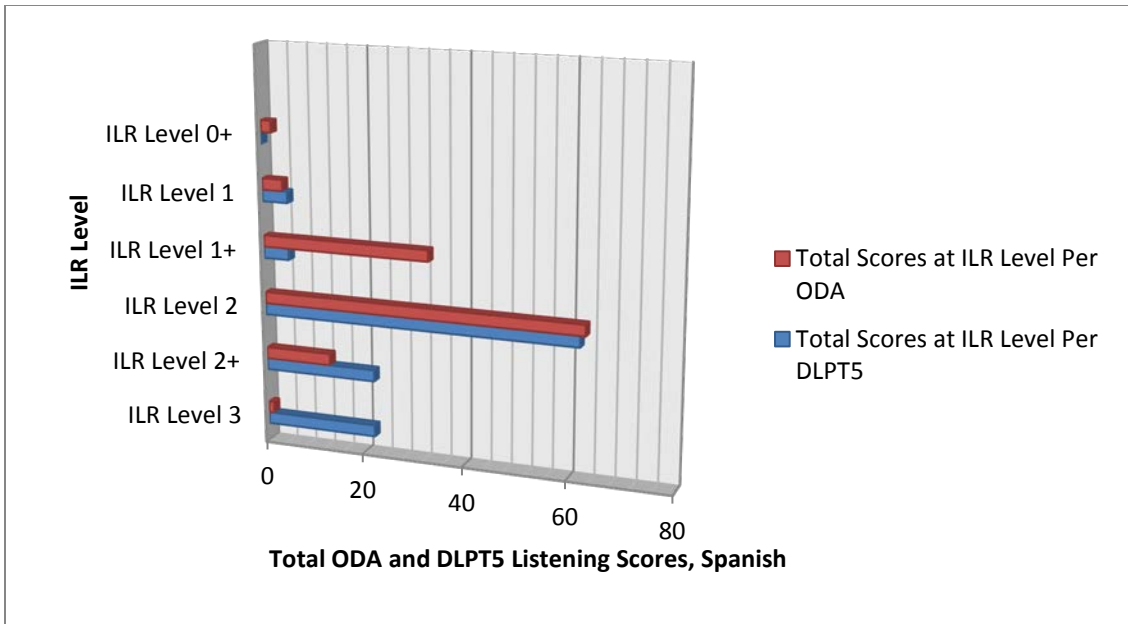


Figure 17. Total Spanish ODA and DLPT5 score comparison—listening.

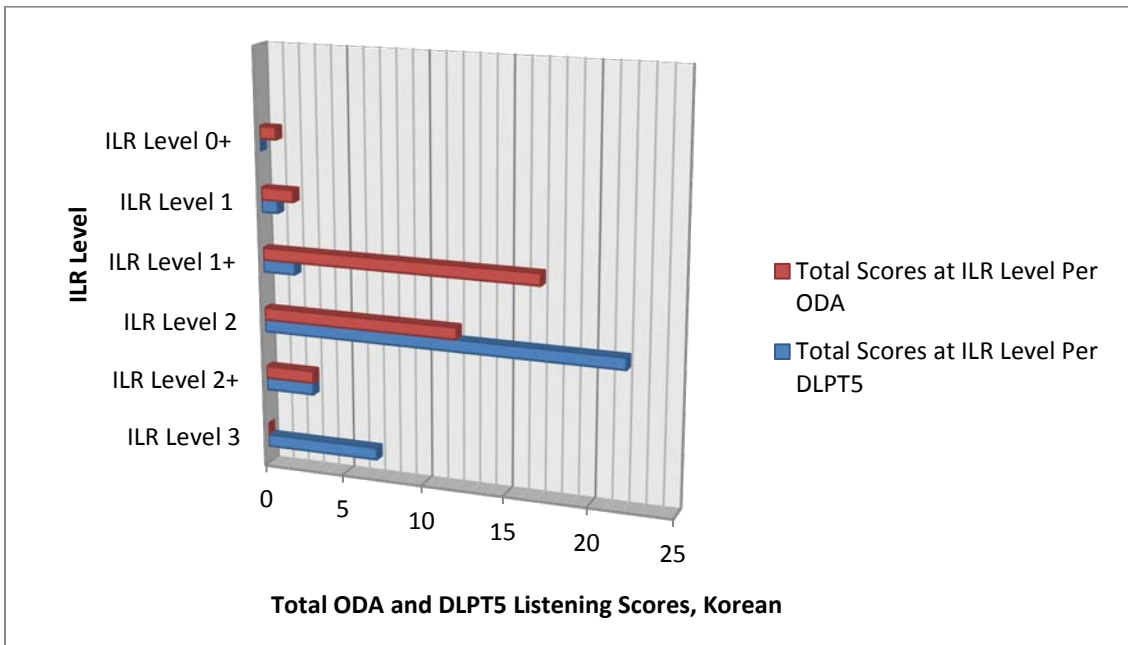


Figure 18. Total Korean ODA and DLPT5 score comparison—listening.

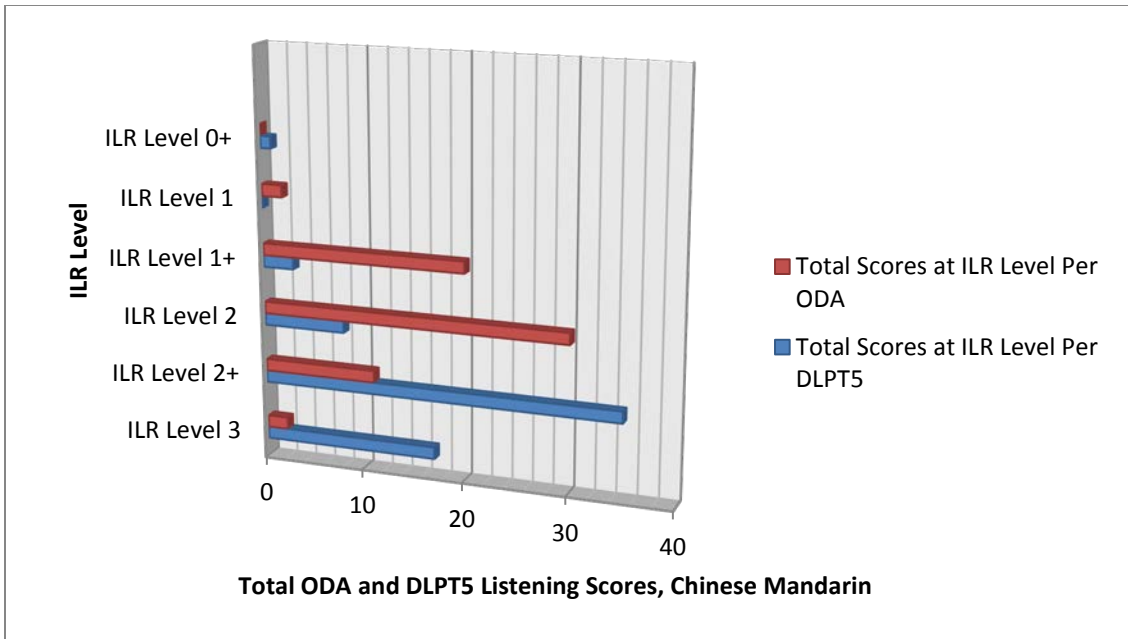


Figure 19. Total Chinese Mandarin ODA and DLPT5 score comparison—listening.

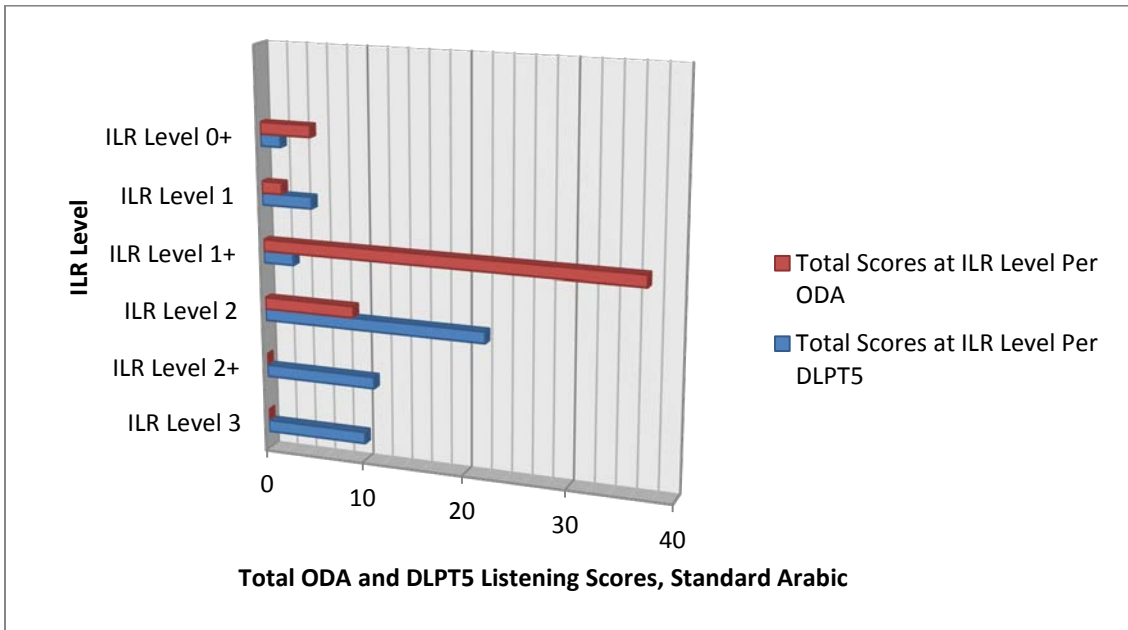


Figure 20. Total Standard Arabic ODA and DLPT5 score comparison—listening.

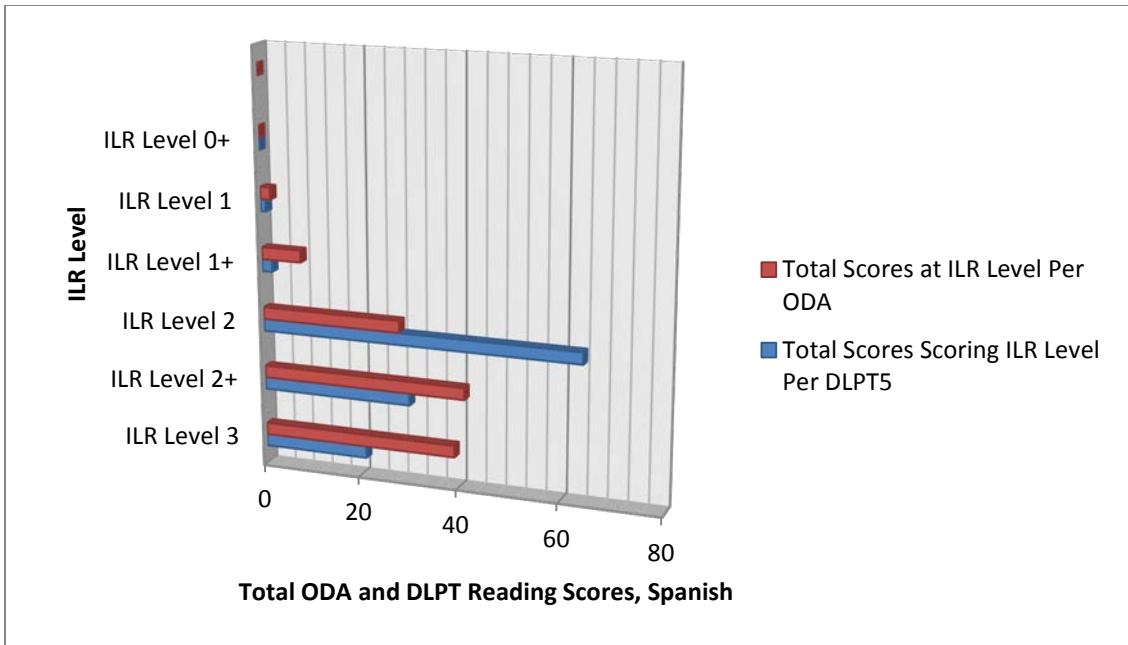


Figure 21. Total Spanish ODA and DLPT5 score comparison—reading.

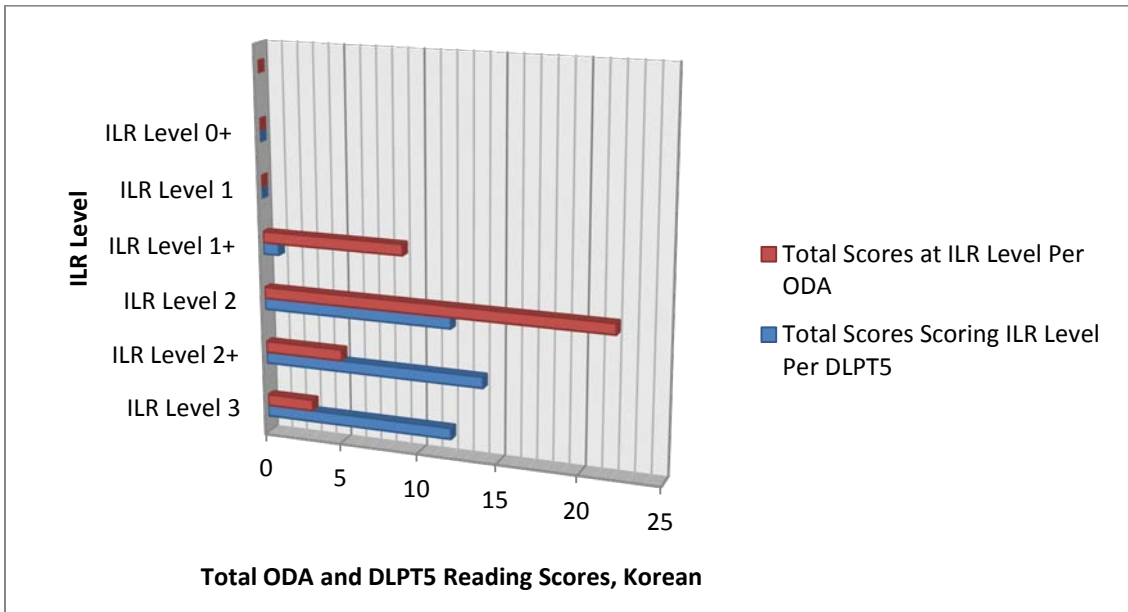


Figure 22. Total Korean ODA and DLPT5 score comparison—reading.

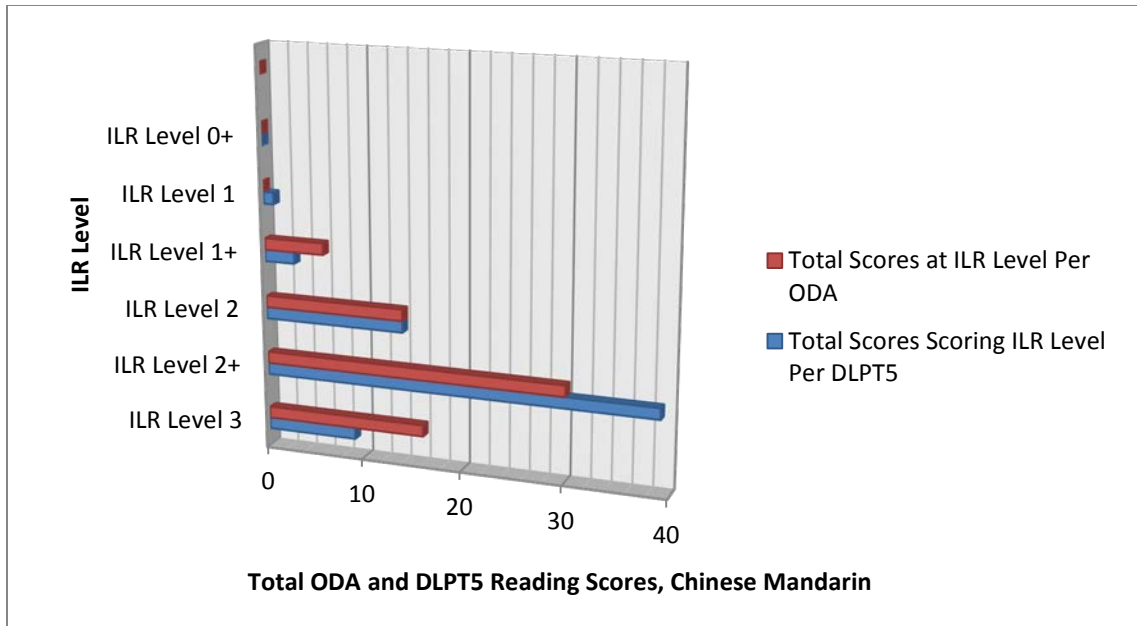


Figure 23. Total Chinese Mandarin ODA and DLPT5 score comparison—reading.

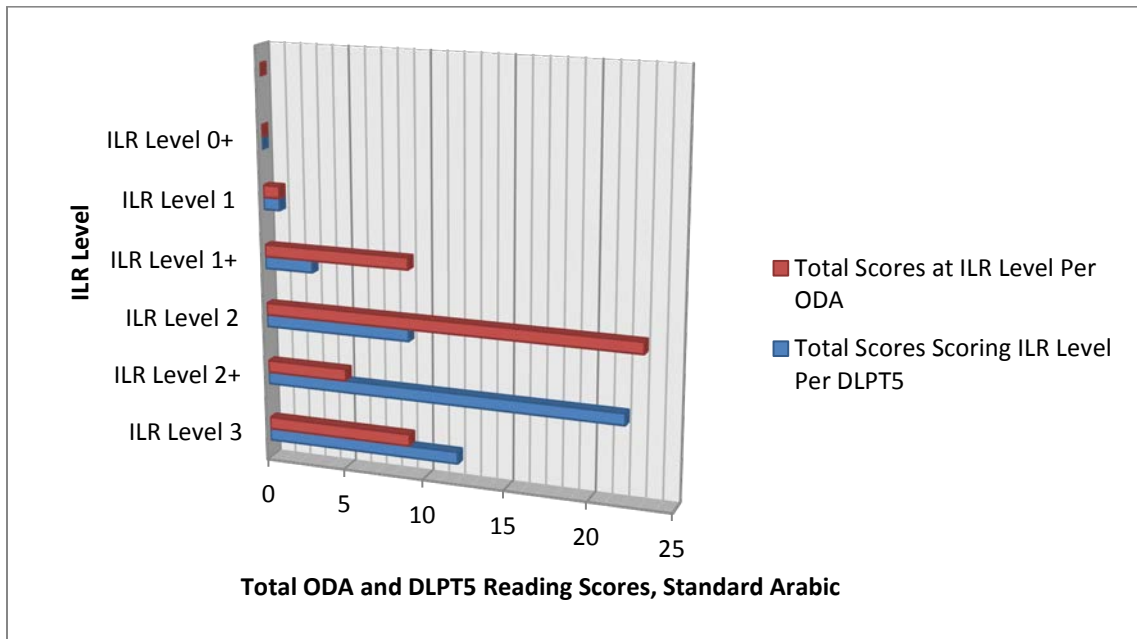


Figure 24. Total Standard Arabic ODA and DLPT5 score comparison—reading

ODA Excel score spreadsheets containing a score distribution for each ILR level per the DLPT5 were completed, along with percentage distribution scores. Column charts and line charts eased the comparison of ODA ILR scores against DLPT5 ILR scores to

identify areas where the ILR ODA levels might more closely align with the DLPT5. Data were also compared to the regression analysis to verify the consistency of the comparison analysis of the ODA ILR results per DLPT5 ILR scores. With the Pearson product–moment correlation serving as the primary criterion, and Excel score distribution as a reference, additional criteria were set to determine the relationship for each ILR level per DLPT5. Strong relationships for ODA scores were predominantly at the target level per DLPT5, with no scores at other ILR levels. Moderate relationships for ODA scores were predominantly at the target level or at one ILR level higher or one ILR level lower than the target ILR level per DLPT5. Weak relationships for ODA scores were predominantly two levels higher or lower than the ILR level per DLPT5 or, for scores with a wide variety of ILR scores, ranges included predominant scores two levels lower or two levels higher than the ILR levels per DLPT5.

Relationship between the ODA and the ILR for listening. For listening, the Pearson product–moment correlation showed a weak correlation for Chinese Mandarin (r value of .20), a moderate correlation for Spanish (r value of .32) a moderate correlation for Korean (r value of .40), and a moderate correlation for Standard Arabic (r value of .56). The correlation did not indicate a strong correlation for any of the languages studied, which requires an r value of .70 to 1.00 to be considered strong. In this context, ILR scores for listening indicated tendencies for scoring at certain ILR levels, with some languages showing higher levels of alignment to the ILR levels than others per the DLPT5. The ODA data distribution per ILR levels 3 and 2+ indicated the weakest relationship to the ILR at Level 3 and Level 2+ for all languages Data also showed that all languages, although with Chinese Mandarin to a lesser extent, had the closest

(moderate) relationship to the ILR at the ILR Level 2 per DLPT5 for listening. Table 19 shows the ODA listening relationship to the ILR levels according to the DLPT5.

Table 19

ODA Listening Relationship to the ILR Levels per DLPT5

Listening	Spanish Moderate (<i>r</i>)	Korean Moderate (<i>r</i>)	Chinese Mandarin Weak (<i>r</i>)	Standard Arabic Moderate (<i>r</i>)
ILR Level 3	Weak	Weak	Weak	Weak
ILR Level 2+	Weak	Weak ^a	Weak	Weak
ILR Level 2	Moderate	Moderate	Moderate to weak	Moderate
ILR Level 1+	Moderate	Strong ^a	Weak ^a	Moderate to weak ^a
ILR Level 1	Moderate	Moderate ^a	N/A	Weak
ILR Level 0+	N/A	N/A	N/A	Strong ^a

^aNot enough scores to identify clear ILR relationship trends.

ODA relationship to the ILR for listening Spanish. For listening, the Pearson product–moment correlation indicated a moderate correlation for Spanish (*r* value of .32). As shown in the score distribution, based on the Spanish sample obtained, the majority of students scored at an ILR level of 2 on the DLPT5, with most of the students who took the ODA scoring at an ILR level of 2 on the ODA as well. The closest ODA alignment to the ILR levels was at Level 2 per the DLPT5, as the percentage of students scoring at the ILR 2 level on the ODA was 61%, with 3% of students scoring one level higher than the ILR level, and 29% scoring one level lower than the ILR level per the DLPT5. At Level 3, data indicated that students who took the ODA tended to score one to two levels lower than the ILR, which indicated a weak alignment. Specifically, 4.5% of students scored at the target level, 41% of students scored one level lower than the ILR level, and 41% of students scored two levels lower than the ILR level per DLPT5. At Level 2+, students scored one to two levels lower than the ILR level, which indicated a weak alignment to the ILR level per DLPT5. Specifically, 9.1% of

students scored at the target level, 59% of students scored one level lower, and 27.4% of students scored two levels lower than the ILR level per the DLPT5. ODA scores at the ILR 1+ level indicated a tendency for students to score one level higher than the ILR level to the target level per the DLPT5, which indicated a moderate alignment. Forty percent of students scored at the target level, and 60% of students scored one level higher than the ILR level per DLPT5. ODA scores at the ILR level of 1 indicated a tendency for students to score one level higher than the ILR level, which indicated a moderate alignment. Specifically, 80% of students scored one level higher and 20% of students scored two levels higher than the ILR level per DLPT5. According to Krejcie and Morgan’s (1970) formula for student sampling, the level of confidence for correlation results is 82% with a .05 margin of error. Figure 25 shows the ODA relationship to the ILR for listening Spanish according to the DLPT5, and Figure 26 shows the ILR percentage distribution for the ODA according to the DLPT5 for listening Spanish.

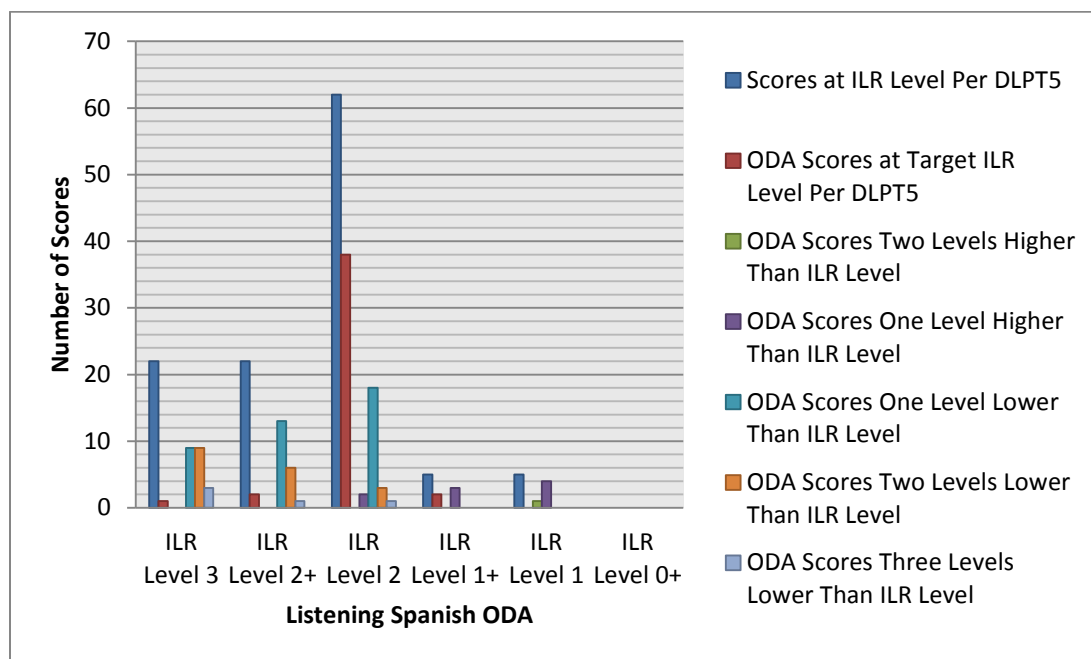


Figure 25. ODA relationship to the ILR—Listening Spanish.

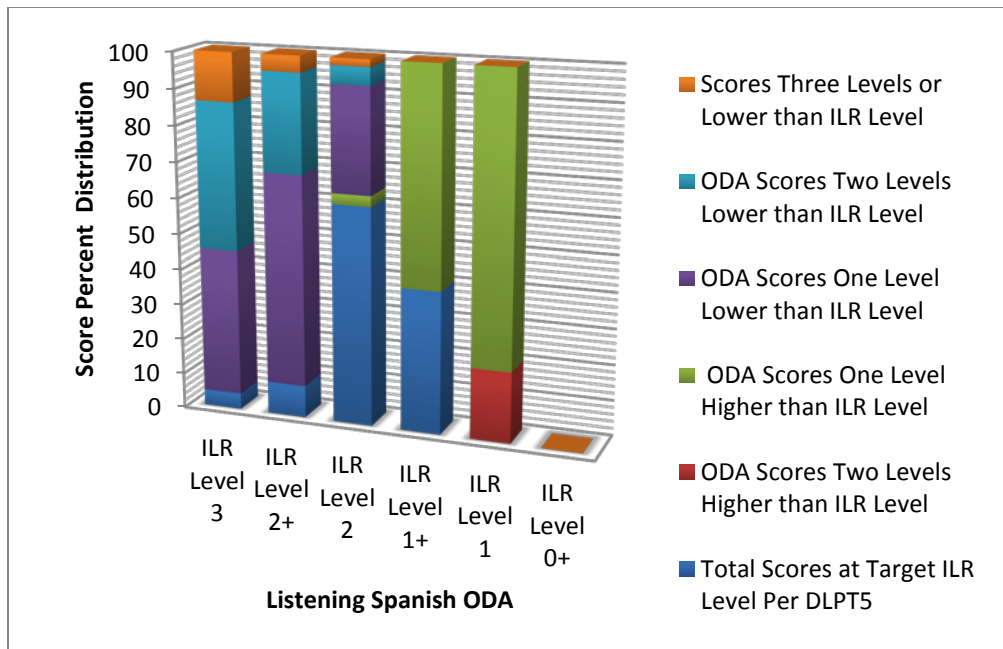


Figure 26. ILR percentage distribution per DLPT5—Listening ODA Spanish.

Relationship between the ODA and the ILR for listening Korean. For listening, the Pearson product–moment correlation indicated a moderate correlation for Korean (r value of .40). Per score distribution, based on the Korean sample obtained, the majority of students scored at an ILR level of 2 on the DLPT5, and the majority of students who took the ODA scored at one level lower, followed by the target level indicating a moderate relationship to the ILR at Level 2. Specifically, 27% of students who took the ODA scored at the ILR target level, 50% of students scored one ILR level lower, 9% scored two ILR levels lower, and 9% scored one level higher than the ILR level per the DLPT5. From the students who scored 2+ or 3 on the DLPT5, data indicated that these students scored two ILR levels lower in the ODA, which indicated a weak alignment to the ILR. Specifically, for Level 3, 86% of students scored two levels lower, and 14% scored one level lower than the ILR level per DLPT5. For Level 2+, 67% of students scored two levels lower than the ILR level per DLPT5, and 33% scored one level lower. At the ILR

Level 1+, there were very few scores to identify clear trends. Data indicated a strong alignment at the target ILR level, as 100% of students scored at the target ILR level per DLPT5. At Level 1, data showed scores one level higher than the DLPT5, which indicated a moderate alignment, although there were insufficient scores at this level to identify clear trends. According to Krejcie and Morgan’s (1970) formula for student sampling, the level of confidence for correlation results is 49% with a .05 margin of error. Figure 27 shows the ODA relationship to the ILR for listening Korean according to the DLPT5, and Figure 28 shows the ILR percentage distribution for the ODA according to the DLPT5 for listening Korean.

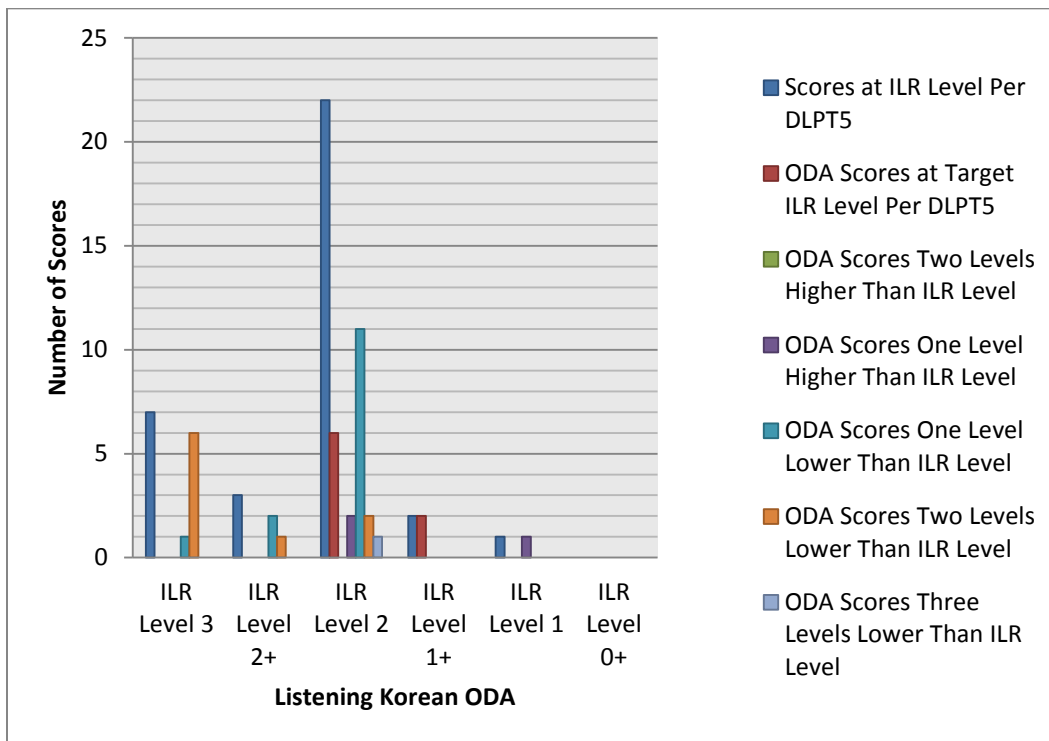


Figure 27. ODA relationship to the ILR— Listening Korean.

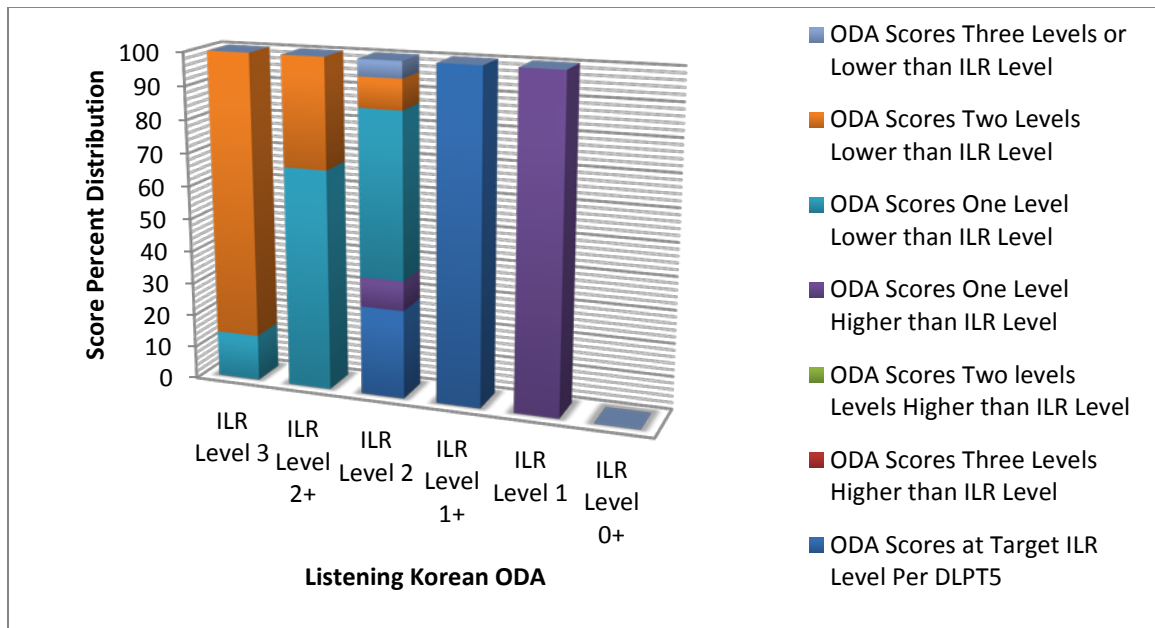


Figure 28. ILR percentage distribution per DLPT5—Listening ODA Korean.

Relationship between the ODA and the ILR for listening Chinese Mandarin. For listening, the Pearson product–moment correlation identified a weak correlation for Chinese Mandarin (r value of .20). Per score distribution, based on the Chinese Mandarin sample obtained, the majority of students scored at an ILR level of 2+ on the DLPT5, and the majority of students who took the ODA scored one or two levels lower per the DLPT5, which indicated a weak relationship to the ILR per DLPT5 at Level 2+. Specifically, 7% of students who took the ODA scored at the 2+ ILR target level, whereas 42% of students scored one level lower and 39% scored two levels lower than the ILR per DLPT5. For Level 3, the majority of students scored two levels lower than the ILR, which indicated a weak relationship to the ILR per DLPT5. Specifically, 18% of students scored one level lower and 59% of students scored two levels lower than the ILR level. The ODA showed the closest relationship to the ILR at Level 2. At this level, the ODA also showed a wide range of scores at other ILR levels, which indicated a moderate to weak relationship to the ILR per DLPT5. Specifically, 50% of students scored at the

ILR level, 25% scored one level lower, 12.5% scored two levels lower, and 12.5% scored two levels higher than the ILR level per the DLPT5. There were no scores at ILR Level 1 per DLPT5, and only one score at Level 0+, with a score two levels higher than the ILR level per the DLPT5, which indicated a possible test-taking irregularity at Level 0+.

According to Krejcie and Morgan’s (1970) formula for student sampling, the level of confidence for correlation results is 61% with a .05 margin of error. Figure 29 shows the ODA relationship to the ILR for listening Chinese Mandarin according to the DLPT5, and Figure 30 shows the ILR percentage distribution for the ODA according to the DLPT5 for listening Chinese Mandarin.

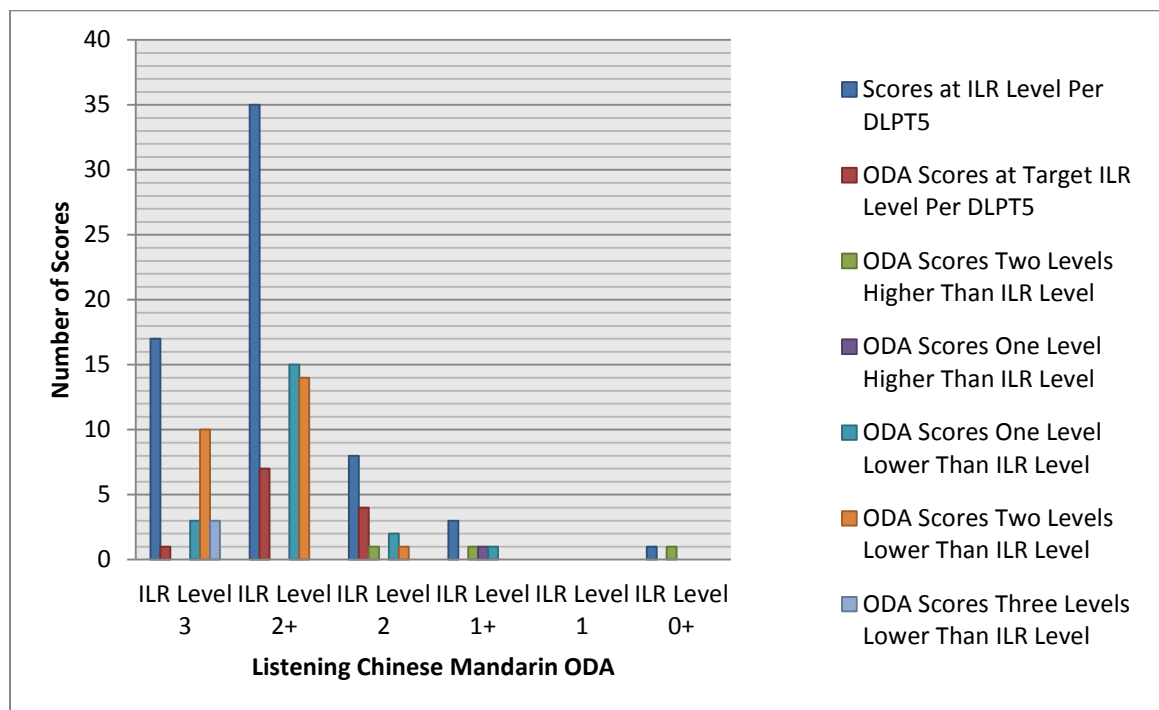


Figure 29. ODA relationship to the ILR—Listening Chinese Mandarin.

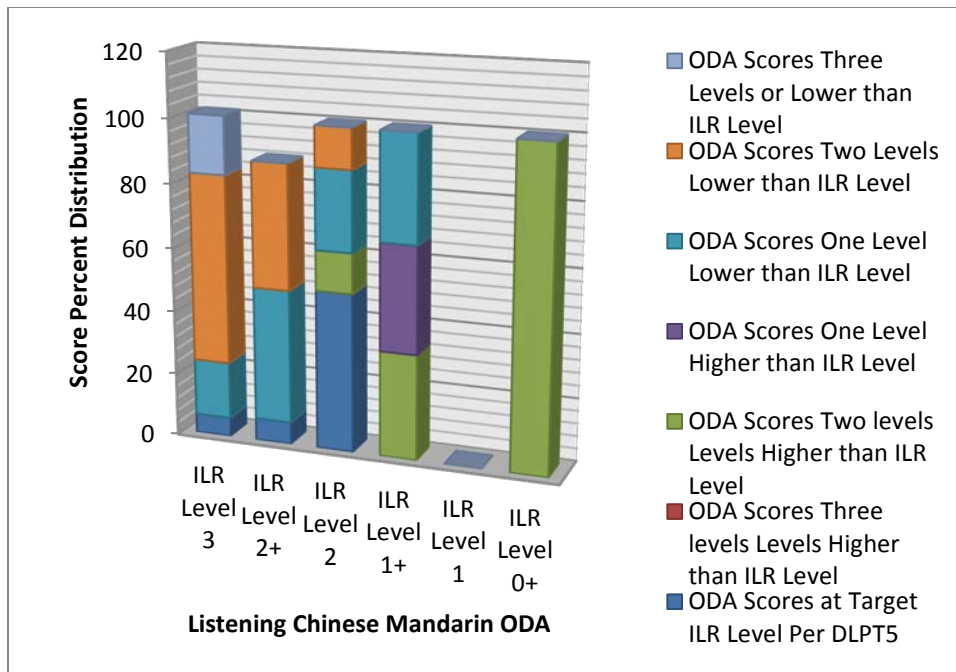


Figure 30. ILR percentage distribution per DLPT5—Listening ODA Chinese Mandarin.

Relationship between the ODA and the ILR for Listening Standard Arabic. For listening, the Pearson product–moment correlation identified a moderate correlation for Standard Arabic (r value of .56). Per score distribution, based on the Standard Arabic sample obtained, the majority of students scored at an ILR level of 2 on the DLPT5, and most of the students who took the ODA scored one level lower. This score distribution indicated a moderate relationship to the ILR for Level 2 per the DLPT5. Specifically, 10% of the students who took the ODA scored at the ILR target level, and 85% of the students scored one level lower. For ILR Level 3, 60% of students scored two levels lower and 40% of students scored three levels lower than the ILR level, which indicated a weak relationship to the ILR per DLPT5. For ILR Level 2+, 91% of students scored two levels lower and 9% of students scored three levels lower than the ILR level, which indicated a weak relationship to ILR per DLPT5. For Level 1+, 1, and 0+, there were only a handful of scores to identify clear trends. From the scores available at Level 1+,

67% of students scored at the ILR target level, and 33% of students scored two levels lower than the ILR level, which indicated a moderate to weak alignment. For Level 1, 20% of the ODA scores were at the target level, 20% of scores were one level lower, 40% of scores were one level higher than the ILR, and 20% of scores were two levels higher than the ILR level per the DLPT5, which indicated a weak alignment at this level. For Level 0+, although there were only a handful of scores, the ODA showed a strong alignment to the ILR, with 100% of scores at the ILR level per the DLPT5. According to Krejcie and Morgan’s (1970) formula for student sampling, the level of confidence for correlation results is 54% with a .05 margin of error. Figure 31 shows the ODA relationship to the ILR for listening Standard Arabic according to the DLPT5, and Figure 32 shows the ILR percentage distribution for the ODA according to the DLPT5 for listening Standard Arabic.

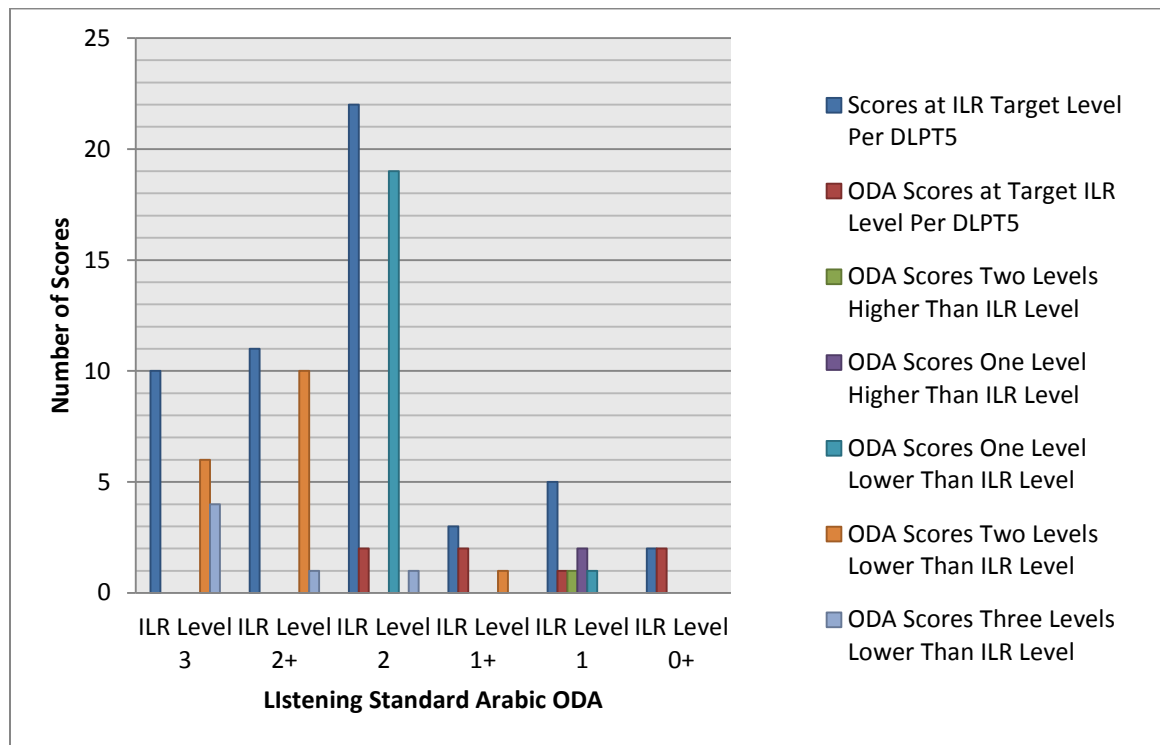


Figure 31. ODA relationship to the ILR—Listening Standard Arabic.

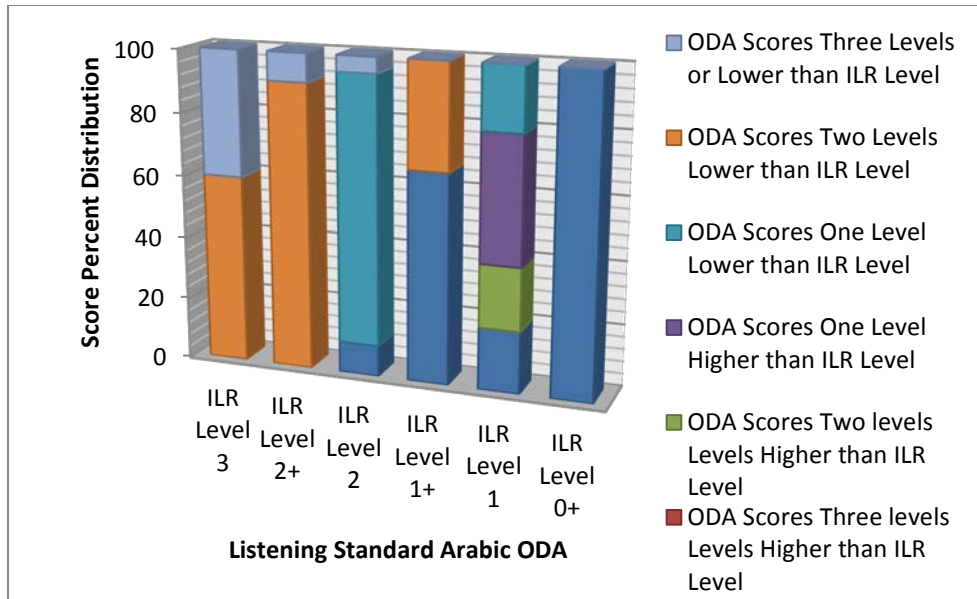


Figure 32. ILR percentage distribution per DLPT5—Listening ODA Standard Arabic.

Relationship between the ODA and the ILR for reading. For reading, the Pearson product–moment correlation indicated a weak correlation for Spanish (r value of .14), a weak correlation for Korean (r value of .23), a moderate correlation for Standard Arabic (r value of .30), and a moderate correlation for Chinese Mandarin (r value of .34). The correlation did not indicate a strong correlation for any of the languages studied, which requires an r value of .70 to 1.00 to be considered strong. For reading, data indicated a different relationship to the ILR depending on the language and depending on the level. Data also indicated that all languages, with the exception of Spanish, had the closest (moderate) relationship to the ILR at the ILR Level 2 per DLPT5 for reading. The ODA score distribution indicated a weak relationship to the ILR at Level 3 for Korean and Standard Arabic, a moderate to weak relationship for Spanish, and a moderate relationship for Chinese Mandarin per the DLPT5. At Level 2+, the ODA score distribution indicated a moderate relationship for Spanish and Chinese Mandarin, a moderate to weak relationship for Standard Arabic, and a weak relationship for Korean.

At Level 2, the ODA score distribution indicated a moderate relationship for Korean, Chinese Mandarin, and Standard Arabic and a weak relationship for Spanish. Although there were few scores available at Levels 1+ and below, at ILR Level 1+, the ODA data indicated a moderate relationship to the ILR for Korean and Standard Arabic and a weak relationship for Chinese Mandarin and Spanish. The sparse scores at ILR Level 1 indicated a moderate relationship to the ILR for Korean and a weak relationship for Chinese Mandarin. More data may be necessary to identify clearer trends. Table 20 shows the ODA relationship to the ILR levels according to the DLPT5.

Table 20

ODA Reading Relationship to the ILR Levels Per DLPT5

Reading	Spanish Weak (<i>r</i>)	Korean Weak (<i>r</i>)	Chinese Mandarin Moderate (<i>r</i>)	Standard Arabic Moderate (<i>r</i>)
ILR Level 3	Moderate to weak	Weak	Moderate	Weak
ILR Level 2+	Moderate	Weak	Moderate	Moderate to weak
ILR Level 2	Weak	Moderate	Moderate	Moderate
ILR Level 1+	Weak ^a	Moderate ^a	Weak ^a	Moderate ^a
ILR Level 1	N/A	N/A	Weak ^a	Moderate ^a
ILR Level 0+	N/A	N/A	N/A	N/A

^aNot enough scores to identify clear ILR relationship trends.

ODA relationship to the ILR—Reading Spanish. For reading, the Pearson product–moment correlation identified a weak correlation for Spanish (*r* value of .14) and the lowest correlation of all languages studied. Per score distribution, based on the Spanish sample obtained, the majority of students scored at an ILR level of 2 on the DLPT5, and the majority of students scored at a 2+ to 3 ILR level on the ODA. The ODA for Level 2 also showed a widespread distribution that included many scores at two levels higher than the ILR per the DLPT5. This information indicated a weak relationship to the ILR per the DLPT5. Specifically, 30% of students who took the ODA scored at the ILR

target level, while 36% of students scored two levels higher than the ILR level and 26% of students scored one level higher than the ILR level. For the ILR level of 3, more scores were distributed at the target level and at one level lower, although there was a fair percentage of scores two and three levels lower, which indicated a moderate to weak relationship to the ILR Level 3 per DLPT5. Specifically, 53% scored at the target ILR level on the ODA, 33% scored one ILR level lower, 5% scored two levels lower, and 10% scored three levels lower or below. For the ILR level of 2+, the ODA scores were distributed across the target level and ILR levels close to the target level, which indicated a moderate ILR relationship to Level 2+ per DLPT5. Specifically, 33% of the students scored at the ILR target level, 33% scored one level higher than the ILR level, and 27% scored one ILR level lower. Data for Level 1+ indicated a weak relationship to the ILR, with 50% of the scores being two levels higher than the ILR and 50% of the scores being at the target ILR level per DLPT5. However, insufficient scores were available to identify clear patterns. According to Krejcie and Morgan's (1970) formula for student sampling, the level of confidence for correlation results is 82% with a .05 margin of error. Figure 33 shows the ODA relationship to the ILR for reading Spanish according to the DLPT5, and Figure 34 shows the ILR percentage distribution for the ODA according to the DLPT5 for reading Spanish.

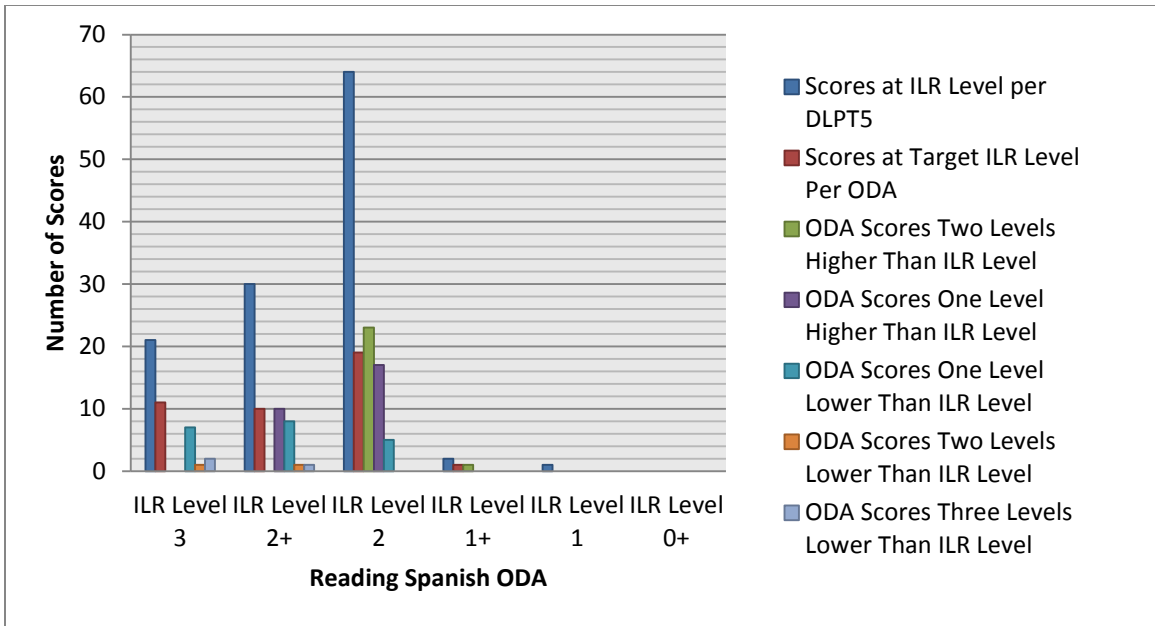


Figure 33. ODA relationship to the ILR—Reading Spanish.

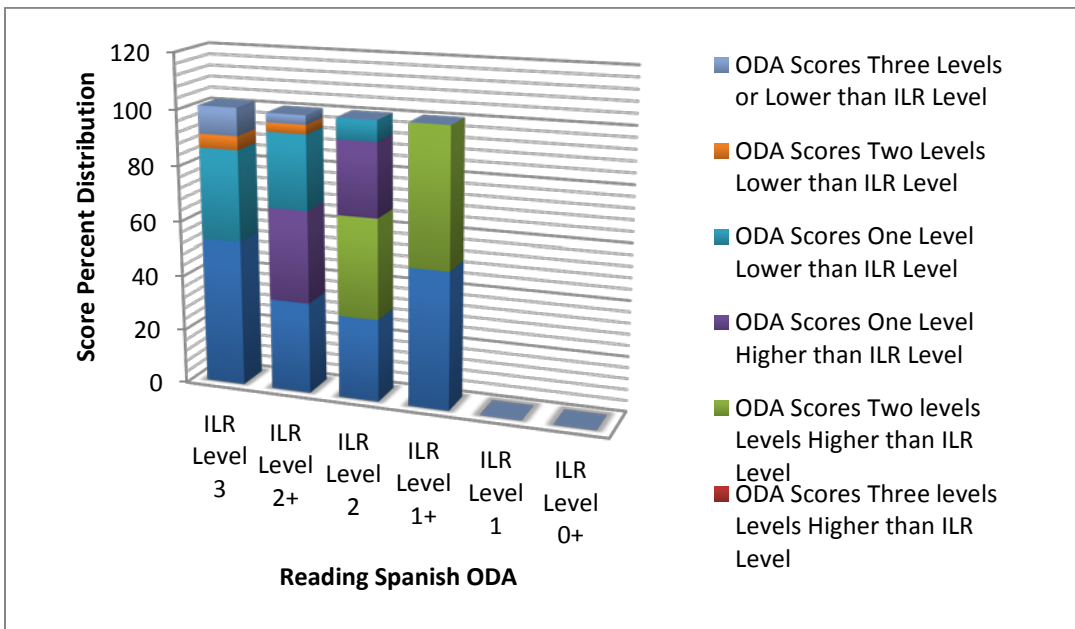


Figure 34. ILR percentage distribution per DLPT5—Reading ODA Spanish.

ODA relationship to the ILR—Reading Korean. For reading, the Pearson product-moment correlation identified a weak correlation for Korean (r value of .23). Per score distribution, based on the Korean sample obtained, the majority of students scored at an

ILR level of 2+ on the DLPT5, while the majority of students scored at Level 2 on the ODA. Level 2 showed the closest relationship to the ILR level, which indicated a moderate relationship to the ILR per the DLPT5. Specifically, 67% of students who took the ODA scored at the target level, 25% of students scored one level lower, and 8% of students scored two levels higher. For ILR Level 2+, ODA scores were spread across different levels with a low percentage of scores at the target level. This score pattern indicated a weak relationship to the ILR level per DLPT5. Specifically, 14% of students scored at the ILR target level, 7% scored one level higher than the ILR level, 57% one level lower, and 22% scored two levels lower than the ILR level per DLPT5. At Level 3, a wide distribution of scores among different ILR levels, including a high number of students scoring two levels lower, indicated a weak relationship to the ILR. Specifically, 8% of students scored at the target ILR level, 25% scored one ILR level lower, 50% of students scored two levels lower, and 2% of students scored three levels lower than the ILR level. The sparse data available at Level 1+ indicated a strong relationship to the ILR, with all scores at the ILR target level. According to Krejcie and Morgan's (1970) formula for student sampling, the level of confidence for correlation results is 49% with a .05 margin of error. Figure 35 shows the ODA relationship to the ILR for reading Korean according to the DLPT5, and Figure 36 shows the ILR percentage distribution for the ODA according to the DLPT5 for reading Korean.

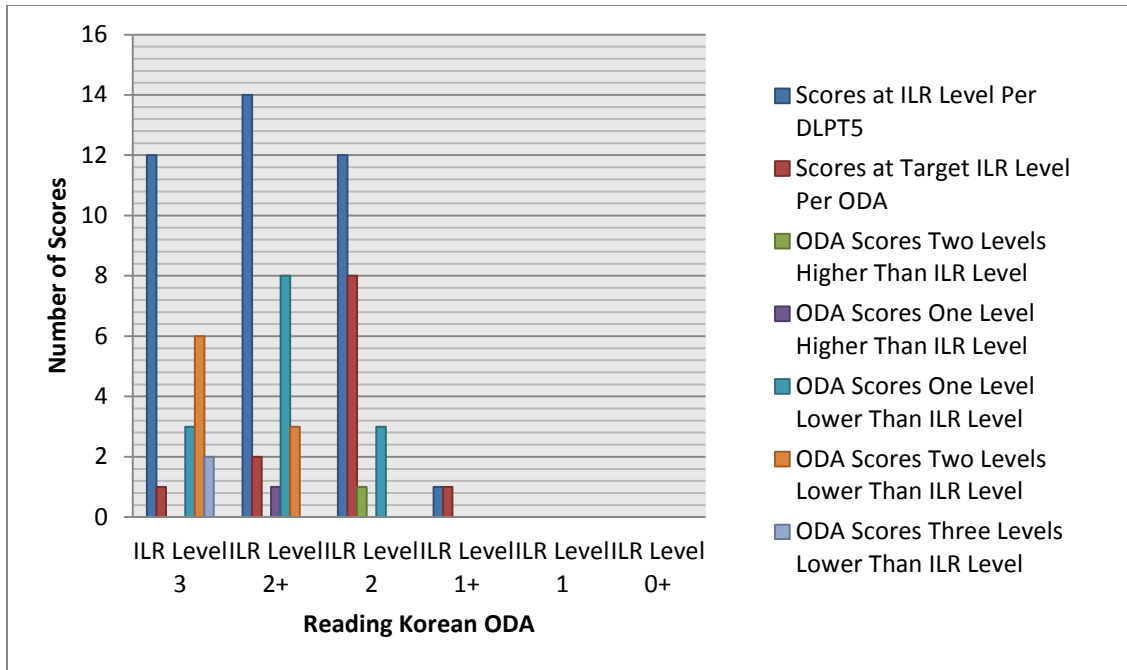


Figure 35. ODA relationship to the ILR—Reading Korean.

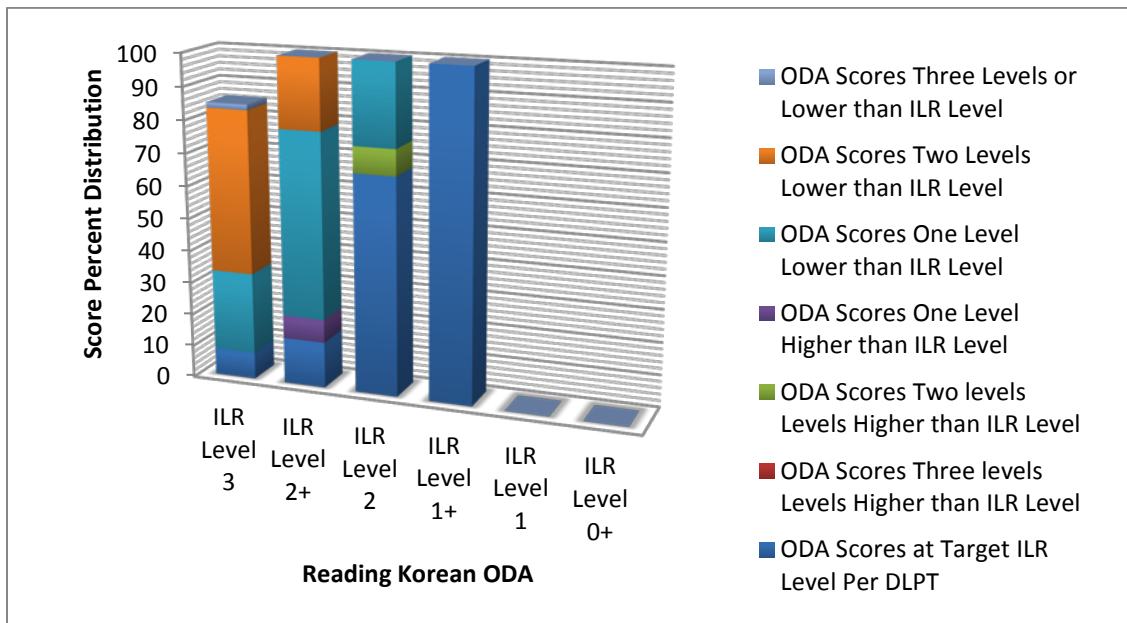


Figure 36. ILR percentage distribution per DLPT5—Reading ODA Korean.

ODA relationship to the ILR—Reading Chinese Mandarin. For reading, the Pearson product–moment correlation identified a moderate correlation for Chinese Mandarin (r value of .34). Per score distribution, based on the Chinese Mandarin sample

obtained, the ODA levels across all levels showed the closest relationship to the ILR levels compared to all other languages per the DLPT5. The majority of students scored at an ILR level of 2+ on the DLPT5 and at a 2+ on the ODA. For Level 2+, data indicated a moderate relationship to the ILR per DLPT5. Specifically, 54% of students scored at the target ILR level, 23% of students scored at one ILR level higher, 13% of students scored one level lower than the ILR, and 10% of students scored two levels lower than the ILR level per DLPT5. At Level 3, data indicated a moderate relationship to the ILR per the DLPT5. Specifically, 56% of students scored at the target ILR level, 33% of students scored one ILR level lower, and 11% students scored two ILR levels lower than the target ILR level per DLPT5. For Level 2, data indicated a moderate relationship to the ILR per DLPT5. Specifically, 50% of students scored at the target level, with 29% of students scoring one level higher and 14% scoring two levels higher than the target ILR level per DLPT5. The few scores at ODA Levels 1+ and 1 showed a weak relationship to the ILR per DLPT5. Specifically, for Level 1+, 33% of students scored at the ILR target level, and 67% of students scored two levels higher. For Level 1, all students scored two levels higher than the target ILR level per DLPT5. According to Krejcie and Morgan's (1970) formula for student sampling, the level of confidence for correlation results is 61% with a .05 margin of error. Figure 37 shows the ODA relationship to the ILR for reading Chinese Mandarin according to the DLPT5, and Figure 38 shows the ILR percentage distribution for the ODA according to the DLPT5 for reading Chinese Mandarin.

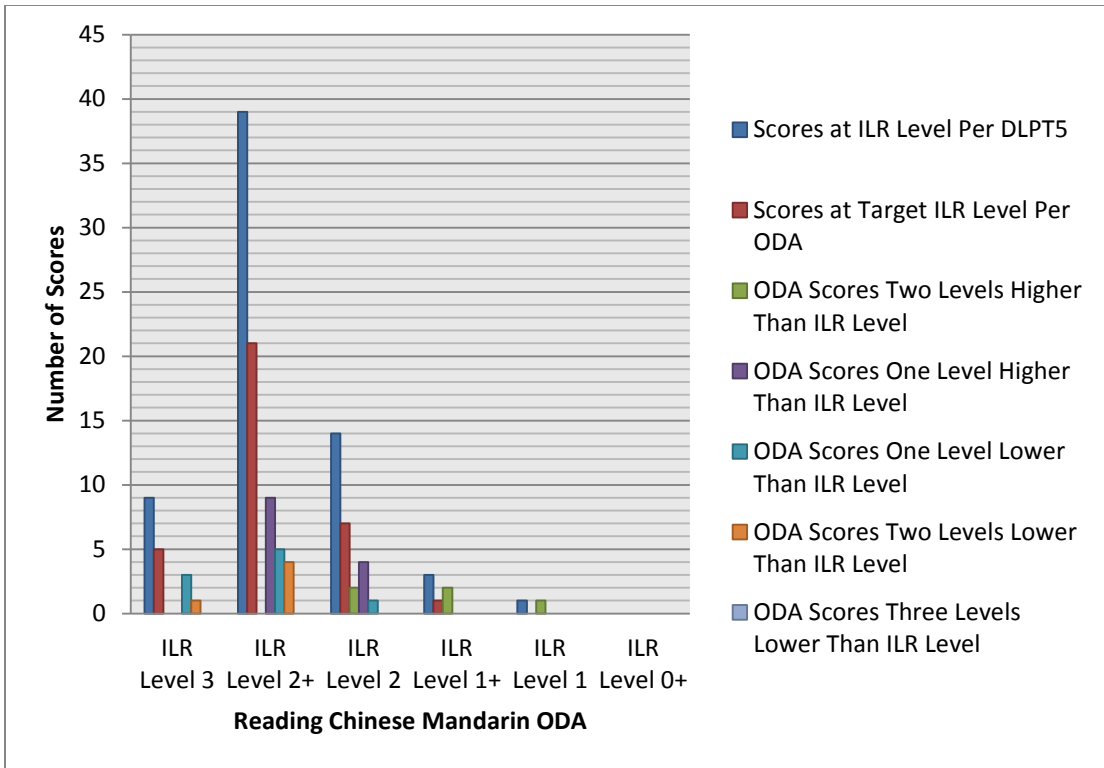


Figure 37. ODA relationship to the ILR—Reading Chinese Mandarin.

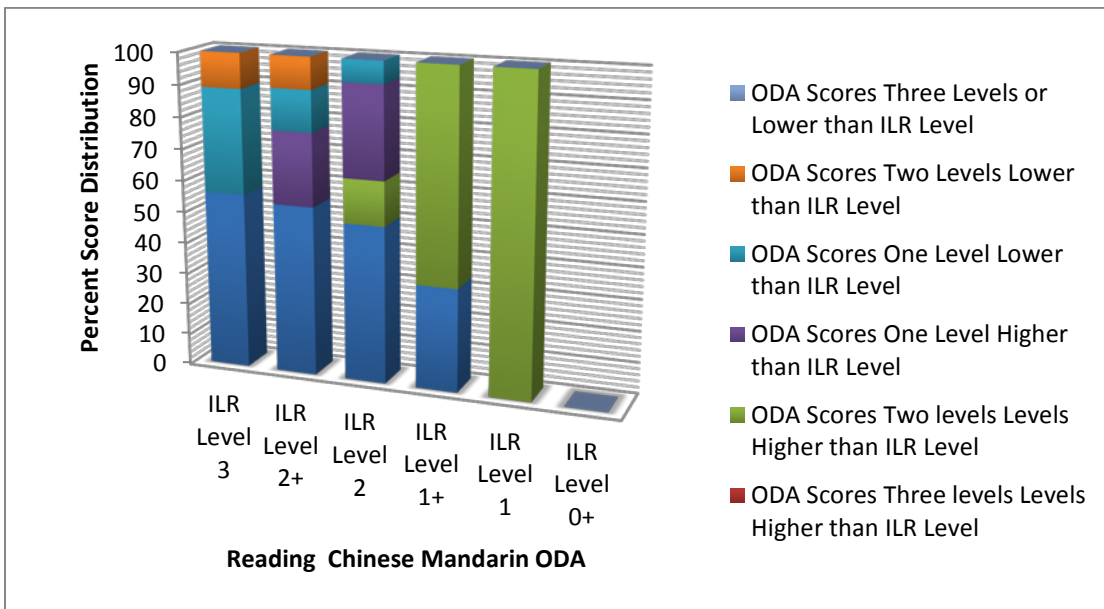


Figure 38. ILR percentage distribution per DLPT5—Reading ODA Chinese Mandarin.

ODA relationship to the ILR—Reading Standard Arabic. For reading, the Pearson product–moment correlation identified a moderate correlation for Standard Arabic (r value of .30). Per score distribution, based on the Standard Arabic sample obtained, the majority of students scored at an ILR level of 2+ on the DLPT5, and the majority of students scored at Level 2 on the ODA. According to the ODA scores, the ILR level of 2 showed the closest relationship to the ILR level. For Level 2 data, the relationship with the ILR was moderate: 56% of students who took the ODA scored at the target ILR level, and 44% of students scored one ILR level lower per DLPT5. For Level 2+, data indicated a low to moderate relationship to the ILR per DLPT5: 4.5% of students scored at the ILR target level, 23% of students scored one level higher than the ILR level, 55% scored one level lower than the ILR level, and 18% scored two levels lower than the ILR level. At Level 3, the ODA student scores showed a wide spread of scores across various levels, including two and three levels lower than the DLPT5, which indicated a weak relationship to ILR level per DLPT5: 33.3% of students scored at the target ILR level, 33.3% scored one ILR level lower, 25% scored two ILR levels lower, and 8.3% students scored three levels lower than the ILR level per DLPT5. The ODA scores for Levels 1+ and 1 indicated a moderate relationship to the ILR. For Level 1+, 33% of students scored at the target level, and 67% scored one level higher than the ILR level. For Level 1, all scores were at one level higher than the ILR level. However, there were not enough data available at Levels 1+ and 1 to identify clear trends of ILR alignment. According to Krejcie and Morgan’s (1970) formula for student sampling, the level of confidence for correlation results is 54% with a .05 margin of error. Figure 39 shows the ODA relationship to the ILR for reading Standard Arabic according to the DLPT5, and Figure

40 shows the ILR percentage distribution for the ODA according to the DLPT5 for reading Standard Arabic.

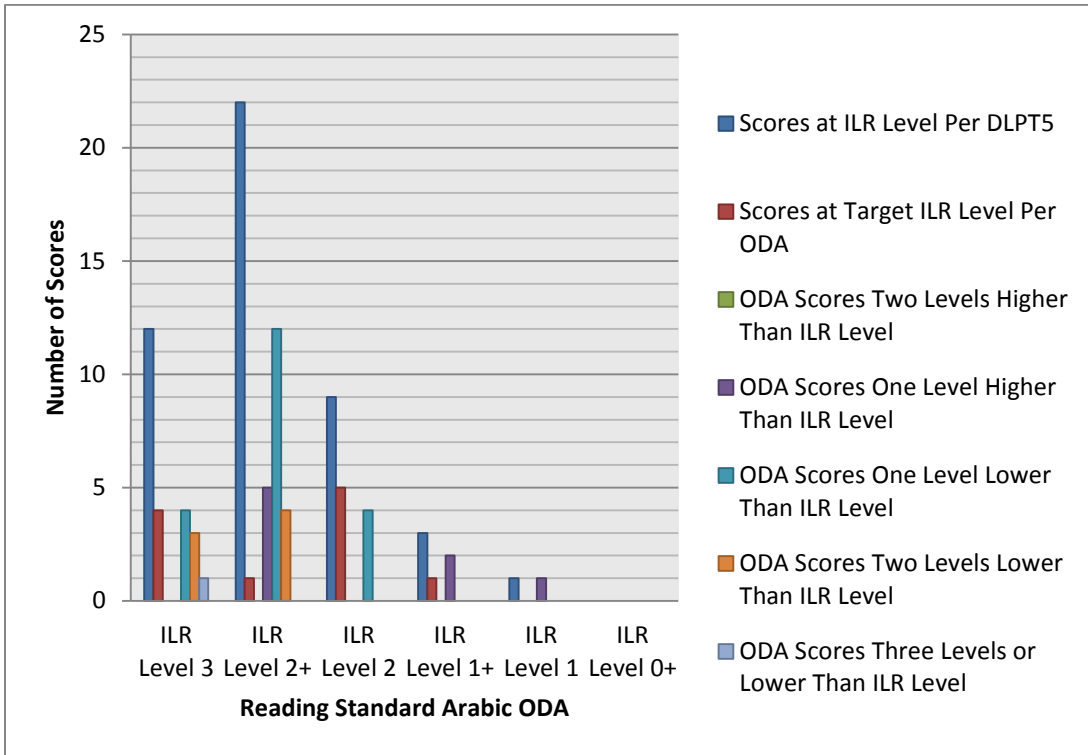


Figure 39. ODA relationship to the ILR—Reading Standard Arabic.

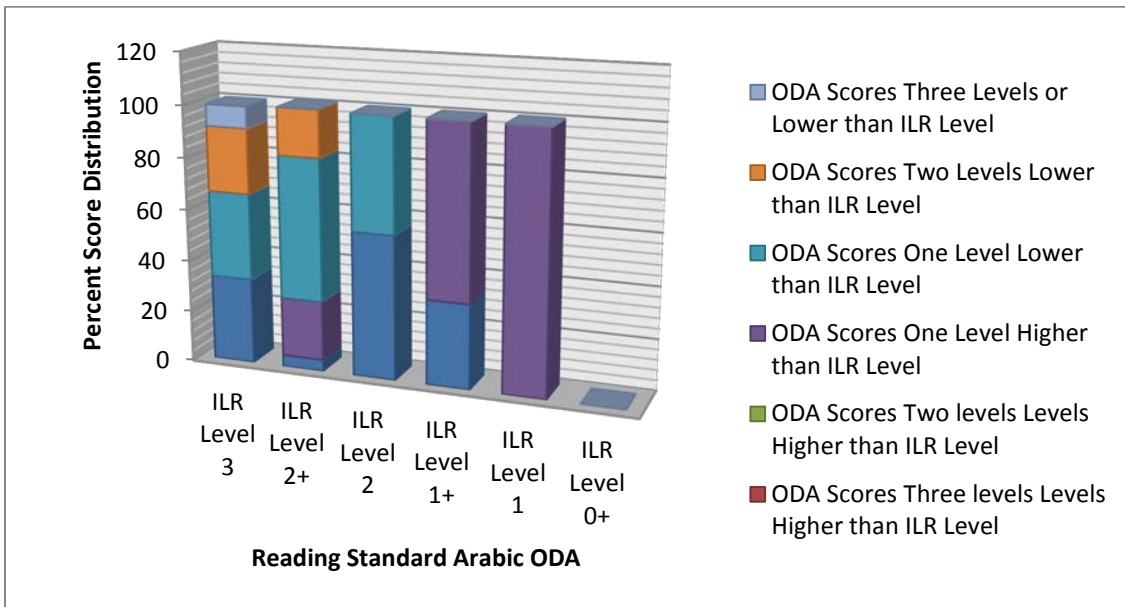
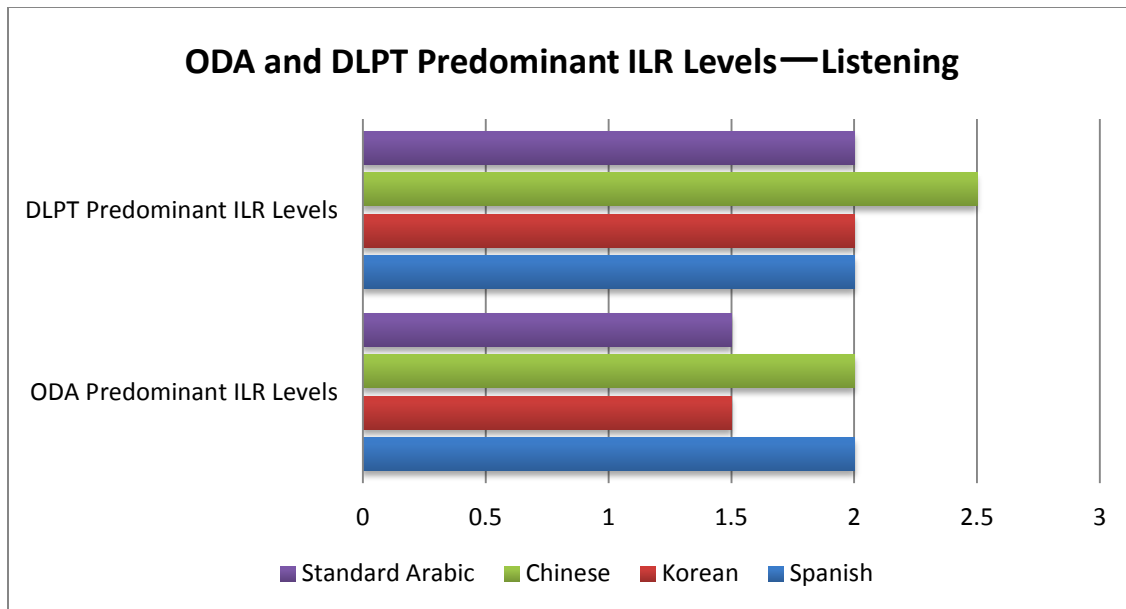


Figure 40. ILR percentage distribution per DLPT5—Reading ODA Standard Arabic.

Research Question 3. Research Question 3 was as follows: Are the relationships found between ODA and DLPT5 for Spanish, Korean, Chinese Mandarin, and Standard Arabic consistent across the levels or is there variance in the relationship depending on the level. Data from Excel files organized by ILR level to identify the areas where ODA levels might align with the DLPT5 more consistently indicated that Standard Arabic had the closest alignment between the ODA and the DLPT5 across all levels for listening. Chinese Mandarin followed by Standard Arabic had the closest alignment between the ODA and the DLPT5 across all levels for reading. For the ODA for listening, at ILR Levels 3 and 2+, students scored one to two levels lower than the ILR target level per DLPT5. This pattern indicated that for the majority of students who had an ILR score of 3 or 2+ on the DLPT5 for listening, there was a variance across all languages studied of one or two levels lower on the ODA. For the ODA for reading at the ILR Level 3 and 2+, there was a closer relationship between the DLPT5 and the ODA at the upper ILR levels for all languages. At these levels, the ODA scores fell at the target level to one level lower, whereas at Level 2+, scores fell at the target level to one level lower or higher depending on the language.

Relationship between the ODA and the DLPT5 for listening. When looking at the predominant scores per level on the DLPT5 and the ODA at a global level, regardless of the specific DLPT5 to ODA level to level relationship, the ODA scores were predominantly one level lower than the DLPT5 scores for all languages studied except for Spanish. For Spanish, the highest number of scores were predominantly at the same ILR level on both the DLPT5 and the ODA. Figure 41 shows the predominant ILR listening levels on the ODA and the DLPT5.



Listening Predominant ILR Levels	Spanish	Korean	Chinese Mandarin	Standard Arabic
DLPT5	LEVEL 2	LEVEL 2	LEVEL 2+	LEVEL 2
ODA	LEVEL 2 <i>Moderate (r)</i>	LEVEL 1+ to LEVEL 2 <i>Moderate (r)</i>	LEVEL 2 to LEVEL 1+ <i>Weak (r)</i>	LEVEL 1+ <i>Moderate (r)</i>

Figure 41. Predominant ILR listening levels on the ODA per DLPT5.

Relationship between the ODA and the DLPT5 for listening per ILR level.

When looking at each level, the listening ODA showed some consistent variances across all languages at certain ILR levels. At ILR Level 3, scores showed one to two ILR levels lower for all languages except for Standard Arabic, which was two to three levels lower than the ILR level per DLPT5. For ILR Level 2+, students scored one to two levels lower than the ILR target level per DLPT5, with Spanish having a few scores at the target level. For ILR Level 2, students predominantly scored at the target level or one level lower. There were fewer scores available at Level 1+ overall, but students predominantly scored at the target ILR level to one level higher except for Chinese, which had different ILR ranges. There were fewer scores available at Level 1 overall, but students predominantly

scored one level higher than the ILR level per DLPT5. Table 21 shows the predominant ILR listening levels on the ODA according to the DLPT5.

Table 21

Predominant ILR Listening Levels on the ODA per DLPT5

	ODA Spanish Predominant Level 2 Moderate (<i>r</i>)	ODA Korean Predominant Level 1+ to 2 Moderate (<i>r</i>)	ODA Chinese Mandarin Predominant Level 2 to 1+ Weak (<i>r</i>)	ODA Standard Arabic Predominant Level 1+ Moderate (<i>r</i>)
ILR Level 3	One to two levels lower Level 2+ to 2	Two levels lower Level 2	Two levels lower Level 2	Two to three levels lower Level 2 to 1+
ILR Level 2+	One to two levels lower Level 2 to 1+	One to two levels lower ^a Level 2 ^a	One to two levels lower Level 2 to 1+	Two levels lower Level 1+
ILR Level 2	Target to one level lower Level 2 to 1+	One level lower to target Level 1+ to 2	Target to one level lower Level 2 to 1+	One level lower Level 1+
ILR Level 1+	One level higher to target Level 2 to 1+	Target Level 1+	Two to one level higher to one level lower ^a Level 2+ to 2 to 1+ ^a	On target to two levels lower Level 1+ to 0+ ^a
ILR Level 1	One level higher Level 1+	One level higher Level 1+ ^a	N/A	One level higher to target to one level lower Level 1+ to 1
ILR Level 0+	N/A	N/A	N/A	On target 0+ ^a

^aNot enough scores to identify clear ILR relationship trends.

When looking at the ILR relationship per level across all languages studied for listening, Standard Arabic indicated the most consistency and the least variance across levels with a higher level of discrimination and a more defined level differentiation.

Chinese Mandarin indicated the highest variance, followed by Korean, with little discrimination and differentiation of student scores at lower ILR levels. Table 22 shows the ODA predominant results at each specific ILR level according to the DLPT5.

Table 22

Predominant ILR Listening Levels on the ODA per DLPT5

Listening ILR Levels per DLPT5	ODA Spanish	ODA Korean	ODA Chinese Mandarin	ODA Standard Arabic
	Predominant Level 2 Moderate (<i>r</i>)	Predominant Level 1+ to 2 Moderate (<i>r</i>)	Predominant Level 2 to 1+ Weak (<i>r</i>)	Predominant Level 1+ Moderate (<i>r</i>)
Level 3	Level 2+ to 2	Level 2	Level 2	Level 2 to 1+
Level 2+	Level 2 to 1+	Level 2	Level 2 to 1+	Level 1+
Level 2	Level 2 to 1+	Level 1+ to 2	Level 2 to 1+	Level 1+
ILR Level 1+	Level 2 to 1+	Level 1+	Level 2 to 1+	Level 1+ to 0+ ^a
ILR Level 1	Level 1+	Level 1+	N/A	Level 1+ to 1
ILR Level 0+	N/A	N/A	N/A	0+ ^a

^aNot enough scores to identify clear ILR relationship trends.

Relationship between the ODA and the DLPT5—Listening Spanish. Specifically, for the listening Spanish ODA, based on the Spanish sample obtained, the highest number of students scored at an ILR level of 2 on the DLPT5, and the highest number of students scored at a level of 2 on the ODA. The relationship found between the DLPT5 and the ODA showed a variance depending on the level. At ILR Level 3, students who took the ODA scored one to two levels lower than on the DLPT5: 41% of students scored one level lower than ILR level and 41% of students scored two levels lower than the ILR level. At ILR Level 2+, students who took the ODA predominantly scored one ILR level lower: 59% of students scored one level lower and 27.4% scoring two levels lower than the ILR level per DLPT5. At ILR Level 2, the ODA showed the closest alignment to the ILR levels per DLPT5: 61% of students scored at the ILR Level 2 and 29% scored one ILR level lower per DLPT5. At ILR Level 1+, scores showed a variance, with 60% of students scoring one level higher than the ILR level and 40 % of student scoring at the ILR target level per the DLPT5. Level 1 also showed a variance depending on the ILR level, with 80% of students scoring one level higher than ILR level and 20% of students

scoring two levels higher than ILR level per DLPT5. Figure 42 shows the total number of listening scores at each ILR level per the ODA and per the DLPT5 for Spanish, and Figure 43 shows the relationship between the ODA and the DLPT5 at each ILR level for listening Spanish.

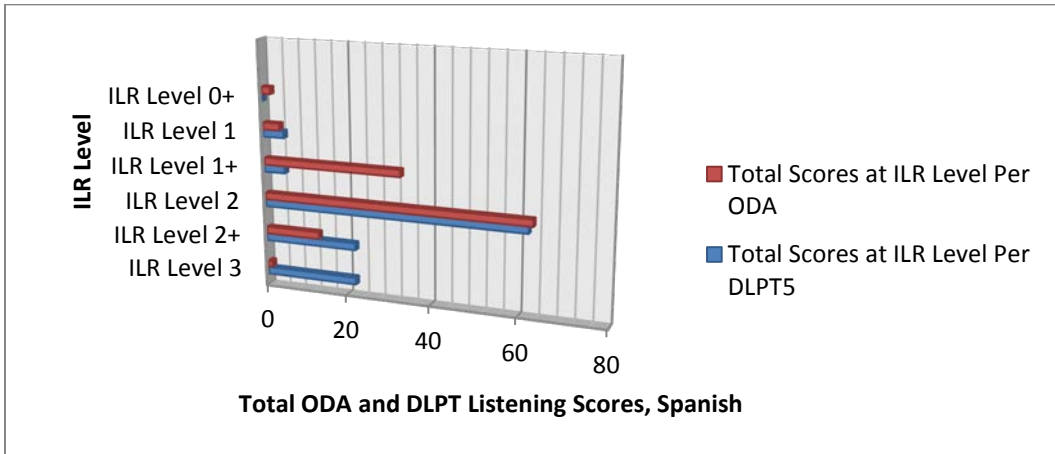


Figure 42. Total ODA and DLPT5 score comparison—Listening Spanish.

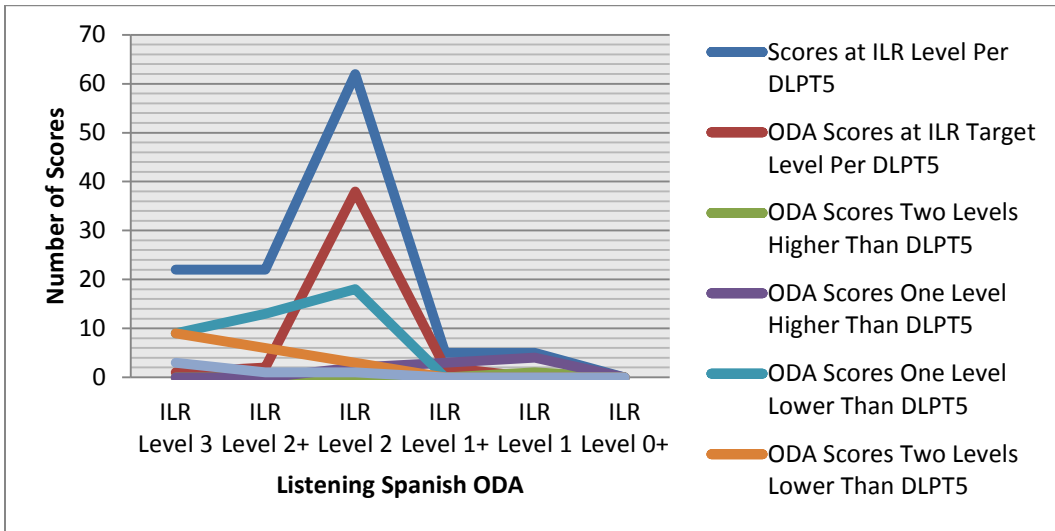


Figure 43. Relationship between the ODA and the DLPT5—Listening Spanish.

Relationship between the ODA and the DLPT5—Listening Korean. Based on the Korean sample obtained, while the highest number of students scored at an ILR level of 2 on the DLPT5, the highest number of students scored at the ILR level of 1+, followed by

Level 2 on the ODA. When looking at all ILR levels, data indicated a variance between the ODA and the DLPT5, depending on the level. At ILR Level 3, students who took the ODA scored two levels lower than the DLPT5: 86% of students scored two levels lower and 14% of scored one level lower than the ILR level per DLPT5. For Level 2+, students scored two levels lower than the DLPT5, which indicated a variance between the ODA and the DLPT5: 67% scored one level lower and 33% scored two levels lower. Level 2 showed moderate variance depending on the level, with 27% of students scoring at the target level and 50% scoring one level lower than the ILR level per DLPT5. At Level 1+, scores showed the least variance, with all scores at the target ILR level. At Level 1, scores indicated variance depending on the level, with ODA scores at one level higher than ILR level per DLPT5. For Levels 1+ and 1, there were not enough scores available to identify a consistent pattern. Figure 44 shows the total number of listening scores at each ILR level per the ODA and per the DLPT5, and Figure 45 shows the relationship between the ODA and the DLPT5 at each ILR level for listening Korean.

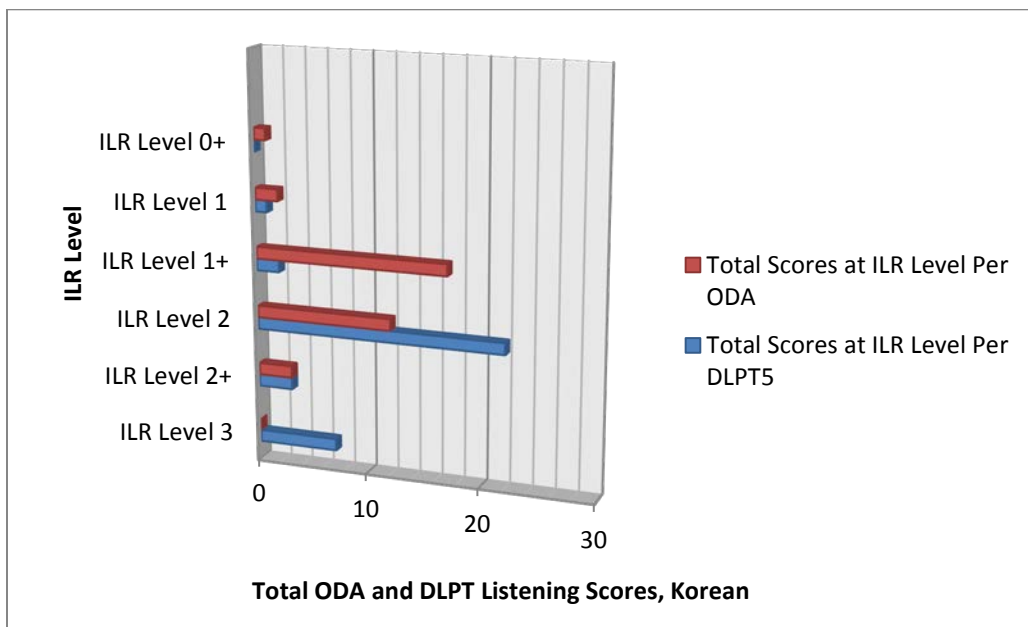


Figure 44. Total ODA and DLPT5 score comparison—Listening Korean.

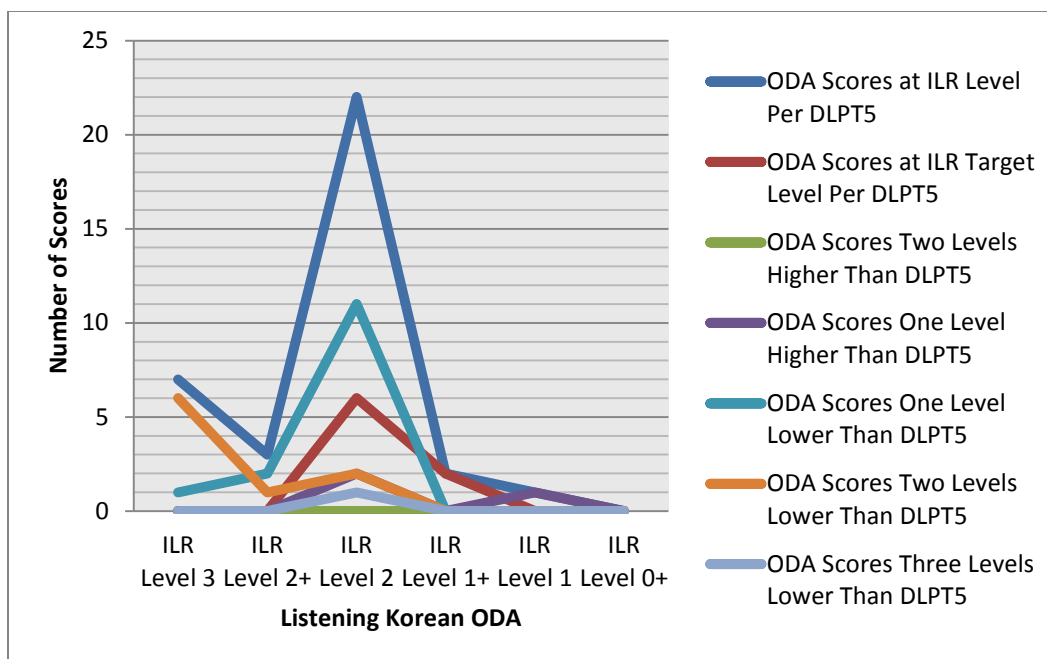


Figure 45. Relationship between the ODA and the DLPT5—Listening Korean.

Relationship between the ODA and the DLPT5—Listening Chinese Mandarin.

Based on the Chinese Mandarin sample obtained, while the highest number of students scored at Level 2+ on the DLPT5, the highest number of students scored one or two ILR levels lower on the ODA. Specifically, for Level 2+, 42% of students scored one level lower and 39% scored two levels lower than the DLPT5. When looking at all ILR levels, sample data indicated a variance depending on the level. For Level 3, data indicated a variance, with 59% of students scoring two levels lower and 18% of students scoring one level lower. The ODA for Level 2 showed the least variance, with 50% of students scoring at the target level on the DLPT5 and 25% of students scoring one level lower than the DLPT5. For Level 1+, data indicated a variance depending on the level, with 33% of students scoring two levels higher than the DLPT5, 33% scoring one level lower than the DLPT5, and 33% scoring one level higher than the DLPT5, although there were sparse data to identify conclusive alignment patterns. Figure 46 shows the total number of

listening scores at each ILR level per the ODA and per the DLPT5 for Chinese Mandarin, and Figure 47 shows the relationship between the ODA and the DLPT5 at each ILR level for listening Chinese Mandarin.

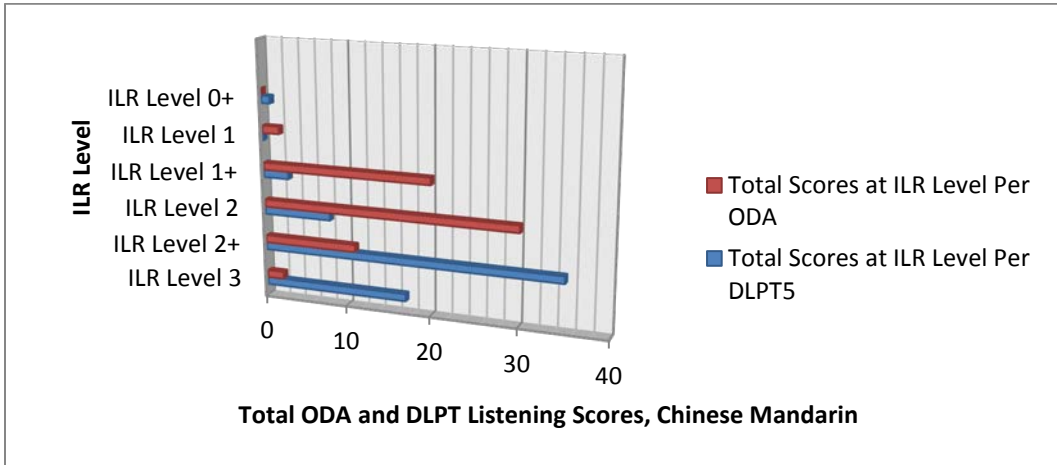


Figure 46. Total ODA and DLPT5 score comparison—Listening Chinese Mandarin.

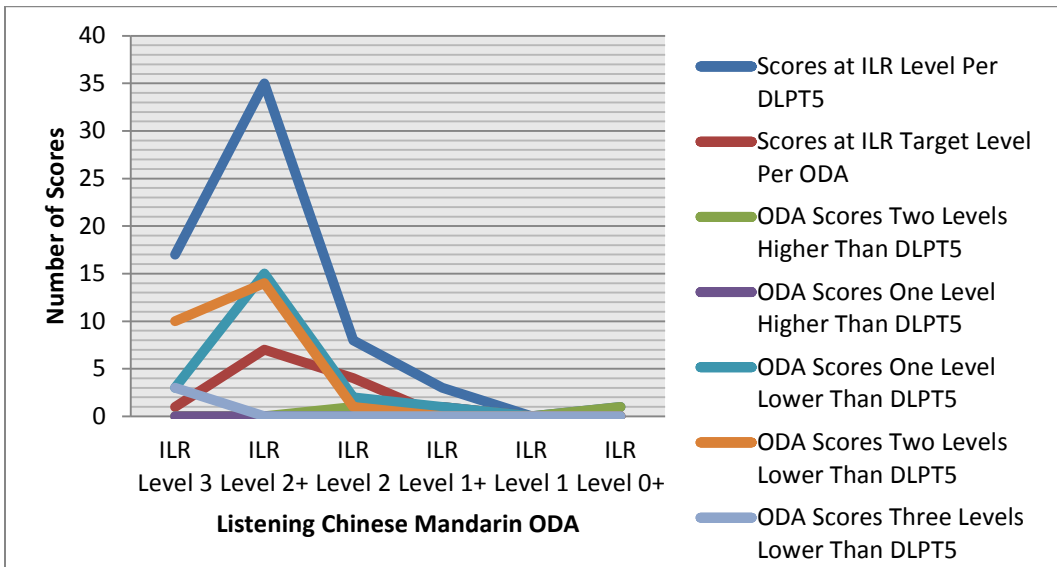


Figure 47. Relationship between the ODA and the DLPT5—Listening Chinese Mandarin.

Relationship between the ODA and the DLPT5—Listening Standard Arabic.

Based on the Standard Arabic sample obtained, while the highest number of students scored at an ILR level of 2 on the DLPT5, the highest number of students scored at a

level of 1+ on the ODA. The relationship found between the DLPT5 and the ODA showed a tendency for students to score two levels lower on the ODA for Levels 3 and 2+. At ILR Level 3, students who took the ODA scored two levels lower than the DLPT5, with 60% of students scoring one level lower and 40% of students scoring two levels lower. At Level 2+, data indicated variance depending on the level, with 91% of students scoring two levels lower than the DLPT5. At Level 2, data indicated a variance depending on the level, with 85% of scores one level lower and 10% of scores at the target level. At Level 1+, data indicated a variance depending on the level, with 67% of students scoring at the ILR target level and 33% scoring two levels lower than the DLPT5. For Level 1, data indicated a variance depending on the level, with the least consistency in student scores: 20% of scores were at the target level, 20% of scores were one level lower, 40% of scores were one level higher, and 20% of scores were two levels higher than the DLPT5. Although there were few data available for 0+, all data were distributed at the target level. Additionally, Standard Arabic showed a higher number of scores at all ILR levels, including the lower levels, thus suggesting a higher level of discrimination at the ILR level, which might have contributed to a higher level of correlation when compared to the other languages studied. Figure 48 shows the total number of listening scores at each ILR level per the ODA and per the DLPT5 for Standard Arabic, and Figure 49 shows the relationship between the ODA and the DLPT5 at each ILR level for listening Standard Arabic.

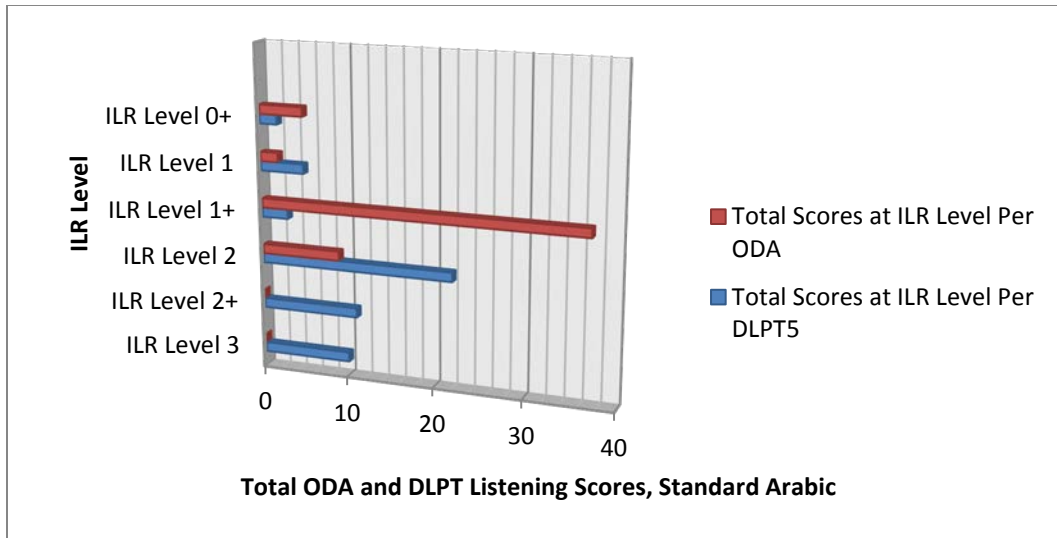


Figure 48. Total ODA and DLPT5 score comparison—Listening Standard Arabic.

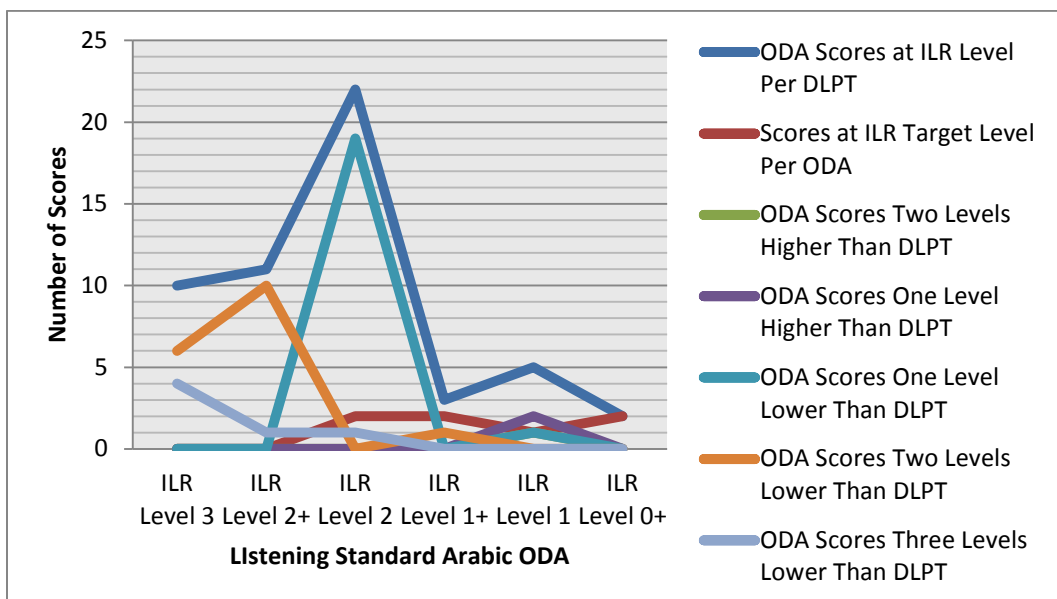


Figure 49. Relationship between the ODA and the DLPT5—Listening Standard Arabic.

Relationship between the ODA and the DLPT5—Reading. When looking at the predominant ILR levels on the DLPT5 and the ODA at a global level, regardless of the specific DLPT5 to ODA level-to-level relationship, students predominantly obtained higher scores on the DLPT5 than on the ODA, except for the Spanish test. For the Spanish test, students predominantly obtained higher scores on the ODA than on the

DLPT5. Figure 50 shows the predominant ILR reading levels on the ODA and the DLPT5.

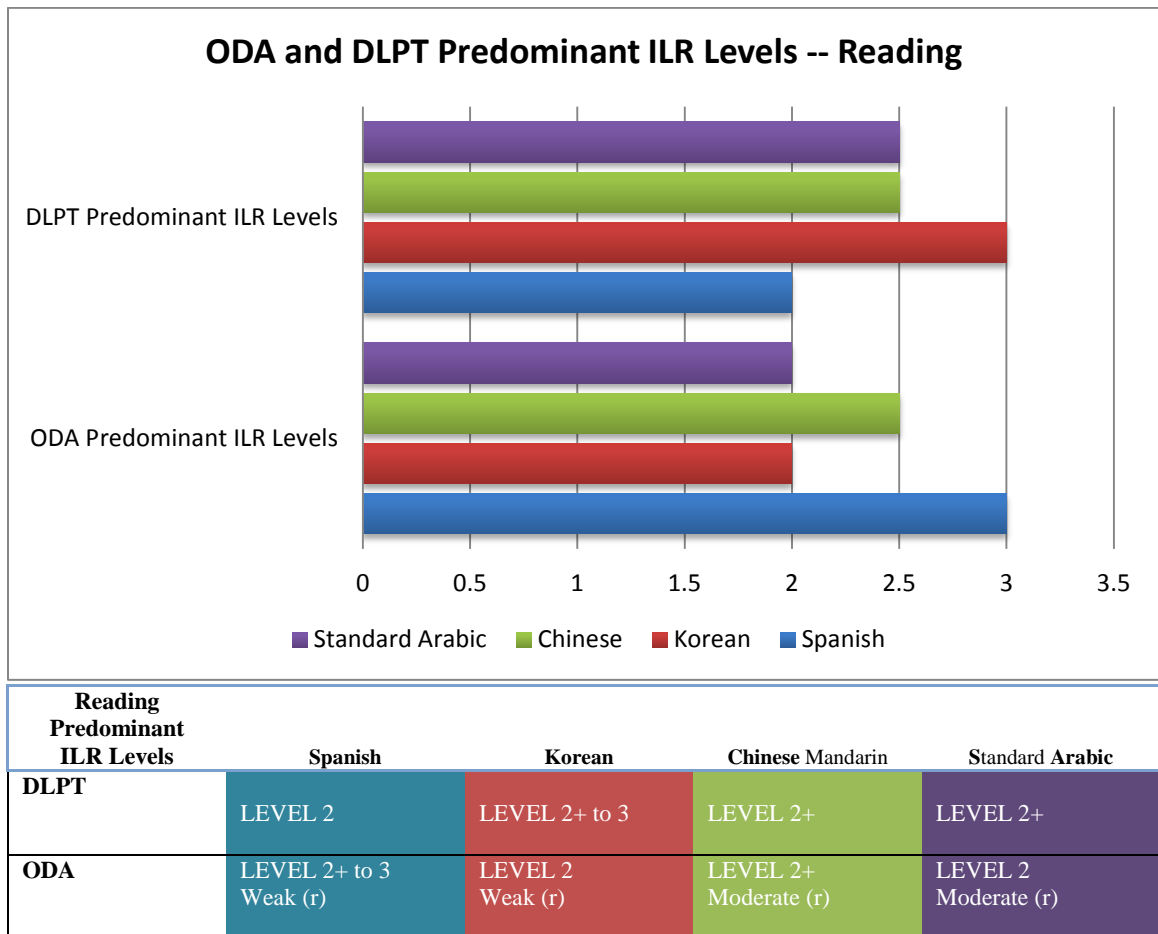


Figure 50. Predominant ILR reading levels on the ODA per DLPT5.

Relationship between the ODA and the DLPT5 for Reading per ILR level. Data organized by ILR level to identify the areas where the ODA levels might more consistently align with the DLPT5 suggested that, for reading, at the ILR Level 3, ODA scores aligned at the target level or at one level lower than the DLPT5 with the exception of Korean, where ODA scores aligned two to one levels lower than the DLPT5. At Level 2+, data indicated a variance depending on the level, with a tendency for Spanish and Chinese Mandarin to score at the target to one level higher than the DLPT5 and Korean

and Standard Arabic to score one to two levels lower than the DLPT5. At Level 2, scores showed the least variance, with data for all languages showing a tendency to score at the target level to one level higher or lower, with the exception of Spanish, where scores fell at the target to two levels higher than the DLPT5. Table 23 shows the predominant ILR reading levels on the ODA according to the DLPT5.

Table 23

Predominant ILR Reading Levels on the ODA per the DLPT5

ILR levels per DLPT5	ODA Spanish Predominant Level 2+ to 3 Weak (r)	ODA Korean Predominant Level 2 Weak (r)	ODA Chinese Mandarin Predominant Level 2+ Moderate (r)	ODA Standard Arabic Predominant Level 2 Moderate (r)
ILR Level 3	Target to one (to two levels lower) Level 3 to 2+ to 2	Two to one level lower Level 2 to 2+	Target to one level lower Level 3 to 2+	Target to one level lower to three levels lower Level 3 to 2+ to 2
ILR Level 2+	Target to one level higher to one lower Level 2+ to 3 to 2	One level lower, one level higher to two levels lower Level 2	Target to one level higher Level 2+ to 3	One level lower (to two levels lower) Level 2
ILR Level 2	Two levels higher to target to one level higher Level 3 to 2 to 2 +	Target to one level lower Level 2 to 1+	Target to one level higher Level 2 to 2+	Target to one level lower Level 2 to 1+
ILR Level 1+	Target to two levels higher Level 1+ to 2+ ^a	Target Level 1+ ^a	Two levels higher to target Level 2+ to 1+ ^a	One level higher to target Level 1+ to 2 ^a
ILR Level 1	N/A	N/A	Two levels higher Level 2 ^a	One level higher Level 1+
ILR Level 0+	N/A	N/A	N/A	N/A

^aNot enough scores to identify clear ILR relationship trends.

When looking at the ILR relationship per level across all languages studied for reading, Chinese Mandarin followed by Standard Arabic had the closest alignment between the ODA and the DLPT5 across all levels and the highest discrimination and score differentiation across ILR levels. At the ILR Level 3 and 2+, there was a closer relationship between the DLPT5 and the ODA at the upper ILR levels compared to listening, with a higher number of scores at Levels 2+ and 3. At these upper levels, Spanish and Chinese Mandarin showed the least discrimination and the lowest score differentiation between ILR Levels 2+ and 3, with Spanish showing the least discrimination and score differentiation across all levels. Table 24 shows the ODA predominant results at each specific ILR level according to the DLPT5.

Table 24

Predominant ILR Reading Levels on the ODA per the DLPT5

ILR levels per DLPT5	ODA Spanish Predominant Level 2+ to 3 Weak (<i>r</i>)	ODA Korean Predominant Level 2 Weak (<i>r</i>)	ODA Chinese Mandarin Predominant Level 2+ Moderate (<i>r</i>)	ODA Standard Arabic Predominant Level 2 Moderate (<i>r</i>)
ILR Level 3	Level 3 to 2 to 2 +	Level 2 to 2+	Level 2+ to 3	Level 3 to 2+ to 2
ILR Level 2+	Level 3 to 2 to 2 +	Level 2	Level 2+ to 3	Level 2
ILR Level 2	Level 3 to 2 to 2 +	Level 2 to 1+	Level 2 to 2+	Level 1+ to 2
ILR Level 1+	Level 1+ to 2+ ^a	Level 1+ ^a	Level 2+ to 1+ ^a	Level 1+ to 2
ILR Level 1	N/A	N/A	Level 2 ^a	Level 1+
ILR Level 0+	N/A	N/A	N/A	N/A

^aNot enough scores to identify clear ILR relationship trends.

Relationship between the ODA and the DLPT5—Reading Spanish. Specifically, for the Spanish ODA for reading, while the highest number of students scored at an ILR level of 2 on the DLPT5, the highest number of students scored at Level 2+ and 3 on the ODA. The relationship found between the DLPT5 and the ODA showed a variance depending on the level. Data for Level 3 indicated a moderate to weak variance between

the ODA and the DLPT5, with 52% of students scoring at the target ILR level, 33% scoring one ILR level lower, and 15% scoring two or three levels lower. For the ILR Level 2+, data indicated a moderate variance, with 33% of the students scoring at the target level, 33% scoring one level higher, and 27% scoring one level lower than the DLPT5. Data for Level 2 indicated a variance depending on the level, with 30% of students who took the ODA scoring at the target level on the DLPT5, 36% of students scoring two levels higher than ILR level, and 26% of students scoring one level higher than the DLPT5. More data may need to be available for Level 1+ and 1, which showed a score at the target level and score two levels higher for Level 1+ and a score three and above levels higher than ILR level for Level 1. The latter indicated an irregular test-taking condition. Figure 51 shows the total number of reading scores at each ILR level per the ODA and per the DLPT5 for Spanish, and Figure 52 shows the relationship between the ODA and the DLPT5 at each ILR level for reading Spanish.

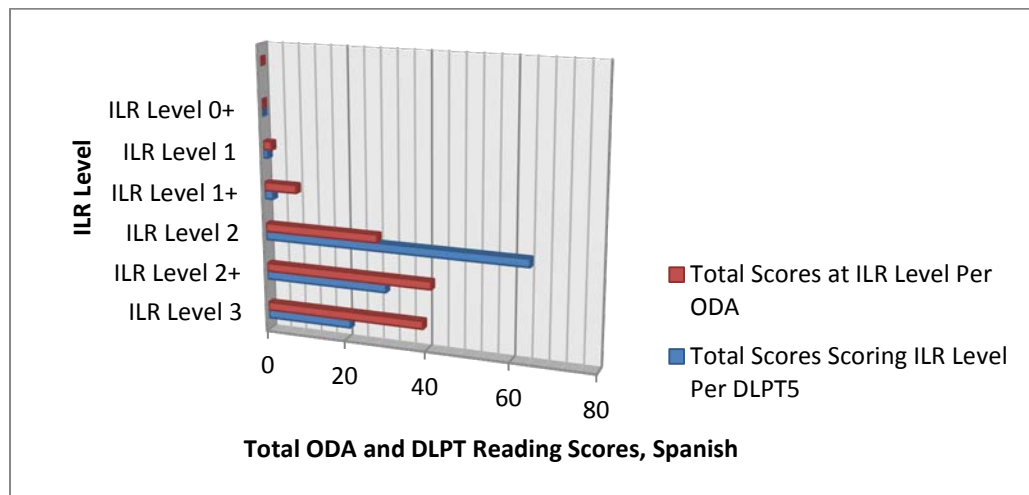


Figure 51. Total ODA and DLPT5 score comparison—Reading Spanish.

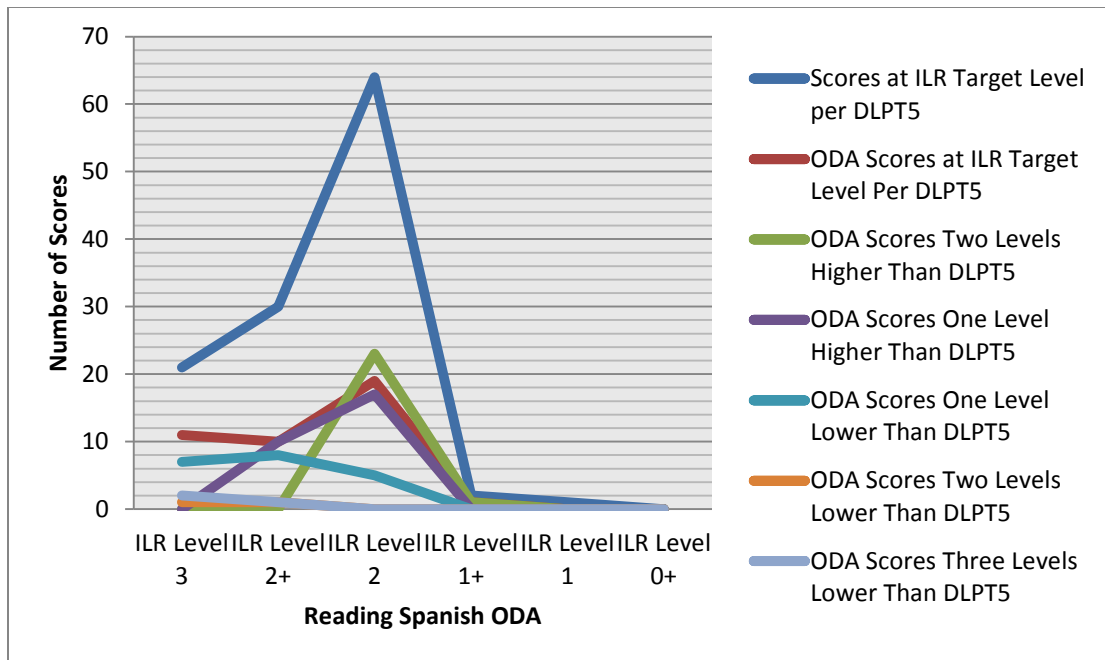


Figure 52. Relationship between the ODA and the DLPT5—Reading Spanish.

Relationship between the ODA and the DLPT5—Reading Korean. For the ODA for Korean, while the highest number of students scored at an ILR level of 2+ and 3 on the DLPT5, the highest number of students scored at a level of 2 ODA. The ODA showed the highest variance at Level 3, with 50% of students scoring two levels lower and 25% scoring one level lower than the DLPT5. ILR Level 2+ also showed a variance, with 57% of students scoring one level lower, and 22% scoring two levels lower, than the DLPLT. ILR Level 2 showed the closest relationship to the DLPT5, with 67% of scores at the target level and 25% of scores one level lower than the DLPT5. There were not enough scores at Level 1 or 1+ to verify patterns of alignment. The data available at Level 1+ indicated a strong relationship to the ILR with all scores at the ILR target level. Figure 53 shows the total number of reading scores at each ILR level per the ODA and per the DLPT5 for Korean, and Figure 54 shows the relationship between the ODA and the DLPT5 at each ILR level for reading Korean.

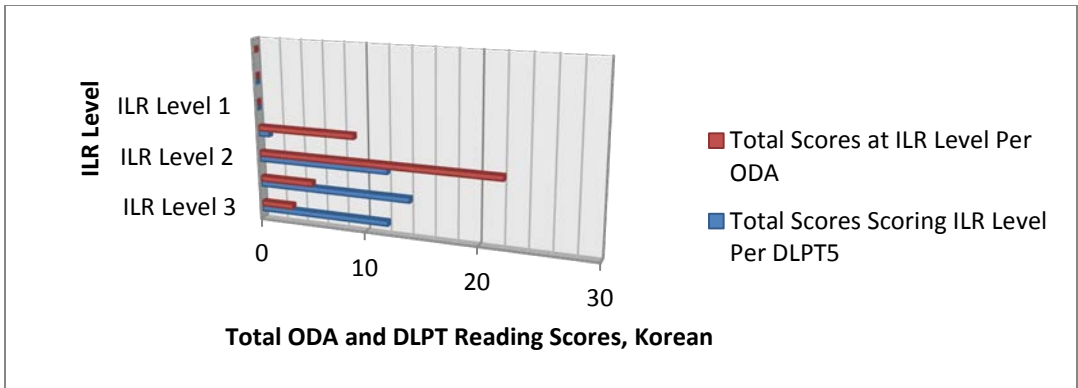


Figure 53. Total ODA and DLPT5 score comparison—Reading Korean.

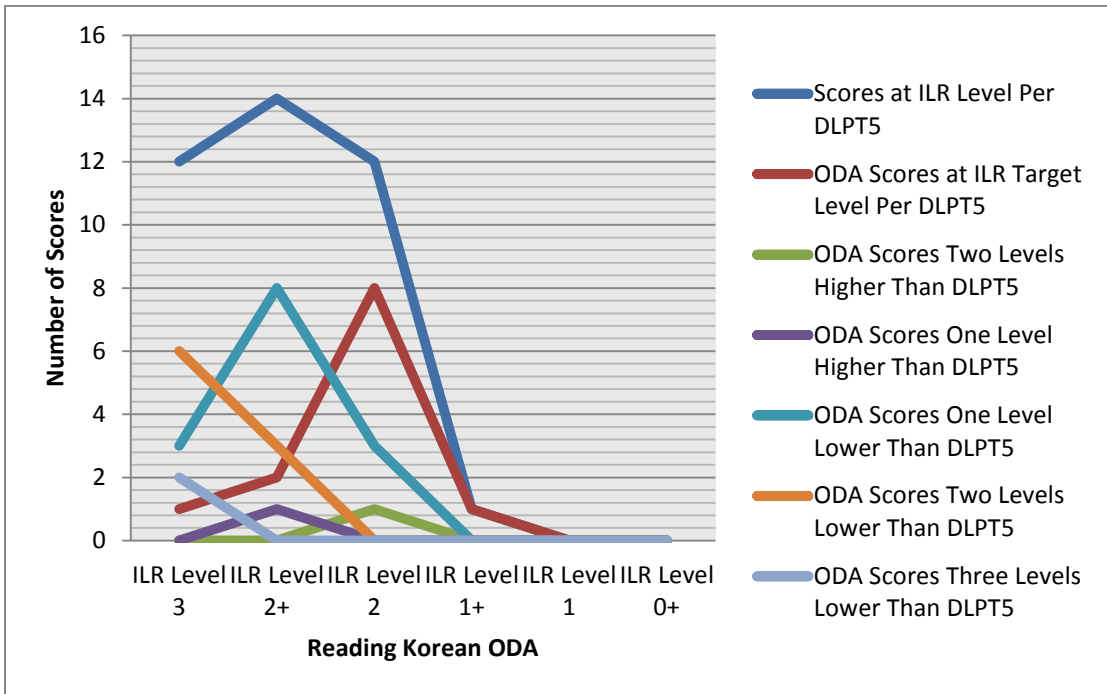


Figure 54. Relationship between the ODA and the DLPT5—Reading Korean.

Relationship between the ODA and the DLPT5 – Reading Chinese Mandarin. For the Chinese Mandarin ODA for reading, based on the Chinese Mandarin sample obtained, the majority of students scored at an ILR level of 2+ on the DLPT5. Similarly, the majority of students scored at an ILR level of 2+ on the ODA. The relationship found between the DLPT5 and ODA showed consistency in the proportion of the DLPT5 and of the ODA scores at the target level for all levels. At Level 3, there is a fair consistency

between the DLPT5 and the ODA, with 56% of students scoring at the target ILR level and 33% of students scoring one ILR level lower than the DLPT5. At the ILR level of 2+, there is a fair consistency between the DLPT5 and the ODA, with 54% of students scoring at the target ILR level and 23% of students scoring one ILR level higher than the DLPT5. For Level 2, there is a fair consistency between the ODA and the DLPT5, with 50% of students scoring at the target level and 29% of students scoring one level higher than the DLPT5. There were not enough scores at ILR Level 1+ or 1 to verify scoring patterns. The few data available showed variance, with 33% of students scoring at the ILR target level and 67% of students scoring two levels higher than the DLPT5. Few data available for Level 1 were distributed two levels higher than DLPT5, which indicated a variance. Figure 55 shows the total number of reading scores at each ILR level per the ODA and per the DLPT5 for Chinese Mandarin, and Figure 56 shows the relationship between the ODA and the DLPT5 at each ILR level for reading Chinese Mandarin.

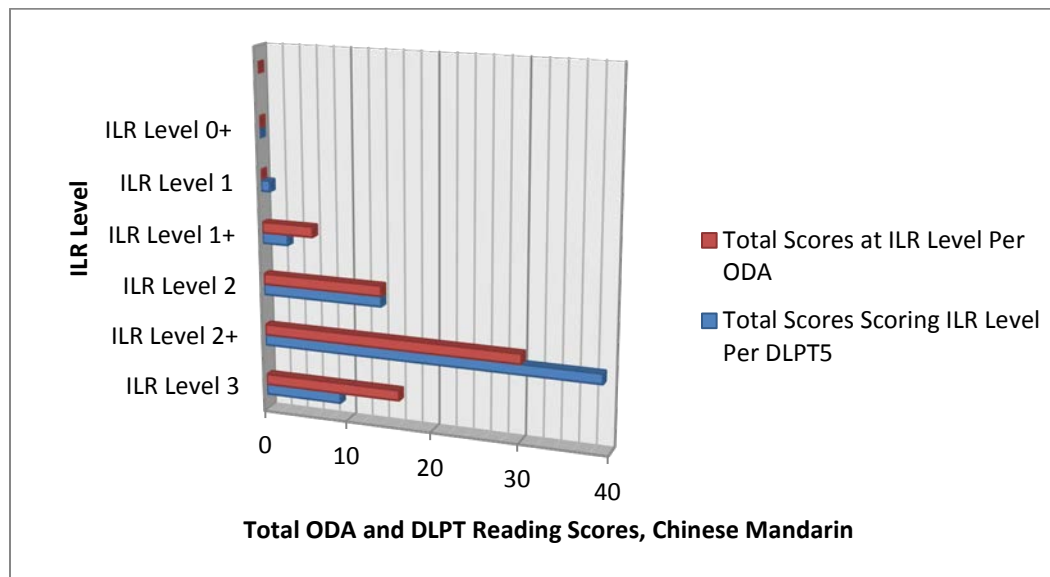


Figure 55. Total ODA and DLPT5 score comparison—Reading Chinese Mandarin.

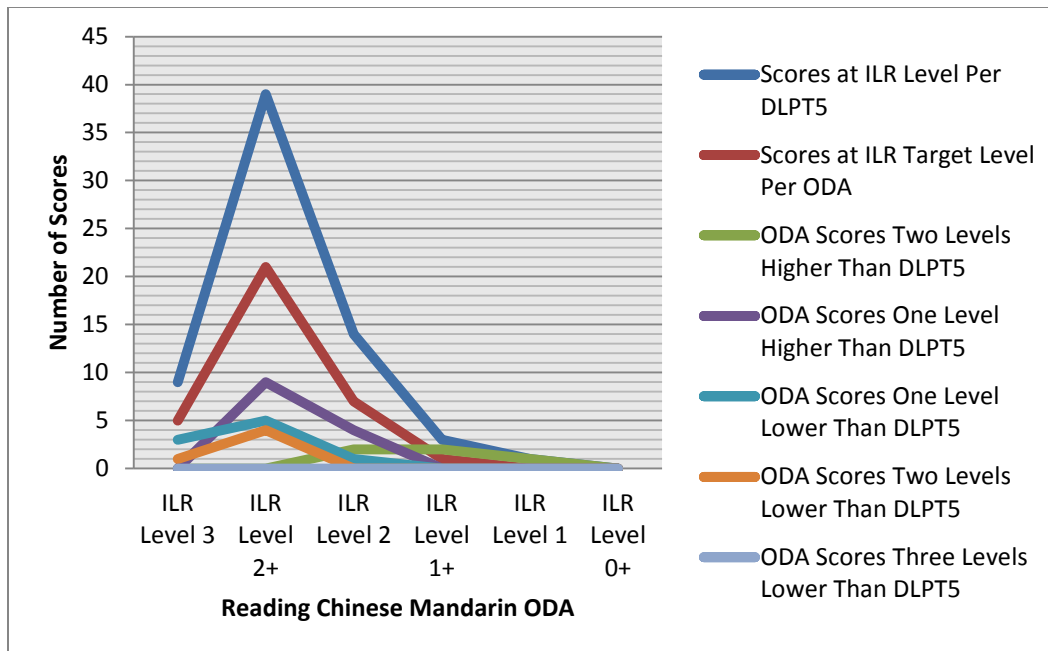


Figure 56. Relationship between the ODA and the DLPT5—Reading Chinese Mandarin.

For the Standard Arabic ODA for reading, based on the Standard Arabic sample obtained, while the majority of students scored at an ILR level of 2+ on the DLPT5, the majority of students scored at an ILR level of 2 on the ODA. There was variance at Level 3, with 33.3% of students scoring at the target ILR level, 33.3% scoring one ILR level lower, and 25% scoring two ILR levels lower. There was variance for Level 2+, with 55% of students scoring one level lower and 23% of students scoring one level higher than the DLPT5. At Level 2, the DLPT5 and ODA showed the closest relationship and least variance, with 56% of the ODA scores at the target level, and 44% of scores one level lower than the DLPT5. The few scores available at Level 1+ indicated a variance between the DLPT5 and the ODA, with 67% of students scoring one level higher and 33% of students scoring at the target level. Data at Level 1 indicated a variance, with ODA scores one level higher than the DLPT5, but there was not enough data to identify clearer relationship patterns. Figure 57 shows the total number of reading scores at each

ILR level per the ODA and per the DLPT5 for Standard Arabic and Figure 58 shows the relationship between the ODA and the DLPT5 at each ILR level for reading, Standard Arabic.

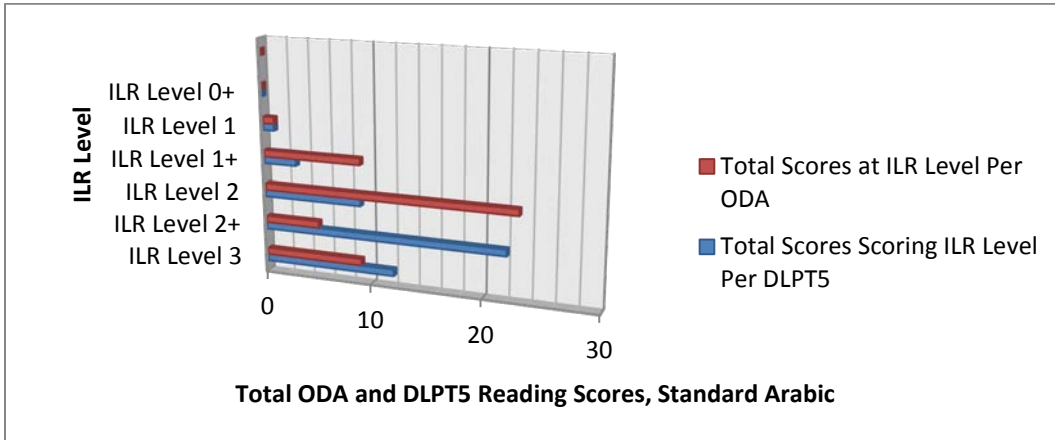


Figure 57. Total ODA and DLPT5 score comparison—Reading Standard Arabic.

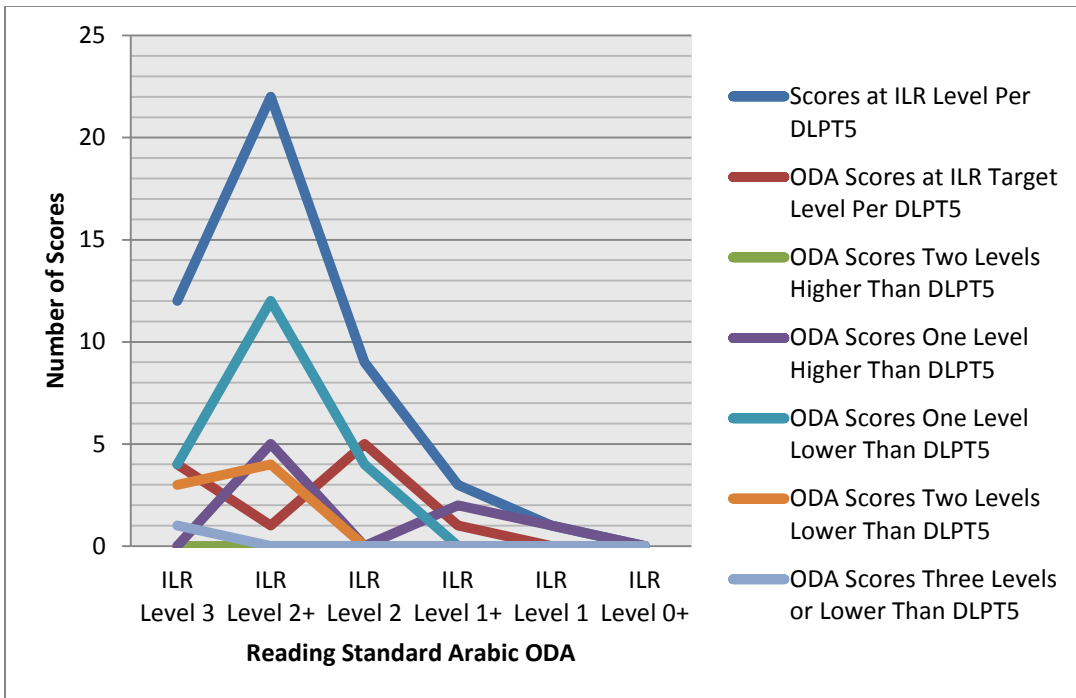


Figure 58. Relationship between the ODA and the DLPT5—Reading Standard Arabic.

Summary

For Research Question 1, a Pearson product–moment correlation for listening indicated a weak correlation for Chinese Mandarin (r value of .20), a moderate correlation for Spanish (r value of .32), a moderate correlation for Korean (r value of .40), and a moderate correlation for Standard Arabic (r value of .56). The correlation was not strong for any of the languages studied, which required an r value of .70 to 1.00 to be considered strong. The Standard Arabic listening ODA test indicated the highest level of correlation to the DLPT5 from all the languages studied. The Chinese Mandarin listening ODA indicated the weakest correlation. A Pearson product–moment correlation for reading indicated a weak correlation for Spanish (r value of .14), a weak correlation for Korean (r value of .23), a moderate correlation for Standard Arabic (r value of .30), and a moderate correlation for Chinese Mandarin (r value of .34). The correlation did not indicate a strong correlation for any of the languages studied, which required an r value of .70 to 1.00 to be considered strong. The Chinese Mandarin ODA for reading indicated the highest correlation compared to the other languages studied, and the Spanish ODA for reading indicated the weakest correlation. According to the student sample based on the total population, the highest level of confidence in the correlation results corresponds to the Spanish sample, with an 82% confidence and a .05 margin of error. The lowest level of confidence in the results corresponded to the Korean sample, with a 39% confidence and a .05 margin of error.

For Research Question 2, data indicated the weakest relationship at the ILR level of 3 and 2+ for listening for all languages studied, with scores one to two levels lower than ILR level per DLPT5. Data for Chinese Mandarin indicated the weakest relationship

across all ILR levels for listening. Conversely, Chinese Mandarin also indicated the closest relationship across all ILR levels for reading compared to the other languages. For listening, Standard Arabic followed by Spanish indicated the closest relationship between the ODA and the DLPT5 across all levels when compared to the other languages. For reading, Chinese Mandarin indicated the closest relationship between the ODA and the DLPT5 across all levels, followed by Standard Arabic. The ODA for reading indicated a pattern of some scores aligning at the target level for Levels 3 and 2+ across all languages, but the listening ODA data did not show any scores at the target level except for a sparse number of scores for Spanish and Chinese Mandarin.

For Research Question 3, for listening, Standard Arabic had the most consistency across levels, with a one level to a target level alignment to the ILR levels except for Levels 3 and 2+. Chinese Mandarin had the highest variance, with scores predominantly two levels followed by one level lower than ILR, followed by Korean, with little discrimination and differentiation of student scores at lower ILR levels. For reading, Chinese Mandarin followed by Standard Arabic had the closest alignment between the ODA and the DLPT5 across all levels and the highest discrimination and score differentiation across ILR levels. At the ILR Level 3 and 2+, there is a closer relationship between the DLPT5 and the ODA at the upper ILR levels for reading compared to listening, with a higher number of scores at Levels 2+ and 3. At these upper levels, Spanish and Chinese Mandarin showed the least discrimination and the least score differentiation between ILR Levels 2+ and 3, with Spanish showing the least discrimination and score differentiation across all levels.

Data indicated that, for all languages studied, the ODA had a closer relationship to the DLPT5 for reading than for listening. Listening aligned predominantly one to two levels lower than DLPT5 at the ILR Levels 3 and 2+. Using Krejcie and Morgan's (1970) formula for student sampling based on the student sample size for each language, several conclusions emerged: (a) the Spanish sample had a 82% level of confidence, (b) the Korean sample had a 49% level of confidence, (c) the Chinese Mandarin sample had a 61% level of confidence, and (d) the Standard Arabic sample had a 54% level of confidence; all levels of confidence had a .05 margin of error.

CHAPTER V: FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

Since September 11, 2001, the DoD has been its own main supplier of foreign language resources to respond appropriately to changing world situations that have fueled an increasing demand for language capabilities. The literature showed that the U.S. government has played a key role in developing standards and accreditation measures for second language acquisition in the United States. In this context, this study involved exploring a technological contribution to education made by DLIFLC in the formative assessment field through the ODA. The studies on second language acquisition online diagnostic assessments are primarily based on the European DIALANG (Clark et al., 2014, Taghizadeh et al., 2014), an online diagnostic test based on the CEFR used by over 12,000 students (Lancaster University) mostly in Europe (Alderson & Huhta, 2011). Although researchers have noted a true foreign language diagnostic test does not exist except for DIALANG (Alderson, 2005; Alderson & Huhta, 2005, 2011; Huhta, 2008), this online diagnostic assessment provides relatively limited diagnostic value because it was designed based on traditional concepts of language use rather than on a theory of foreign language acquisition and use (Alderson & Huhta, 2011). By contrast, the ODA employs the ACTFL criteria and the ILR standards, and over 35,000 users take it each year. Although researchers know about the DLIFLC predictive test DLAB and the summative DLPT5 through published research studies, little is known about the properties of the ODA as a formative diagnostic test through published correlation or validation studies.

Literature indicated a disconnect exists between theory and practice when looking at formative and summative assessments in a more integrated manner, and limited

research addressed the correlation between formative assessments and summative assessments (Crooks, 2011; Croteau, 2014; Knight, 2000; Taras, 2005). The current study contributes to research literature by (a) integrating the ODA to the body of research on online diagnostic assessments in second language acquisition, (b) assessing the correlation of the formative ODA to the summative DLPT5 to assess validity, and (c) incorporating the ODA to the body of research associated with the correlation of formative and summative tests.

Purpose Statement

The purpose of this nonexperimental correlational study was to identify the relationship between online formative (ODA) and summative (DLPT5) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success in a Basic Course program for adult students at the DLIFLC.

Research Questions

1. What is the relationship between the Spanish, Korean, Chinese Mandarin, and Standard Arabic ODA formative test results administered at the end of the course and students' final summative DLPT5 scores?
2. What is the relationship between the ODA and the ILR levels for Spanish, Korean, Chinese Mandarin, and Standard Arabic as measured by the DLPT5?
3. Are the relationships found between ODA and DLPT5 for Spanish, Korean, Chinese Mandarin, and Standard Arabic consistent across the levels or is there variance in the relationship depending on the level?

Research Methods and Data Collection Procedures

The nonexperimental study included a standard regression model to determine the relationships between two variables: (a) end-of-course ODA scores and (b) DLPT5 final scores. The study involved performing several statistical analysis tests to identify correlations between ODA scores and DLPT5 final scores using a multiple regression analysis. The data collection instruments used in this research study consisted of archived data from eight formative ODA and eight summative DLPT5 assessments developed by DLIFLC:

- Archival scores for listening and reading from students who took the formative ODA at the end of the 36-week course in Spanish and archival scores of the same students who took the DLPT5 at the end of this program.
- Archival scores for listening and reading from students who participated in a formative ODA at the end of the 64-week course in Korean, Chinese Mandarin, and Standard Arabic and archival scores of the same students who took the summative DLPT5 at the end of this program.

Population

Each calendar year, approximately 3,500 students attend the Basic Course programs available at the DLIFLC for 17 languages (DLIFLC, 2015c). For the languages studied, the total population in 2015 and 2016 at the Basic Course program consisted of 342 students for Spanish, 426 students for Korean, 571 students for Chinese Mandarin, and 912 students for Standard Arabic.

Sample

Two hundred sixty-nine listening archived scores and 270 reading archived scores from 276 students for four languages represented 7.7% of the total population in 1 year. These scores also represented 35% of the total Spanish school population, 8% of the total Korean school population, 12% of the total Chinese Mandarin school population, and 6% of the total Standard Arabic school population in 2015 and 2016.

Major Findings

Finding 1: Research Found Evidence of ODA Content Validation Procedures

The literature review of the content development and validation process of the ODA (Chapter II, Appendix B) indicated that this online diagnostic tool provides substantiated documentation regarding the ODA standardized procedures for the development of items and stimuli, as well as for their quality control and validation procedures. It also showed evidence that the ODA generates diagnostic profiles and provides individualized diagnostic information. This information helps to identify the specific areas of strength and growth that allow a second language learner to acquire the skills at the next level of language proficiency. Literature research also indicated that the ODA follows standardized development and quality control procedures consistent with assessment literacy standards to develop formative assessment materials, along with the correct application of protocols that ensure the validity, reliability, and fairness of an assessment instrument. Additional research is necessary to verify content validity, which was not studied in this research. An essential aspect of the content validity for well-designed online diagnostic tests after items and testlets become operational is monitoring items. The ODA database includes a feature labeled “item–user correlation” and data

statistics that help identify the level of discrimination between items and testlets across levels as well as the validation of all possible correct answers for open-ended items. Through this monitoring process, some items may be replaced or updated because content may have become outdated, societal and cultural exposure to certain content may elicit prior knowledge responses over time, items may not provide the expected outcomes, or a need arises to develop new content on an area or skill where gaps exist (DLIFLC, 2015b). This research was not able to verify the content validity of content of ODA or the item-to-item correlation and item–user correlation feature of the ODA client side. Evidence of data or statistical information resulting from the item-to-item correlation and item–user correlation may further enhance the content validity of the ODA.

Finding 2: Evidence of Irregular ODA Administrations at the Basic Course

The importance of delivering diagnostic information with areas of strength and growth cannot be underestimated. Although it was not the intent in this study to address how instructors’ perceptions may affect the implementation and impact of an assessment, it is important to recognize instructors’ essential contribution to the success of an assessment (Fox, 2009; Jang, 2005, 2009). In this context, it is relevant to recognize that ODA archived data received compared with total student population in 2015 and 2016 indicated that of the languages studied, the ODA has different degrees of regularity in administration, with some schools administering the ODA to a large extent and others to a smaller extent.² The effectiveness of a formative assessment depends on the successful

² Archived data received by DCSIT indicate the possibility that there might be a higher number of ODA administrations, but some students may have written incomplete names during ODA enrollment.

implementation of the formative test results into relevant instruction (Frohbeiter et al., 2011; S. McManus, 2008; Pellegrino, 2014). Therefore, the effectiveness of the ODA through the successful implementation of formative results into relevant instruction needs further study.

Finding 3: Evidence of Moderate or Low Correlations to the DLPT5

For Research Question 1 for listening, a Pearson product–moment correlation showed a weak correlation for Chinese Mandarin (r value of .20), a moderate correlation for Spanish (r value of .32), a moderate correlation for Korean (r value of .40), and a moderate correlation for Standard Arabic (r value of .56). The listening correlation did not indicate a strong correlation for any of the languages studied, which required an r value of .70 to 1.00 to be considered strong. For listening, the Standard Arabic ODA test indicated the highest correlation to the DLPT5 compared to the other languages studied. The Chinese Mandarin ODA listening indicated the weakest correlation to the DLPT5 compared to the other languages studied.

For Research Question 1 for reading, a Pearson product–moment correlation indicated a weak correlation for Spanish (r value of .14), a weak correlation for Korean (r value of .23), a moderate correlation for Standard Arabic (r value of .30), and a moderate correlation for Chinese Mandarin (r value of .34). The correlation did not indicate a strong correlation for any of the languages studied, which required an r value of .70 to 1.00 to be considered strong. For reading, Chinese Mandarin had the strongest correlation to the DLPT5 compared to the other languages studied and indicated the weakest correlation to the DLPT5. Tables 25 and 26 show the correlation results for listening and for reading.

Table 25

Correlation Results for Listening

Listening	Correlation	Strength of the relationship
Spanish	0.32	Moderate
Korean	0.40	Moderate
Chinese Mandarin	0.20	Weak
Standard Arabic	0.56	Moderate

Table 26

Correlation Results for Reading

Language	Correlation	Strength of the relationship
Spanish	0.14	Weak
Korean	0.23	Weak
Chinese Mandarin	0.34	Moderate
Standard Arabic	0.30	Moderate

Finding 4: Evidence of Weak Relationship to the ILR Levels Across All Languages for Listening

For Research Question 2, a Pearson product–moment correlation and an analysis of the ODA score distribution of ILR scores per DLPT5 indicated the weakest relationship to the ILR at Level 3 and Level 2+ for all languages for listening, with scores one to two levels lower than ILR level per DLPT5. While the ODA for reading indicated a pattern of scores at the target level for Levels 3 and 2+ across all languages, the listening ODA data showed a sparse to nonexistent occurrence of scores at the target level for these levels. Data for Chinese Mandarin indicated the weakest relationship to the ILR levels for listening compared to the other languages. Conversely, Chinese Mandarin also showed the closest relationship to the ILR levels for reading compared to the other languages studied. For listening, Standard Arabic indicated the closest relationship to the ILR levels compared to the other languages. For reading, Chinese Mandarin had the

closest relationship to the ILR levels compared to the other languages, followed by Standard Arabic. While the ODA for reading indicated a pattern of some scores aligning at the target level for Levels 3 and 2+ across all languages, the listening ODA data did not show any scores at the target level except for a sparse number of scores for Spanish and Chinese Mandarin. This research increased the confidence level in its results, particularly for Listening, because all languages, regardless of the sample size, showed a consistent pattern at levels 3 and 2+, with scores one to two levels lower than ILR level per DLPT5. Tables 27 and 28 show the listening and reading relationship to the ILR levels according to the DLPT5.

Table 27

ODA Listening Relationship to the ILR Levels per DLPT5

Listening	Spanish	Korean	Chinese Mandarin	Standard Arabic
ILR Level 3	Weak	Weak	Weak	Weak
ILR Level 2+	Weak	Weak ^a	Weak	Weak
ILR Level 2	Moderate	Moderate	Moderate to weak	Moderate
ILR Level 1+	Moderate	Strong ^a	Weak ^a	Moderate to weak ^a
ILR Level 1	Moderate	Moderate ^a	N/A	Weak
ILR Level 0+	N/A	N/A	N/A	Strong ^a

^aNot enough scores to identify clear ILR relationship trends.

Table 28

ODA Reading Relationship to the ILR Levels per DLPT5

Reading	Spanish	Korean	Chinese Mandarin	Standard Arabic
ILR Level 3	Moderate to weak	Weak	Moderate	Weak
ILR Level 2+	Moderate	Weak	Moderate	Moderate to weak
ILR Level 2	Weak	Moderate	Moderate	Moderate
ILR Level 1+	Weak ^a	Moderate ^a	Weak ^a	Moderate ^a
ILR Level 1	N/A	N/A	Weak ^a	Moderate ^a
ILR Level 0+	N/A	N/A	N/A	N/A

^aNot enough scores to identify clear ILR relationship trends.

Finding 5: Evidence of Variance in the Relationship to the DLPT5 Depending on the Language and Depending on the Level

For Research Question 3, according to the data obtained, the relationship between the DLPT5 and the ODA showed a variance depending on the language and depending on the level. However, the variance seems to have a degree of consistency across languages. For listening, ODA scores consistently fell lower than DLPT5 scores at ILR Levels 3 and 2+ (one to two levels lower than the DLPT5), with Standard Arabic at a higher degree of variance (two to three levels lower than the DLPT5). For reading, ODA scores consistently aligned at the target level to one level lower at ILR Level 3 with the exception of Korean (two to one ILR level lower). For listening, taking the variance at Levels 3 and 2+ into account, Standard Arabic had the closest relationship between the ODA and the DLPT5 of all languages studied, with a higher level of discrimination and a more defined level of differentiation. For listening, Chinese Mandarin had the highest variance, followed by Korean, with little discrimination and differentiation of student scores at lower ILR levels. For reading, Chinese Mandarin had the closest alignment between the ODA and the DLPT5 across all levels and the highest discrimination and score differentiation across ILR levels. At the ILR Level 3 and 2+ for reading, there was a closer relationship between the DLPT5 and the ODA at the upper ILR levels compared to listening, with a higher number of scores at Levels 2+ and 3. Data indicated that, for all languages studied, the ODA had a closer relationship to the DLPT5 for reading than for listening. Listening aligned predominantly one to two levels lower than DLPT5 at the ILR Levels 3 and 2+ with a high number of scores two ILR levels down. Figure 59 shows the predominant ILR listening levels on the ODA per DLPT5, and Figure 60 shows the

predominant ILR reading levels on the ODA per DLPT5. Tables 29 and 30 show the predominant scores on the ODA at each specific ILR level. The first column shows the ILR levels per DLPT5 results.

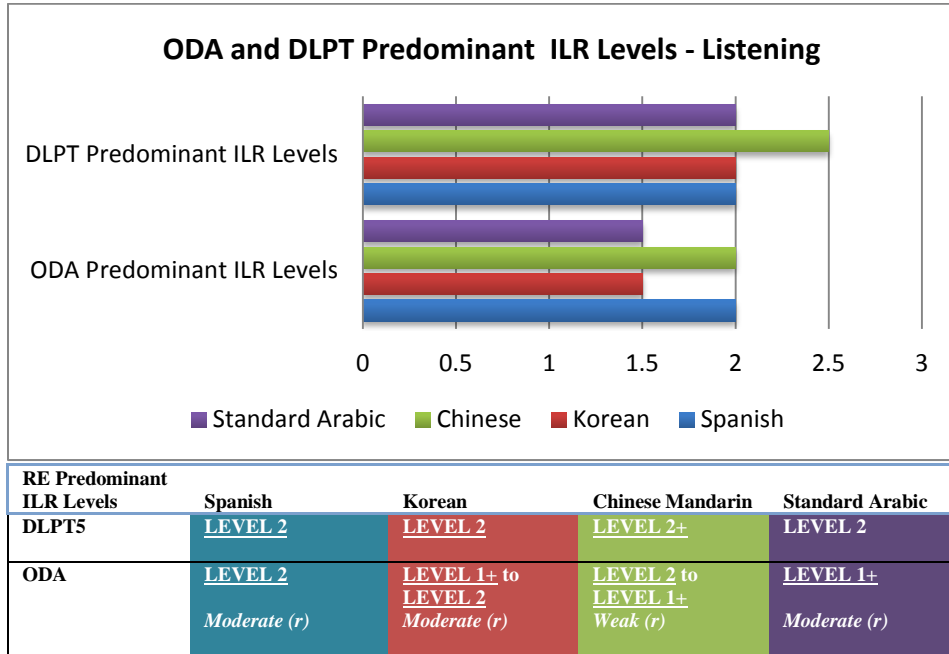


Figure 59. Predominant ILR listening levels on the ODA per DLPT5.

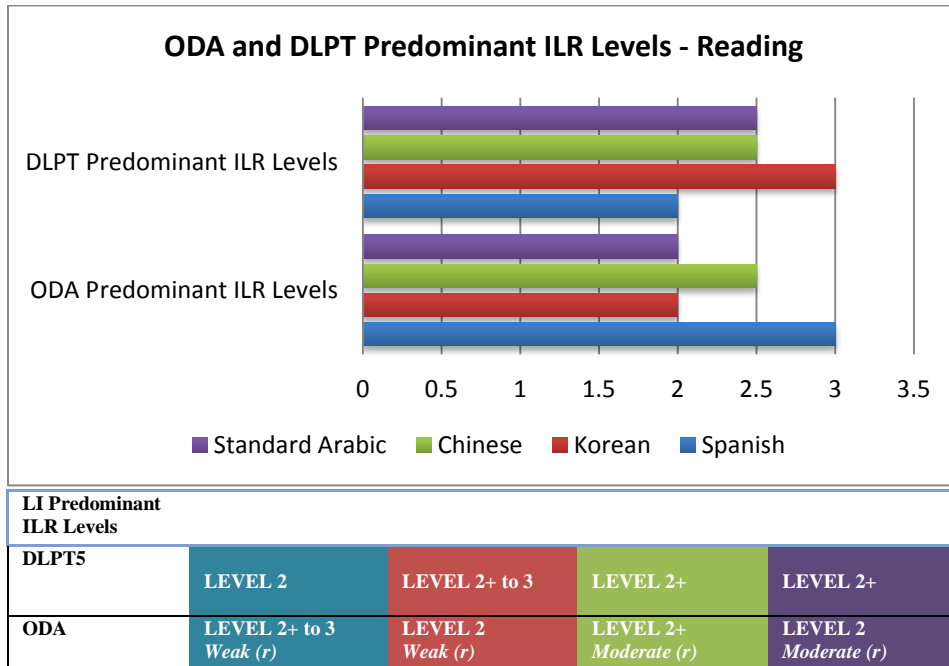


Figure 60. Predominant ILR reading levels on the ODA per DLPT5.

Table 29

Predominant ILR Listening Levels on the ODA per the DLPT5

ILR levels per DLPT5	Spanish Level 2 Moderate (<i>r</i>)	Korean Level 1+ to 2 Moderate (<i>r</i>)	Chinese Mandarin Level 2 to 1+ Weak (<i>r</i>)	Standard Arabic Level 1+ Moderate (<i>r</i>)
ILR Level 3	Level 2+ to 2	Level 2	Level 2	Level 2 to 1+
ILR Level 2+	Level 2 to 1+	Level 2	Level 2 to 1+	Level 1+
ILR Level 2	Level 2 to 1+	Level 1+ to 2	Level 2 to 1+	Level 1+
ILR Level 1+	Level 2 to 1+	Level 1+	Level 2 to 1+	Level 1+ to 0+ ^a
ILR Level 1	Level 1+	Level 1+	N/A	Level 1+ to 1
ILR Level 0+	N/A	N/A	N/A	0+ ^a

^aNot enough scores to identify clear ILR relationship trends.

Table 30

Predominant ILR Reading Levels on the ODA per the DLPT5

ILR levels per DLPT5	Spanish Level 2+ to 3 Weak (<i>r</i>)	Korean Level 2 Weak (<i>r</i>)	Chinese Mandarin Level 2+ Moderate (<i>r</i>)	Standard Arabic Level 2 Moderate (<i>r</i>)
ILR Level 3	Level 3 to 2 to 2 +	Level 2 to 2+	Level 2+ to 3	Level 3 to 2+ to 2
ILR Level 2+	Level 3 to 2 to 2 +	Level 2	Level 2+ to 3	Level 2
ILR Level 2	Level 3 to 2 to 2 +	Level 2 to 1+	Level 2 to 2+	Level 1+ to 2
ILR Level 1+	Level 1+ to 2+ ^a	Level 1+ ^a	Level 2+ to 1+ ^a	Level 1+ to 2
ILR Level 1	N/A	N/A	Level 2 ^a	Level 1+
ILR Level 0+	N/A	N/A	N/A	N/A

^aNot enough scores to identify clear ILR relationship trends.

Finding 6: Evidence of the Closest Relationship to the DLPT5 at ILR Level 2

For Research Question 3, data indicated that both reading and listening had the closest relationship to the ODA and the DLPT5 at Level 2 for all languages studied.

Therefore, it is possible to devise assessments with dissimilar design constructs—

formative and summative—but common ILR requirements that, if designed

appropriately, lead to comparable ILR results. The closest relationship observed between

the DLPT5 and the ODA at ILR Level 2 could be meaningful. However, a high

correlation at a given ILR level needs to be further assessed in the context of the

correlation to all other ILR levels. For example, the Spanish ODA for reading had a closer relationship at ILR Level 2. However, this high correlation did not necessarily result in a high correlation to the DLPT5. This is because for the Spanish ODA, many other students also scored at Level 2 on the ODA while receiving a different ILR level on the DLPT5. For this reason, the relationship between the DLPT5 and the ODA at Level 2 or at any other ILR level needs to be assessed in the context of the specific language studied and in the context of the correlation to the rest of the ILR levels.

Unexpected Findings

Unexpected Finding 1: Higher Discrimination in Category IV Languages

The researcher estimated that Spanish, a Category I language, might have a higher correlation and a higher level of discrimination across ILR levels compared to the Category IV languages studied. However, the majority of students who took the ODA Spanish reading test scored predominantly at ILR Level 2+ or Level 3, even though their scores on the DLPT5 might have ranged across different ILR levels, predominantly Level 2. Conversely, Category IV languages showed higher correlation and discrimination and a more delineated continuum across levels.

Unexpected Finding 2: Regular ODA Administrations and Higher Sample Size Does Not Necessarily Lead to a Higher Correlation

Of equal interest was the finding that a larger sample size and a higher level of regularity in the ODA administration did not necessarily result in a higher level of consistency in the ODA and DLPT5 correlation results. While the Spanish Basic Course administered the ODA more frequently and consistently than the other languages studied

at 3 months to 1 week before the DLPT5 administration, the correlation results did not necessarily lead to a higher number of closely correlated ODA and DLPT5 scores.

Conclusions

Conclusion 1: Irregular Administrations Hinder the Full Diagnostic

Potential of the ODA

The purpose of this nonexperimental correlational study was to identify the relationship between online formative (ODA) and summative (DLPT5) assessments in foreign language instruction in Spanish, Korean, Chinese Mandarin, and Standard Arabic to determine their relationship to student success in a basic course program for adult students at the DLIFLC. Although it was not the intent in this study to address how the instructors' perceptions of the ODA affect the full implementation or success in the application of diagnostic information resulting from this instrument, it is relevant to recognize that ODA archived data received compared with total student population in 2015 and 2016 indicated that, of the languages studied, the ODA has different degrees of regularity in administration, with some schools administering the ODA to a large extent and others to a smaller extent.³ Because literature indicated that the effectiveness of a formative assessment depends on the successful implementation of the formative test results into relevant instruction (Frohbeiter et al., 2011; S. McManus, 2008; Pellegrino, 2014), the inconsistent administration of the ODA at specific phases of the language course for some of the languages studied might hinder the full potential of this diagnostic

³ Archived data received by DCSIT indicate the possibility of a higher number of ODA administrations, but some students may have written incomplete names during ODA enrollment.

instrument. Additional research may be necessary to verify the consistency of ODA administrations.

Conclusion 2: While Irregular Administrations Hinder the ODA's Full Diagnostic Potential, Regular Administrations Do Not Necessarily Lead to Comparable ODA and DLPT5 Scores or a Closer Correlation

Based on the analysis of archived data, and with regard to Research Question 1, a Pearson product–moment correlation (r) for listening indicated a moderate correlation existed for Spanish, Korean, and Standard Arabic and a weak correlation existed for Chinese Mandarin. A Pearson product–moment correlation (r) for reading indicated a moderate correlation for Chinese Mandarin and Standard Arabic and a weak correlation for Spanish and Korean. The confidence for these results per Krejcie and Morgan's formula to estimate confidence in results given the size of the sample sizes is 82% confidence for Spanish, 49% for Korean; 61% for Chinese Mandarin; and 54% for Standard Arabic with a .05 margin of error. The researcher found that the consistency in ODA administrations did not necessarily lead to comparable ODA and DLPT5 scores or a closer correlation, as in the case of the Spanish ODA, which represented the language studied with the most regular ODA administrations, as well as the largest sample size. Whereas the Spanish correlation represented the results with the highest level of confidence (82%), the regularity in the administration of the ODA for Spanish did not result in a higher correlation between the ODA and the DLPT5. In fact, the ODA for Spanish reading represented the lowest correlation of all languages and content areas studied (correlation of .14). Therefore, the resulting conclusion is that the low correlation

of the ODA for Spanish is not the result of an irregular administration, but of other factors that need further study.

Conclusion 3: Higher Correlation of Category IV Languages Over a Category I Language Might Be the Result of an Intrinsic Test-Taking Advantage for Category I Test Takers

With regard to Research Question 1, the research indicated that Spanish, as a Category I language, had a moderate relationship to the DLPT5 for listening and a low relationship to the DLPT5 for reading. While the majority of students who took the ODA Spanish reading test scored predominantly at ILR Level 2+ or Level 3, scores on the DLPT5 might have ranged across several ILR levels, predominantly Level 2. Conversely, the Category IV languages studied showed higher correlation to the DLPT5, higher discrimination, and a more delineated score differentiation across ILR levels. For reading, the differential functioning of items for Category I languages versus Category IV languages might be the result of an intrinsic test-taking advantage for Category I test takers. Category I languages might lead to intrinsic test-taking advantages for test takers whose first language is English when required to write open-ended responses in their native language. The test-taking advantages for Category I test takers might include the use of the same Roman or Latin alphabet for writing extended responses in the native language, as well as the number of cognates between Category I languages and the natively used English language.

Conclusion 4: Because ODA Levels 3 and 2+ Are Difficult to Reach, Students Who Reach These Levels Are Very Likely to Be Ready for the DLPT5

For Research Question 2, data organized by ILR level compared to regression analysis indicated a variance across all languages at the ILR level of 3 and 2+ for listening, with scores one to two levels lower than ILR level per the DLPT5 and a high number of scores two ILR levels down. Although the ODA for reading indicated a pattern of some scores aligning at the target level for Levels 3 and 2+ across all languages, the listening ODA data showed sparse scores at these upper levels. The reading and listening data indicated that the ODA is a difficult test across all languages, particularly for listening. With the exception of the Category I language studied, because ILR Levels 2+ and 3 are difficult to reach on the ODA, students who can effectively reach ILR Levels 2+ or 3 on the ODA are very likely to reach the desired 2+ level on the DLPT5.

Conclusion 5: It Is Possible to Devise Assessments With Dissimilar Design Constructs—Formative and Summative—but Common ILR Requirements That, if Designed Appropriately, Lead to Comparable ILR Results

For Research Question 2, data for Chinese Mandarin indicated the weakest relationship across all ILR levels for listening. Conversely, Chinese Mandarin also showed the closest relationship across all ILR levels for reading. For listening, considering the high difficulty at the upper levels, Standard Arabic had the closest relationship between the ODA and the DLPT5 across all levels. For reading, Chinese Mandarin had the closest relationship between the ODA and the DLPT5 across all levels, followed by Standard Arabic. Because the literature review revealed a disconnect

between theory and practice when looking at formative and summative assessments in an integrated manner, the findings from this research are meaningful. At least one test showed a higher degree of correlation and score differentiation for reading (the Chinese Mandarin ODA followed by the Standard Arabic ODA), and at least one test showed a higher degree of correlation and score differentiation for listening (the Standard Arabic ODA). Therefore, it is possible to devise assessments with dissimilar design constructs—formative and summative—but common ILR requirements that, if designed appropriately, lead to comparable ILR results. These findings have the potential to not only validate additional ODA assessments but also confirm the validation procedures established for the DLPT5.

Conclusion 6: The ODA Learning Progression Design, the Logarithmic Function, and the ODA Test Design May Contribute to the ODA Variance in the Relationship to the DLPT5

Data obtained for Research Question 3 indicated the relationship between the DLPT5 and the ODA showed a variance depending on the language and depending on the level. Data indicated that for all languages studied, when looking at the predominant ODA scores, the ODA had a closer relationship to the DLPT5 for reading than for listening (listening fell one to two levels lower than DLPT5 at Levels 3 and 2+). Both reading and listening had the closest relationship between the ODA and the DLPT5 at Level 2 for all languages studied. Category IV languages had a higher correlation and ILR score differentiation than a Category I language. Therefore, there might be technical factors, content development factors, and intrinsic differences between Category I and Category IV languages that contribute to higher degrees of a DLPT5–ODA correlation on

a given ODA test, and the differences in correlation might be the result of (a) the ODA learning progression design needing an adjustment; (b) a test design construct that contains stimuli needing an adjustment particularly for listening; (c) a logarithmic function that needs to be adjusted, particularly at the upper levels for listening; and (d) an intrinsic difference between Category I and Category IV languages that requires a differentiated test design depending on the language for which a reduced number of extended response items and a higher number of multiple choice items for Category I languages may be more appropriate.

Conclusion 7: Although Data Indicated Low or Moderate Correlations of Varying Degrees for All Languages, This Study Did Not Find Any Evidence of Strong Correlations

Although the results from this study indicated varying degrees of correlation between the ODA and the DLPT5, ranging from low to moderate, none of the results showed evidence of strong correlations for any of the languages studied, which required an r value of .70 to 1.00 to be considered strong. This might be the result of an intrinsic difference between formative, open-ended, classroom-based assessments and summative, multiple-choice assessments that is predictable and expected. A projection correlation study is appropriate for assessment instruments that have a common framework but varied tasks, testing conditions, or purposes or are conducive to a different level of student motivation (Mislevy, 1992). The DLPT5 and the ODA meet these characteristics. Although these two assessments have assessment construct goals in common, they have different tasks, testing conditions, and differences in outcomes because of their respective summative and formative characteristics. This study could not implement a projection

correlation because of the limited sample available, along with the very limited studies available on the practical application of projection correlation studies.

Implications for Action

What are the implications for the ODA for listening, which indicates a more defined ILR level differentiation for listening in Standard Arabic, and for reading in Chinese Mandarin ? What can the DLIFLC do with the knowledge that scores fell one to two levels lower or even more across all listening languages at ILR Levels 3 and 2+ or that a high differentiation and discrimination is observed on Category IV languages such as Chinese Mandarin followed by Standard Arabic, while a Category I language shows fewer degrees of discrimination and score differentiation across all levels?

Implication 1: DLPT5 Validation Procedures

DLIFLC has been at the forefront in the implementation of innovative processes that ensure the increased validation of foreign language assessment instruments, as in the DLPT5. With the development of the DLPT5, innovative methods were introduced to ensure greater validity and calibration procedures. These procedures included the configuration of standard-setting panels for setting DLPT5 cut scores. As part of the DLPT5 validation, a panel of ILR experts from different languages ensured ILR consistency during the item development process. Each passage and item went through an independent review by the Proficiency Standards division to ensure a consistent interpretation of the ILR performance-level descriptors across languages. Because this study did not involve analyzing the specific stimuli and item development content of the ODA, and because of the ILR variance at Level 3 and 2+ for listening, ODA developers need to review stimuli and open-ended items, particularly at the upper levels, including

independent reviewers from DLIFLC instructors and language experts from universities across the United States. ODA leaders and developers need to select a panel of ILR proficiency standards experts that ensures consistency in ODA ILR levels at the language level and across languages. This requires that each passage and item go through an independent review by this panel to ensure a consistent interpretation of the ILR performance-level descriptors across languages.

After the panel of ILR experts completes and verifies ODA test development, a pre-standard-setting and a standard-setting panel need to be introduced, as was introduced to the DLPT5, with ILR experts from different languages participating in the process of interpreting the ILR performance-level descriptors in the context of the specific requirements of the ODA. The pre-standard-setting panel will ensure the ILR levels are implemented in a more systematic way across all languages and content areas. Lastly, a standard-setting phase with ODA test scores available from the database is needed as a crucial step in the validation process. A standard-setting phase should include standardized procedures that use the ILR performance-level descriptor statements in a clearly organized and categorized process across languages, as well as examples of student responses to ensure a clearly differentiated level of discrimination among different ILR levels and student scores to ensure greater validity of the ODA.

Implication 2: Correlation Studies

DLIFLC leaders need to develop future DLPT5–ODA correlation studies for all ODA tests as part of the standardized validation procedures for the ODA. This common strategy used for primary language academic assessments validates the content of lower stakes assessments of a formative nature and is incorporated into the test design, item

selection, and content validation process by correlating the results of low-stakes tests to high-stakes summative tests. Because of the high-stakes nature of the DLPT5, which has gone through a strenuous and highly monitored process of item selection, standardization, and validation, a DLPT5–ODA correlation introduced as part of the standardized procedures for validating the ODA ensures the ODA leverages from the extent of the DLPT5 validation by incorporating a systemic correlation procedure for appropriate ILR level verification.

Implication 3: Leveraging From ODA’s Internal Assets

ODA developers need to meet and study each other’s ODA content and test design to identify the content development factors that contributed to higher degrees of DLPT5–ODA correlation on a given ODA test and assess if (a) the differences in correlation might be the result of closer or farther alignment to the ILR specifications at each level; (b) there are test-taking advantages for test takers when writing open-ended responses in English for a Category I language versus a Category IV language; and (c) the ODA authoring system’s settings might lead to a higher level of content difficulty at the upper levels of listening and might result in an ILR variance when compared to the DLPT5. Developers should consider if the variance in correlation at the upper levels of listening is (a) the result of the ODA learning progressions design; (b) a test design construct with listening stimuli at a higher level of difficulty than ILR specifications; or (c) an intrinsic difference between formative, open-ended classroom-based assessments and summative, multiple-choice assessments that is predictable and expected.

Implication 4: Practical Implementation of the ODA in the Classroom

DLIFLC leaders need to study the factors that might be hindering the full implementation of the ODA results into applicable instructional strategies in the classroom. The leaders should perform usability studies of the ODA individual diagnostic profile information to address the level of buy-in of DLIFLC instructors toward the complete implementation of the ODA as a tool that contributes to student success and mastery of a secondary language. ODA developers should develop ODA guides for instructors and ODA administrators, as well as ancillary materials that could include DVDs and manuals on the use of the ODA, as well as training for instructors and students on the appropriate interpretation of individual diagnostic profiles. ODA administration manuals and DVDs should also include recommendations for practical applications of student results into appropriate instruction. It might be worth considering the applicability of issuing ODA profiles for instructors in addition to the profiles already available for students.

Implication 5: Preparation for Success at the Upper Levels

For listening, at Levels 3 and 2+, students score one or two levels lower than the DLPT5. Additional studies are necessary to identify if this difference is the result of the intrinsic difference between formative and summative assessments. These results have great potential for action, as they assure students, instructors, test developers, and the DLIFC that the listening ODA was designed at a high level of content difficulty at the upper levels, which could be more realistic and cost-effective to make adjustments if necessary. Scores at levels 3 and 2+ also reassure the institution leaders, tests takers, and instructors about the ODA results for those students who were able to reach Levels 2+

and 3 on the ODA listening and Level 3 on the ODA for reading. For these students, the likely chances of comparability to similar DLPT5 scores are high. Because research results showed the closest relationship at ILR Level 2, these results could also be meaningful. However, ILR Level 2 results do not necessarily ensure a correspondence to the DLPT5 because many other students also scored a Level 2 on the ODA while receiving a different ILR level on the DLPT5, particularly for the reading Spanish ODA.

Implication 6: Validation of Chinese Mandarin for Reading and Standard Arabic for Listening

For a diagnostic test that has never been validated or correlated before, it is remarkable that while following a different test design construct with formative characteristics different from the DLPT5, the ODA demonstrated higher levels of correlation with the DLPT5 for the Chinese Mandarin ODA for reading and the Standard Arabic ODA for listening. The ODA design follows many of the recommended features for foreign language diagnostic instruments and meets many of the requirements suggested for online diagnostic assessment and instruction. Instructors, schools, students, and DLIFLC leaders need to use the ODA results to inform instruction not only at the beginning of the school program but also during the last semester to identify if students have reached expected levels of 2+ in listening and 2+ in reading. Because these levels are difficult to reach on the ODA, students who can effectively reach ILR Levels 2+ or 3 on the ODA are very likely to be ready for the DLPT5. At these upper levels, meaningful instructional strategies for students who are unable to reach upper ODA levels of listening test difficulty should be implemented during the last semester of the school program to ensure appropriate DLPT5 graduation scores are achieved with relevant

instruction. Additionally, instructors, schools, and students should leverage from the results of this study, particularly for the languages that showed the closest relationship across all ILR levels, such as Chinese Mandarin for reading and Standard Arabic for listening, which indicated the closest relationship between the ODA and the DLPT5 across all levels. However, in the case of Standard Arabic, the upper-level difficulty for ILR Levels 3 and 2+ needs to receive consideration.

Recommendations for Further Research

Recommendation 1: Validation of the ODA for the Intermediate and Advanced Instructional Programs

The researcher recommends conducting ODA validation studies to evaluate the effectiveness of the ODA at these levels and to determine if there is a need to develop alternate forms, particularly given the fact that students take the ODA and the DLPT5 several times during their military career. Originally developed to address the language maintenance and enhancement needs of military staff who had already graduated from the Monterey Basic Course program (nonresident linguists), the ODA has grown to support the formative diagnostic requirements of DLIFLC resident students as well as nonresident students at the basic, intermediate, and advanced levels. The focus of this study was on the correlation of the ODA in the context of the Basic Course program, and the study does not include any insight into validating the ODA at the intermediate and advanced levels. The ODA and DLPT5 correlation might vary at the intermediate and advanced instructional programs because of the possible familiarity of students with the DLPT5 or the ODA.

Recommendation 2: Applicability of the ODA Into Appropriate Instruction

The literature indicated that the effectiveness of a formative assessment depends on the successful implementation of the formative test results into relevant instruction (Frohbeiter et al., 2011; S. McManus, 2008; Pellegrino, 2014). The researcher therefore recommends a study of instructors' and students' perceptions of the ODA that address the factors that might be hindering the full implementation of the ODA results into applicable instructional strategies. Included in this research, the researcher recommends studying the perceptions of the usability of the ODA individual diagnostic profile information, including its practical implementation into instructional activities. From the implementation perspective, such a study should include a survey on the level of understanding of the test administration sections, diagnostic profiles, and features of the ODA, as well as the level of buy-in of DLIFLC instructors toward the complete implementation of the ODA as a tool that contributes to student success and mastery of a secondary language.

Recommendation 3: Analysis of Variance at Level 3 and 2+ for Listening

According to the archived data available, while the Spanish Basic Course instructors administered the ODA more frequently and consistently than the instructors of the other languages studied at the end of the course program, the consistency in ODA administration did not necessarily lead to comparable ODA and DLPT5 scores or a closer correlation. The researcher therefore recommends future studies on the relationship between consistent ODA administration and instruction and the rate of student success, including the study of specific factors that could have contributed to the variance in correlation. Such factors include (a) open-ended responses written in the English native

language; (b) the characteristics of the ODA semiadaptive features; (c) testing times; (d) idiosyncratic differences between formative, classroom-based assessments and summative, large-scale assessments; and (e) idiosyncratic requirements specific to the measurement of second language skills in listening, which include speed rate of listening stimuli, the length of the recordings, quality of recordings, accents, and cognitive skills involved in short- and long-term memory (Buck, 2011).

Recommendation 4: Effect of Open-Ended Responses on Second Language

Acquisition Tests

The researcher recommends further studies on the factors that account for the variance in the correlation for listening, particularly at the upper levels, considering the relationship to (a) open-ended responses written in the English native language; (b) characteristics of the semiadaptive features; (c) testing times; (d) idiosyncratic differences between formative, classroom-based assessments and summative, large-scale assessments; and (e) idiosyncratic differences between listening and reading second language assessments.

Recommendation 5: Study of Cultural Factors That Affect Predictability of Assessment Constructs

In the context of writing questionnaires, Turner (1993) cautioned about the cultural background of a second language learner as a factor that may have an effect on the responses obtained, which may lead to inordinate response distributions. The researcher recommends future studies related to cultural factors that may defer the predictability of expected responses in second language assessment such as the ODA, including a projection correlation study for assessment instruments that have varied tasks,

testing conditions, or purposes or are conducive to a different level of student motivation, as with the DLPT5 and the ODA. This study could identify if the variance in listening at the upper levels might be the result of a needed projection correlation adjustment also appropriate to address cultural background differences. Within this context, other studies to consider include nonparametric statistics or distribution free tests that are often recommended for second language correlations to account for dissimilar characteristics and inordinate response distributions.

Recommendation 6: Correlation Studies Using the Low-Range DLPT5 and ODA

Because of the sparse scores available at the lower ILR levels, additional studies are recommended to verify variances in correlation with a larger pool of students scoring at Levels 0+, 1, and 1+. It is unlikely that DLPT5 and ODA end-of-course administrations will lead to sufficient data at the lower levels. For this reason, the researcher recommends DLPT5 and the ODA correlation studies with beginning second language learners using the low-range DLPT5 and the ODA.

Concluding Remarks and Reflections

In the 8th century, Charlemagne is attributed to saying that to speak another language is to possess a second soul. The study of linguistics and language communication has been my passion since I graduated with a degree in communications sciences from a large university in Mexico City. A few years later, I immigrated to the United States, and I felt as if a part of my being—a second soul—developed when I learned English, which I fine-tuned as I developed assessment items for CTB/McGraw-Hill, now part of Data Recognition Corporation (DRC). Assessment development, just as the mastery of English had done, became part of my passion, my life, and my nature.

While working for CTB for over 16 years, I learned subtleties in the development of multiple-choice items versus constructed responses and extended response items. I was assigned to work on the Spanish counterpart of the Tests of Adult Basic Education, TABE Español, and of TerraNova, TerraNova SUPERA. I was as proud of these and other assessments developed for CTB as if they were my own children and, just as for a newborn, I helped name one of these tests: the Spanish TerraNova SUPERA. *Supera* is a Spanish word in command form for “to achieve” or “to overcome obstacles.” (Probably very few would remember that the name SUPERA was also created as an acronym: *Su Preferido Examen de Referencia Académica*, “Your Preferred Exam of Academic Reference”). I later became the project manager for these and many other assessment products, which contributed to my gratitude to the United States that helped me to achieve the American Dream.

Overcoming obstacles is something akin to assessment development.

CTB/McGraw-Hill was later acquired by DRC. When I started to work as a Spanish language instructor for the DLIFLC Distance Learning Division, I learned about DLIFLC’s commitment to foreign language instruction through its worldwide deployment of instructors. My assignments took me to distant and unusual places where I could work on my dissertation, which was any location that had Wi-Fi, and included a charming oyster restaurant next to the Alabama River, a funky coffeehouse in a converted garage in Atlanta, and the exquisite Joslyn Museum’s Café Durham in Omaha, Nebraska. An unexpected joy for my new job arose the moment I went back to a generational family trade: teaching. I remembered that the ultimate goal in education is the success of generations of students, and teaching military students who come from diverse

backgrounds helped me recognize that, for many of these pupils, this is the only way to achieve the American Dream. I recalled that my colleagues at CTB often said that the ultimate goal of any assessment is to help students succeed. While working for DLIFLC, I discovered the ODA. The ODA is akin to the hidden gems I encountered during my teaching assignments for Distance Learning. Not very well known in the United States, it is the only online formative assessment available that competes in scope, design, and complexity with its European counterpart DIALANG. Because I witnessed the tremendous effort in resources and technology in the private sector to develop adaptive and diagnostic assessment instruments, I recognized that developing the ODA was not a simple matter, and I immediately adopted the ODA, as I have done with assessments in the past, as if it were my own child. Originally developed to address the language maintenance and enhancement needs of military staff who had already graduated from the Monterey Basic Course program (nonresident linguists), the ODA has grown to support the formative diagnostic requirements of DLIFLC resident students as well as nonresident students at the basic, intermediate, and advanced levels.

This research is a labor of love for assessment development and foreign language instruction. I hope that this study can help bring recognition to the worthwhile contribution of DLIFLC to foreign language instruction and assessment through the ODA and be the first step in future correlation and validation procedures to help military students succeed and achieve their dreams, as well as to contribute to the fulfillment of the rigorous goals for linguists at DLIFLC.

REFERENCES

- Ableeva, R. (2010). *Dynamic assessment of listening comprehension in second language learning* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3436042)
- Al-Adawi, S. A., & Al-Balushi, A. K. (2016). Investigating content and face validity of English language placement test designed by colleges of applied sciences. *English Language Teaching, 9*, 107-121. ISSN 1916-4742 E-ISSN 1916-4750
- Alade, A. J., & Buzzetto-More, N. A. (2006). Best practices in e-assessment. *Journal of Information Technology Education, 5*, 251-269.
- Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1-24). Harlow, England: Longman.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.
- Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*, 301-320.
- Alderson, J. C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? *EUROSLA Yearbook, 11*, 30-52.
- Alvarez, M., & Rice, J. (2006). *Web-based tests in second/foreign language self-assessment*. Paper presented at the 29th National Convention of the Association for Educational Communications and Technology, Dallas, TX.

- American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines 2012*. Retrieved from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anderson, R. (1997). *Study of initial entry student attrition from Defense Language Institute Foreign Language Center*. Retrieved from <https://archive.org/details/studyofinitialen00ande>
- Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education, 17*, 199-214.
- Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service, Policy Information Center.
- Antón, M. (2009). Dynamic assessment of advanced second language learners. *Foreign Language Annals, 42*, 576-598.
- Ash, D., & Levitt, K. (2003). Working within the zone of proximal development: Formative assessment as professional development. *Journal of Science Teacher Education, 14*, 1-26.

- Assessment Reform Group. (2007). *Assessment for learning*. Retrieved from <http://dera.ioe.ac.uk/7600/1/1f1ab286369a7ee24df53c863a72da97-1.pdf>
- Atkin, J., Black, P., & Coffey, J. (2001). *The relationship between formative and summative assessment—In the classroom and beyond*. Retrieved from http://books.nap.edu/html/classroom_assessment/index.html
- Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476. doi:10.119/0265532202lt240oa
- Bachman, L. (2013). How is educational measurement supposed to deal with test use? *Measurement: Interdisciplinary Research and Perspectives, 11*, 19-23. doi:10.1080/15366367,2013.784150
- Bachman, L., & Clark, J. L. D. (1987). The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science, 490*, 20-33. doi:10.1177/0002716287490001003
- Bachman L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.
- Bax, C., Branford-White, C., Heugh, S., & Jacoby, J. (2014). Enhancing learning through formative assessment. *Innovations in Education and Teaching International, 51*, 72-83. doi:10.1080/14703297.2013.771970
- Belfield, C., & Crosta, P. M. (2012). *Predicting success in college: The importance of placement tests and high school transcripts* (CCRC Working Paper No. 42). Retrieved from <http://ccrc.tc.columbia.edu/Publication.asp?UID=1030>

- Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Educational Policy Analysis Archives*, 9(5), 1-26.
- Bennett, R. E. (2004). Reinventing assessment: How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, 9(5), 1-26.
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18, 5-25.
- Bergin, S. (2002). *Speaking in tongues*. Retrieved from <http://magazine.byu.edu/?act=view&a=845>
- Berman, S. J., Whitt, S., Krol, M., & Salyer, S. (2008). *Comparison of Versant for Spanish™ test scores with DLIFLC Spanish program student performance data* (Research and Analysis Division Report FY08–12). Monterey, CA: Defense Language Institute Foreign Language Center.
- Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal of Research in Mathematics Education*, 24, 442-459.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7-71.
- Black, P., & Wiliam, D. (2003). In praise of educational research: Formative assessment. *British Educational Research Journal*, 29, 623-637.
- Black, P. J., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability*, 21, 5-31.
- Blake, R. (2009). The use of technology for second language distance learning. *Modern Language Journal*, 93, 822-835.

- Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign language skills. *Language Learning, 39*, 313-340.
- Bley-Vroman, R. (1988). The fundamental character of foreign language learning. In W. Rutherford & M. Shanwood Smith (Eds.), *Grammar and second language teaching: A book of readings* (pp. 19-30). New York, NY: Newbury House.
- Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.
- Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21-33.
- Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation, 8*(9), 1-4.
- Bower, M. (2005). Online assessment feedback: Competitive, individualistic, or... preferred form! *Journal of Computers in Mathematics and Science Teaching, 24*, 121-147.
- Brown, H. D. (2004). *Language assessment principles and classroom practice*. White Plains, NY: Longman.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*, 653-675.
- Brown, N. A. (2009). Argumentation and debate in foreign language instruction: A case for the traditional classroom facilitating advanced-level language uptake. *Modern Language Journal, 93*, 534-549.

- Brunfaut, T. (2008). *Foreign language reading for academic purposes: Students of English (native speakers of Dutch) reading English academic texts* (Unpublished doctoral dissertation). University of Antwerp, Antwerp.
- Buck, G. (2011). *Assessing listening*. Cambridge Univ. Press.
- Burwell, G., González-Lloret, M., & Nielson, K. (2009). *Assessment in a TBLT Spanish immersion course*. Paper presented at 3rd Biennial International Conference on Task Based Language Teaching, Lancaster, UK.
- Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27, 5-31.
- Carpenter, T., Fennema, E., & Franke, M. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *Elementary School Journal*, 97, 3-20.
- Center for Advanced Study of Language. (2017). *Improving DLAB's prediction*. Retrieved from <https://www.casl.umd.edu/publications/improving-dlabs-prediction-3/>
- Chalhoub-Deville, M. (2003). L2 interaction: Current perspectives and future trends. *Language Testing*, 20, 369-383.
- Chapelle, C. A., & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27, 301-315.
- Chapelle, C. A., Chung, Y., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27, 443-469.

- Charman, D., & Elmes, A. (1998). A computer-based formative assessment strategy for a basic statistics module in geography. *Journal of Geography in Higher Education*, 22, 381-385.
- Chen, J., Belkada, S., & Okamoto, T. (2004). How a Web-based course facilitates acquisition of English for academic purposes. *Language Learning & Technology*, 8(2), 33-49.
- Christensen, R. B. (2013). *Paying for language skills: The department of defense foreign language incentive program*. Available from ProQuest Dissertations & Theses learning Global. (UMI No. 3563610)
- Clark, M. (2013). Assessment across languages. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 147-155). Malden, MA: Wiley-Blackwell.
- Clark, M., Green, C., Miller, C., Vatz, K., Tare, M., Bonilla, C., . . . Jones, E. (2014). *Assessment challenges in online instruction: Appropriate assessment in online*. Retrieved from http://www.casl.umd.edu/sites/default/files/DO43_7.1_Assessment_Challenges.pdf
- Cohen, L., Manion, L., & Morrison, K. (2003). *Research methods in education*. London, England: RoutledgeFalmer.
- Common Core State Standards Initiative Development Process. (2016). Retrieved from <http://www.corestandards.org/about-the-standards/development-process/>
- Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report No. RR-63). New York, NY: Center on Continuous Instructional Improvement, Teachers College.

- Council of Europe. (2001). *Common Europe framework of reference for languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Pearson.
- Cromeey, A., & Hanson, M. (2000). *An exploratory analysis of school-based student assessment systems* (Report No. TM-032-506). Oak Brook, IL: North Central Regional Educational Laboratory.
- Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.
- Crooks, T. (2011). Assessment for learning in the accountability era. *Studies in Educational Evaluation*, *37*, 71-77.
- Croteau, J. L. (2014). *Online formative assessments as predictors of student academic success* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3630568)
- Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Princeton, NJ: Educational Testing Services.
- Data Recognition Corporation. (2011-2012). *Technical report for the 2011-2012 classroom diagnostic tools*. Retrieved from <http://www.datarecognitioncorp.com/Pages/default.aspx>
- Data Recognition Corporation. (2013). *Assessment update*. Retrieved from <https://pa.drceirect.com>

Defense Intelligence Agency. (2015). Foreign languages. Retrieved from <http://www.dia.mil/Careers/ForeignLanguages.aspx>

Defense Language Institute Foreign Language Center. (2009, January 1). *Online Diagnostic Assessment*. Retrieved from <http://oda.lingnet.org/>

Defense Language Institute Foreign Language Center. (2011). *Online Diagnostic Assessment*. Retrieved from <https://vimeo.com/16633421>

Defense Language Institute Foreign Language Center. (2014). *Online Diagnostic Assessment CONOPS*. Monterey, CA: Author.

Defense Language Institute Foreign Language Center. (2015a). *About*. Retrieved from <http://www.dliflc.edu/about/>

Defense Language Institute Foreign Language Center. (2015b). *Defense Language Proficiency Test 5 system familiarization guide for multiple-choice format*. (2015). Retrieved from <http://www.dliflc.edu/wp-content/uploads/2015/03/Generic-Fam-Guide-MC-CBu-updated.pdf>

Defense Language Institute Foreign Language Center. (2015c). *General catalog*. Retrieved from <http://www.dliflc.edu>

Defense Language Institute Foreign Language Center. (2015d). *Online Diagnostic Assessment team program review*. Monterey, CA: Author.

Defense Language Institute Foreign Language Center. (2015e, February 1). *Two Plus requirements*. Retrieved from <https://vimeo.com/119477856>

Defense Language Institute Foreign Language Center. (2017). *General Catalog*. Retrieved from http://www.dliflc.edu/wp-content/uploads/2016/12/GeneralCatalog_Online-Color_FY17-18.pdf

- DeKeyser, R. M., & Sokalski, K. J. (2001). The differential role of comprehension and production practice. *Language Learning*, 51, 81-112. doi:10.1111/j.1467-1770.2001.tb00015
- Deming, W. E. (1980). *Scientific methods in administration and management* (Course No. 617). Washington, DC: George Washington University.
- Devlin, K. (2015). *Learning a foreign language a 'must' in Europe, not so in America*. Retrieved from <http://www.pewresearch.org/fact-tank/2015/07/13/learning-a-foreign-language-a-must-in-europe-not-so-in-america/>
- Durp, R. P. (2011). Ensuring valid educational assessments for ELL students: Scores, score interpretation, and assessment uses. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 115-142). New York, NY: Routledge.
- Ehrman, M., & Leaver, B. L. (2003). Cognitive styles in the service of language learning. *System* 31, 393-415. doi:10.1016/S0346-251X(03)00050-2
- Ehrman, M., Leaver, M., & Skekhtman, B. (2002). *E&L Learning Style Questionnaire V.2.0*. Retrieved from http://www.uniurb.it/docenti/florasisti/comunicazione_interculturale/2009-10/Quest_StiliCogni.pdf
- Elliott, J. G. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review*, 55, 15-32.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Educational Testing Service. (2011). *ETS item bank: Measuring student success*. Retrieved from <http://www.schoolcity.com/docs/ets-itembank-flyer-2011.pdf>

- Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes*, 8, 26-42.
- Frohbeiter, G., Greenwald, E., Stecher, B., & Schwartz, H. (2011). *Knowing and doing: What teachers learn from formative assessment and how they use information* (CRESST Report 802). Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Gardner, J., Harlen, W., Hayward, L., Stobart, G., & Montgomery, M. (2010). *Developing teacher assessment*. Berkshire, England: Open University Press.
- Gibbons, S. (2010). Collaborating like never before: Reading and writing through a wiki. *English Journal*, 99(5), 35-39.
- Gierl, M. J. (1997). Comparing the cognitive representations of test developers and examinees on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research*, 9, 26-32.
- Glisan, E., & Phillips, J. (1996). *Making the standards happen: A new vision for foreign language teacher preparation*. Yonkers, NY: ACTFL.
- The Glossary of Education Reform. (n.d.). Retrieved from <http://edglossary.org/>
- Goodman, K. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist*, 6, 126-135.
- Goodman, K. (1996). *On reading*. Portsmouth, NH: Heinemann.
- Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education*, 3, 1-49.

- Grigorenko, E. (2009). Dynamic assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities, 42*, 111-132.
- Gulikers, J., Biemans, H. J. A., Wesselink, R., & van der Wel, M. (2013). Aligning formative and summative assessments: A collaborative action research challenging teacher conceptions. *Elsevier Studies in Educational Evaluation, 39*, 116-124.
- Guskey, T. R. (2005). Formative classroom assessment and Benjamin Bloom: Theory, research, and implications. Unpublished manuscript.
- Guskey, T. R. (2010). Formative assessment: The contributions of Benjamin S. Bloom. In H. Andrade & C. Cizek (Eds.), *Handbook of formative assessment* (pp. 106-124). New York, NY: Routledge.
- Haahr, J. H., & Hansen, M. E. (2006). *Adult skills assessment in Europe: Feasibility study*. Denmark: Danish Technological Institute.
- Haertel, E. H. (2006) Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: Praeger.
- Hambelton, R., & Slater, R. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment, 13*, 21-28.
- Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*, 317-336. doi:10.1177/0265532214564505

- Harlen, W. (2005). Teachers' summative practices and assessment for learning: Tensions and synergies. *Curriculum Journal, 16*, 207-223.
doi:10.1080/09585170500136093
- Harris, K., Bauer, M., & Redman, M. (2008). *Cognitive based developmental models used as a link between formative and summative assessment*. Princeton, NJ: Educational Testing Service.
- Henly, D. (2003). Use of web-based formative assessment to support student learning in a metabolism/nutrition unit. *Dental Education, 7*, 116-122.
- Henly, D., & Reid, A. E. (2001). Use of the web to provide learning support for a large metabolism and nutrition class. *Biochemistry and Molecular Biology Education, 29*, 229-233.
- Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.
- Herman, J. L. (2010). *Coherence: Key to next generation assessment success* (AACC report). Los Angeles, CA: University of California.
- Herzog, M. (2015). *An overview of the history of the ILR language proficiency skill level descriptions and scale*. Retrieved from <http://www.govtilr.org/skills/irl%20scale%20history.htm>
- Hintze, J., & Silbergitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372-386. Retrieved from <http://www.nasponline.org>
- Hogan, T. E. (2013). *Using a computer-adaptive test simulation to investigate test coordinators' perceptions of a high-stakes computer-based testing program*

- (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global database. (UMI No. 3583649)
- Horn, R. V. (2003). Computer adaptive tests and computer-based tests. *Phi Delta Kappan*, 567, 630-631.
- Hsueh, S. (2008). *An investigation of the technological, pedagogical and content knowledge framework in successful Chinese language classrooms* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3342724)
- Hubbard, P., & Levy, M. (Eds.). (2006). The scope of CALL education. In P. Hubbard & M. Levy (Eds.), *Teacher education on CALL* (pp. 3-20). Amsterdam, The Netherlands: John Benjamins.
- Hwang, G.-J., & Chang, H.-F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, 56, 1023-1031.
- Interagency Language Roundtable. (2015). *Description of proficiency levels*. Retrieved from <http://www.govtilr.org/skills/ILRscale1.htm>
- Internet@Schools. (2005). ETS develops online item bank of K-12 test questions aligned to state standards. Retrieved from <http://www.internetatschools.com/Articles/News/Breaking-News/ETS-Develops-Online-Item-Bank-of-K-12-Test-Questions-Aligned-to-State-Standsards-58380.aspx>
- Izquierdo, J., & Collins, L. (2008). The facilitative role of L1 influence in tense aspect marking: A comparison of Hispanophone and Anglophone learners of French. *Modern Language Journal*, 92, 350-368. doi:10.1111/j.1540-4781.2008.00751

- Jamieson, J., Grgurovic, M., & Becker, T. (2008). Using diagnostic information to adapt traditional textbook-based instruction. In C. A. Chapelle, Y. R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 25-39). Ames: Iowa State University.
- Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (Unpublished doctoral dissertation). University of Illinois at Urbana, Champaign.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, 26, 31-73. doi:10.1177/0265532208097336
- Jorgensen, M. (2003). *Can the testing industry meet growing demand?* Retrieved from <http://www.issues.Org/19.2/jorgensen.htm>
- Justham, D., & Timmons, S. (2005). An evaluation of using a web-based statistics test to teach statistics to postregistration nursing students. *Nurse Education Today*, 25, 156-163.
- Kane, M. (2011). Book Review: Language assessment in practice: Developing language assessments and justifying their use in the real world. *Language Testing*, 28, 581-587. doi:10.1177/0265532211400870
- Keesling J. W. (2007). *Validity and reliability of DLPT 5 multiple-choice tests*. Retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf

- Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-207). Mahwah, NJ: Erlbaum.
- Kim, A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 32, 227-258.
- Kingsbury, G. G., & Hauser, C. (2004, April). *Computer adaptive testing and No Child Left Behind*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Klinger, D., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy*, 76, 1-34. Retrieved from <http://www.umanitoba.ca/publications/cjeap>
- Knight, P. (2000). The value of a program-wide approach to assessment. *Assessment and Evaluation in Higher Education*, 24, 237-251.
- Koschmann, T. (1999, December). Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. In C. Hoadley & J. Roschelle (Eds.), *Proceedings of the Computer Support for Collaborative Learning (CSCL) 1999 conference* (pp. 308-313). Mahwah, NJ: Erlbaum.
- Kozulin, A., & Garb, E. (2001, August). *Dynamic assessment of EFL text comprehension*. Paper presented at the 9th Conference of the European Association for Research on Learning and Instruction. Fribourg, Switzerland.

- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford, England: Pergamon.
- Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London, England: Longman.
- Krashen, S. D. (1994). The input hypothesis and its rivals. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 45-77). London, England: Academic Press.
- Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*, 607-610.
- Kuh, G., Janowski, N., Ikenberry, S., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities*. Champaign, IL: National Institute for Learning Outcomes Assessments.
- Lam, R. (2013). Formative use of summative tests: Using test preparation to promote performance and self-regulation. *Asia-Pacific Education Researcher, 22*, 69-78. doi:10.1007/s40299-012-0026-0
- Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics, 1*(2), 49-72.
- Lantolf, J. P., & Thorne, S. (2006). *Sociocultural theory and the genesis of second language development*. Oxford, UK: Oxford University Press.
- Lantolf, J. P., & Thorne, S. L. (2007). Sociocultural theory and second language learning. In B. Van Patten & J. Williams (Eds.) *Theories in second language acquisition: An introduction* (pp. 201-224). Mahwah, NJ: Erlbaum.

- Lee, Y., & Sawaki, Y. (2009). Cognitive diagnosis and Q-matrices in language assessment. *Language Assessment Quarterly*, 6, 169-171.
- Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.
- Lidz, C. (1987). *Dynamic assessment*. New York, NY: Guilford Press.
- Lin, J-W., Lai, Y-C., Szu, Y-C., Lai, C-N., Chuang, Y-S., & Chen, Y-H. (2014). Development and evaluation of across-unit diagnostic feedback mechanism for online learning. *Journal of Educational Technology & Society*, 17(3), 138-153.
- Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come*. Retrieved from <http://www.rasch.org/memo69.pdf>
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6). doi:10.1097/00006199-198611000-00017
- Machado de Almeida Mattos, A. (2000). A Vygotskian approach to evaluation in foreign language learning contexts. *ELT Journal*, 54, 335-345.
- McCartney, E., & Perchaud, S. (2014). *ODA/DLPT data analysis for DLIFLC French basic course program* (ODA Action Research Project). Unpublished manuscript.
- McClanahan, L. (2014). Training using technology in the adult ESL classroom. *Journal of Adult Education*, 43, 22-27.
- McManus, K. (2015). L1-L2 differences in the acquisition of form-meaning pairings in a second language. *Canadian Modern Language Review*, 71(2), 51-77.
- McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers.

- McMillan, J. H. (2010). The practical implications of educational aims and contexts for formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 41-58). New York, NY: Routledge.
- McMillan, J., & Schumacher, S. (2009). *Research in education: Evidence-based inquiry* (7th ed.). Boston, MA: Pearson/Allyn and Bacon.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, 23, 13-23.
- Miller, S. M. (2007). English teacher learning for new times: Digital video composing as multimodal literacy practice. *English Education*, 40, 61-83.
- Miller, S. T. (2009). *Formative computer-based assessments: The potentials and pitfalls of two formative computer-based assessments used in professional learning programs* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 305048958)
- Minick, N. (1987). Implications of Vygotsky's theories of dynamic assessment. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 116-140). New York, NY: Guilford Press.
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Washington, DC: Educational Testing Service.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *On the structure of educational assessments* (CSE Technical Report 597). Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.
- Morris, R. D., Stuebing, K. K., Fletcher, J. M., Shaywitz, S. E., Lyon, G. R., Shankweiler, D. P., . . . Shaywitz, B. A. (1998). Subtypes of reading disability:

- Variability around a phonological core. *Journal of Educational Psychology*, 90, 347-373.
- Mozgalina, A., & Ryshina-Pankova, M. (2015). Meeting the challenges of curriculum construction and change: Revision and validity evaluation of a placement test. *Modern Language Journal*, 99, 346-370.
- Myers, S. (2008). *Formative and summative assessments* (Research starters). Great Neck, NY: Great Neck Publishing.
- National Research Council. (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment*. Washington, DC: National Academies Press.
- National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.
- National Research Council. (2014). *Developing assessments for the next generation*. Washington, DC: National Academies Press.
- Olson, L. (2005). ETS to enter formative-assessment market at K-12 level. *Education Week*, 24(25), 11.
- Onwuegbuzie, A. J., & Collins, K. M. (2007). A typology of mixed methods sampling designs in social and science research. *Qualitative Report*, 12, 281-316. Retrieved from <http://tqr.nova.edu/>
- Organisation for Economic Co-operation and Development. (2005). *Formative assessment: Improving learning in secondary classrooms* [Policy brief]. Printed by OECD

- Oxford, R. L. (2017). *Teaching and researching language learning strategies: Self-regulation in context* (2nd ed.). New York, NY: Routledge Taylor & Francis Group.
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Erlbaum.
- Panetta, L. (2011). Foreign language education: If “scandalous” in the 20th century, what will it be in the 21st century? Retrieved from <https://www.stanford.edu/dept/lc/language/about/conferencepapers/panettapaper.pdf>
- Pellegrino, J. W. (1999, November 17). The evolution of educational assessment: Considering the past and imagining the future. Retrieved from <https://www.ets.org/Media/Research/pdf/PICANG6.pdf>
- Pellegrino, J. W. (2004). *The evolution of educational assessment: Considering the past and imagining the future*. Princeton, NJ: Educational Testing Service, Policy Evaluation and Research Center, Policy Information Center.
- Pellegrino, J. W. (2006, November). *Rethinking and redesigning curriculum, instruction, and assessment: What contemporary research and theory suggests*. Washington, DC: National Center on Education and the Economy.
- Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, 20, 65-77.
doi:10.1016/j.pse.2014.11.002
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the “two disciplines” problem: Linking theories of cognition and learning with assessment and

- instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307-353). Washington, DC: American Educational Research Association.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Retrieved from https://www.nap.edu/login.php?record_id=10019&page=http://www.nap.edu/download.php?record_id=10019#
- Pellegrino, J. W., & Hickey, D. (2006). Educational assessment: Towards better alignment between theory and practice. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present and future trends. Sixteen essays in honour of Erik De Corte* (pp. 169-189). Oxford, England: Elsevier.
- Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2013). *Development assessments for the next generation of science standards*. Washington, DC: National Academies Press.
- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5-13.
- Petersen, C. R., & Cartier, F. A. (1975). *Some theoretical problems and practical solutions in proficiency test validity in book by Jones, R.L, & Spolsky, B. (1975) Testing language proficiency*. Arlington, VA: Center for Applied Linguistics.

- Peterson, C. R., & Al-Haik, A.R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement*, 36, 369-380. doi:10.1177/001316447600216
- Phelan, C., & Wren, J. (n.d.). *Assessment issues: Exploring reliability in academic assessment*. Retrieved from <https://www.uni.edu/assessment/issues.shtml>
- Pinckey, R. D., Mealy, M. J., Thomas, C. B., & MacWilliams, P. S. (2001). Impact of a computer-based auto-tutorial program on parasitology test scores of four consecutive classes of veterinary medical students. *Journal of Veterinary Medical Education*, 28, 136-139.
- Pitt, S. J., & Gunn, A. (2004). The value of computer-based formative assessment in undergraduate biological science teaching. *Bioscience Education E-Journal*, 3, Article 1. Retrieved from <http://bio.ltsn.ac.uk/journal/vol3/beej-3-1.htm>
- Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice*, 33(4), 4-12.
- Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French* (Unpublished dissertation). Pennsylvania State University, University Park.
- Poehner, M., & Lantolf, J. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, 9, 233-265.
- Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research*, 17, 323-342.

- Popham, J. W. (2009). Diagnosing the diagnostic test. *Educational Leadership*, 66(6), 90-91. Retrieved from <http://www.ascd.org/publications/educational-leadership/mar09/vol66/num06/Diagnosing-the-Diagnostic-Test.aspx>
- Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5). Retrieved from <http://www.marcprensky.com/writing/prensky%20-%20digital%20natives,%20digital%20immigrants%20-%20part1.pdf>
- Radford, B. W. (2014). *The effect of formative assessments on language performance* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3622118).
- Ragini, S. (2016). Measuring L2 Learning Preferences through Ehrman and Leaver Learning Styles Questionnaire. *Language in India*, 16(4), 58-63.
- Rea-Dickinson, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing*, 17, 215-243.
- Reber, A. S. (1985). *Dictionary of psychology*. New York, NY: Penguin Books.
- Ronan, A. (2015). Every teacher's guide to assessment. Retrieved from <http://www.edudemic.com/summative-and-formative-assessments/>
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York, NY: McGraw-Hill.
- Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39, 369-393.
- Sadler, D. R. (1989a). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.

- Sadler, D. R. (1989b). Formative assessment: Revisiting the territory. *Assessment in Education, 5*, 77-84.
- Salaberry, M. R. (2008). *Marking past tense in second language acquisition: A theoretical model*. London, England: Continuum.
- Sambell, K., Sambell, A., & Sexton, G. (1999). Student perceptions of the learning benefits of computer-assisted assessment: A case study in electronic engineering. In S. Brown, P. Race, & J. Bull (Eds.), *SEDA staff educational and development series: Computer assisted assessment* (pp. 179-191). London, UK: Kogan Page.
- Sato, M., & Atkin, J. M. (2006). Supporting change in classroom assessment. *Educational Leadership, (4)*, 76-79.
- Schneider, E., & Ganschow, L. (2000). Dynamic assessment and instructional strategies for learners who struggle to learn a foreign language. *Dyslexia, 6*, 72-82.
- Schultz, R. K. (2012). *Investigating the effectiveness of classroom diagnostic tools* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3535428)
- Schum, D. (1987). *Evidence and inference for the intelligence analyst*. Lanham, MD: University Press of America.
- Scott-Clayton, J. (2012). *Do high-stakes placement exams predict college success* (CCRC Working Paper No. 41) Retrieved from <http://ccrc.tc.columbia.edu/media/k2/attachments/high-stakes-product-success.pdf>
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (AERA Monograph Series on Curriculum Evaluation, No. 1) (pp. 39-83). Chicago, IL: Rand McNally.

- Shavelson, R. J., Black, P. J., Wiliam, D., & Coffey, J. (2007). *On linking formative and summative functions in the design of large-scale assessment systems*. Stanford, CA: Stanford Graduate School of Education.
- Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 13-26). Rotterdam, The Netherlands: Sense Publishers.
- Shaw M., Jackson L., & Lett A. (1993). The effects of length of service and prior language study at DLIFLC on DLPT attainment. Retrieved from http://calhoun.nps.edu/bitstream/handle/10945/1296/04Dec_Wong.pdf?sequence=1
- Shin, I. (1999). *Analysis of a learning organization: The Korean Language School in the Defense Language Institute*. Available from ProQuest Dissertations & Theses Global. (UMI No. 304557542)
- Shiotsu, T., & Weir, C. J. 2007. The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24, 99-128.
- Shirai, Y., & Kurono, A. (1998). The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning*, 48, 245-279. doi:10.1111/1467-9922.00041
- Sieber, V. (2009). Diagnostic online assessment of basic IT skills in 1st-year undergraduates in the Medical Sciences Division, University of Oxford. *British Journal of Educational Technology*, 40, 215-226.

- Silye, M. F., & Wiwczaroski, T. B. (2002). *A critical review of selected computer assisted language testing instruments*. Retrieved from <http://www.date.hu/acta-agraria/2002-01i/fekete1.pdf>
- Singer, P. (2014). *Federally supported innovations: 22 examples of major technology advances that stem from federal research support*. Retrieved from http://www2.itif.org/2014-federally-supported-innovations.pdf?_ga=1.5314359.565401261.1472449909
- Skehan, P. (2014). *Processing perspectives on task performance*. Amsterdam, The Netherlands: John Benjamins Publishing.
- Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185-211.
- Sly, L. (1999). Practice tests as formative assessment improve student performance on computer-managed learning assessments. *Assessment and Evaluation in Higher Education, 24*, 339-343.
- Son, J.-B. (2008). Using Web-based language learning activities. *International Journal of Pedagogies and Learning, 4*(4), 34-43.
- Sparks, R. L., & Ganschow, L. 1993. The impact of native language learning problems on foreign language learning: Case study illustrations of the linguistic coding deficit hypothesis. *Modern Language Journal, 77*, 58-74.
- Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2006). Native language predictors of foreign language proficiency and foreign language aptitude. *Annals of Dyslexia, 56*, 129-160.

- Standards in Your State. (2016). Retrieved from <http://www.corestandards.org/standards-in-your-state/>
- Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86*, 24–53.
- Sternberg, R. J., & Grigorenko, E. L. (2001). All testing is dynamic testing. *Issues in Education, 7*, 137-170.
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing*. New York, NY: Cambridge University Press.
- Sternberg, R., Grigorenko, E., & Zhang, L. (2008). Styles of learning and thinking matter in instruction and assessment. *Perspectives on Psychological Science, 3*, 486-506. doi:10.1111/j.1745-6924.2008.00095
- Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*, 699-715.
- Stiggins, R. J. (1997). *Student-centered classroom assessment*. Upper Saddle River, NJ: Prentice-Hall.
- Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2009). *Classroom assessment for student learning: Doing it right—Using it well*. Portland, OR: Assessment Training Institute.
- St. Pierre, C. N., III. (2008). *Foreign language learning and the efficacy of preparatory course interventions* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3297027)

- Sugaya, N., & Shirai, Y. (2007). The acquisition of progressive and resultative meanings of the imperfective aspect marker by L2 learners of Japanese: Transfer, universals or multiple factors? *Studies in Second Language Acquisition*, 29, 1-38.
doi:10.1017/S0272263107070015
- Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction: Toward a theory of teaching. *Educational Researcher*, 41(5), 147-156.
- Taghizadeh, M., Alavi, S., & Rezaee, A. (2014). Diagnosing L2 learners' language skills based on the use of a web-based assessment tool called DIALANG. *Journal of Distance Education*, 29(2), 1-28.
- Takala, S. (1998). *Language testing: Recent developments and persistent dilemmas*. Retrieved from ERIC database. (ED460636)
- Tanimoto, S. L. (1987). *The elements of artificial intelligence*. Rockville, MD: Computer Science Press.
- Taras, M. (2005). Assessment—summative and formative—some theoretical reflections. *British Journal of Educational Studies*, 53, 466-478.
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36. doi:10.1017/S0267190509090035
- Teschner, R. (Ed.). (1991). *Assessing foreign language proficiency of undergraduates*. Boston, MA: Heinle & Heinle.
- Test Development Division, Evaluation and Standardization, Defense Language Institute Foreign Language Center. (2007). *Defense Language Proficiency Testing System*

- 5 framework*. (2007). Retrieved from http://www.dliflc.edu/wp-content/uploads/2015/11/20090910_VLR_DLPT_Framework_Doc.pdf
- Tharp, R., & Gallimore, R. (1991). *The instructional conversation: Teaching and learning in social activity* (Research Report 2). Santa Cruz, CA: The National Center for Research on Cultural Diversity and Second Language Learning, University of California, Santa Cruz.
- Tharpe, T. L. (2010). Wiki, wiki, wiki—what? Assessing online collaborative writing. *English Journal*, 99(5), 40-46.
- Thayer, K. K. (2013). *The diffusion of innovations in education: A study of secondary English language arts teachers' classroom technology integration* (Doctoral dissertation, Florida State University). Retrieved from <https://fsu.digital.flvc.org/>
- Torgesen, J. K., Morgan, S. T., & Davis, C. (1992). Effects of two types of phonological awareness training on word learning in kindergarten children. *Journal of Educational Psychology*, 84, 364-370.
- Trumbull, E., & Lash, A. (2013). *Understanding formative assessment insights from learning theory and measurement theory*. Retrieved from <https://www.wested.org/resources/understanding-formative-assessment-insights-from-learning-theory-and-measurement-theory/>
- Tucker, M. S. (2010). *An assessment system for the United States: Why not build on the best?* Retrieved from <http://www.k12center.org/rsc/pdf/TuckerSystemModel.pdf>
- Tucker, M. S., & Coddling, J. (2002). *The principal challenge: Leading and managing change in an era of accountability*. New York, NY: Wiley.

- Turner, J. (1993). Using Likert scales in L2 research: Another researcher comments. *TESOL Quarterly*, 27, 736-739. doi:10.2307/3587409
- University of Illinois at Urbana-Champaign. (2012). *English Placement Test (EPT)*. Retrieved from <http://www.publications.uiuc.edu>
- Urciouli, B. (2005). The language of higher education assessment: Legislative concerns in a global context. *Indiana Journal of Global Legal Studies*, 12, 183-204.
- U.S. Department of the Army. (1994a) *Prediction of language learning success at DLIFLC: LSCP II*. Monterey, CA: Author.
- U.S. Department of the Army. (1994b). *Prediction of language learning success at DLIFLC: LSCP III*. Monterey, CA: Author.
- U.S. Department of the Army. (2015). *U--Online diagnostic assessment testlets* (Solicitation No. W9124N-15-R-0001). Retrieved from https://www.fbo.gov/index?s=opportunity&mode=form&id=b026a0e31d70cfbc484cb2f712db01a3&tab=core&_cview=0
- U.S. Department of the Army. (2016). *Army Regulation 11-6*. (2015). Retrieved from http://www.apd.army.mil/epubs/DR_pubs/DR_a/pdf/web/r11_6.pdf
- U.S. Department of Defense. (2009). DoD language testing program (Instruction No. 5160.71). Retrieved from https://dlseo.org/sites/default/files/DoDI_5160.71.pdf
- U.S. Department of Defense. (2013). *Military foreign language skill proficiency bonuses* (Instruction No. 1340.27). Retrieved from <http://www.dtic.mil/whs/directives/corres/pdf/134027p.pdf>

- Valenti, S., Nitko, A., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319-329.
- Vendlinski, T., & Stevens R. (2002). Assessing student problem-solving skills with complex computer based tasks. *Journal of Technology, Learning and Assessment*, 1(3). Available at <http://scholarship.bc.edu/jtla/vol/3>
- Vygotsky, L. S. (1963). Learning and mental development at school age (J. Simon, Trans.). In B. Simon & J. Simon (Eds.), *Educational psychology in the U.S.S.R.* (pp. 21-34). London, England: Routledge & Kegan Paul.
- Vygotsky, L. S. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Vygotsky, L. (1998). The problem of age. In R. W. Reiber (Ed.), *The collected works of L. S. Vygotsky* (Vol. 5, pp. 187-205). New York, NY: Plenum.
- Walqui, A., & van Lier, L. (2010). *Scaffolding the academic success of adolescent English language learners*. San Francisco, CA: WestEd.
- Wang, M., Peng, J., Cheng, B., Zhou, H., & Liu, J. (2011). Knowledge visualization for self-regulated learning. *Educational Technology & Society*, 14(3), 28-42.
- Warschauer, M., & Meskill, C. (2000). Technology and second language learning. In J. Rosenthal (Ed.), *Handbook of undergraduate second language education* (pp. 303-318). Mahwah, NJ: Erlbaum.
- Watson, D. M. (2001). Pedagogy before technology: Re-thinking the relationship between ICT and teaching. *Education and Information Technologies*, 6(4), 251-266. doi:10.1023/A:1012976702296

- Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155-180). Mahwah, NJ: Erlbaum.
- Wilen-Daugenti, T. (2009). *Edu: Technology and learning environments in higher education*. New York, NY: Peter Lang.
- Wilson, M. R., & Bertenthal, M. W. (2006). Executive summary. In M. R. Wilson & M. W. Bertenthal (Eds.), *Systems for state science assessment* (pp. 1-10). Washington, DC: National Research Council, National Academies Press.
- Wisconsin Center for Education Research. (2009). WIDA focus on formative assessment. Retrieved from <https://www.wida.us/get.aspx?id=215>
- Wong, C. H. (2004), An analysis of factors predicting graduation of students at Defense Language Institute Foreign Language Center. Retrieved from http://calhoun.nps.edu/bitstream/handle/10945/1296/04Dec_Wong.pdf?sequence=3
- Yatzkanic, R. (2015). *The relationship between formative assessments and PSSA performance* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3700695)
- Zhang, T., Mislevy, M. J., Haertel, G., Javitz, H., Murray, E., Gravel, J., & Hansen, E. G. (2010). *A design pattern for a spelling assessment for students with disabilities* (Assessment for Students With Disabilities Technical Report 2). Menlo Park, CA: SRI International.
- Zhou, J. (2010). *Estimating attribute-based reliability in cognitive diagnostic assessment* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 305234931)

Zou, K., O'Malley, A., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*, 654-657.

doi:10.1161/CIRCULATIONAHA.105.594929

APPENDICES

APPENDIX A

Department of the Army Letter of Consent



DEPARTMENT OF THE ARMY
DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER
PRESIDIO OF MONTEREY
MONTEREY, CALIFORNIA 93944-6000

March 15, 2016

Office of the Commandant

Office of Human Subject Research Protection
Chapman University
One University Drive
Orange, CA 92866


To Whom it May Concern:

This letter is to express our willingness to grant for Ms. Aima S. Castro, a doctorate student at the Chapman University, permission to use the requested student assessment data for her dissertation research project tentatively titled, "*Online Formative Assessments as Valid Predictors of ILR Foreign Language Proficiency As Measured By Summative Tests.*" Ms. Castro's proposal was reviewed by our Scientific Review Board and acquired the Board's support. It is our pleasure to support this qualified academic research activity that is relevant to our institute's mission.

The permission is contingent on Chapman University's IRB review and DLIFLC's administrative review. Per our understanding, the Chapman University IRB will conduct the institutional review and will maintain oversight for this research project. Following the IRB approval at the Chapman University, DLIFLC's Human Research Protection Program (HRPP) Office will conduct an administrative review in accordance with DoD requirements for supported research regardless of its exempted status. The administrative review ensures compliance with DoDI 3216.02, "Protection of Human Subjects and Adherence to Ethical Standards in DoD-Supported Research" in addition to "the Common Rule." The requested data cannot be released before DLIFLC completes the administrative review.

If you have any question, please contact Dr. Heejong Yi, Scientific Review Board Chair, (831) 242-7245 or heejong.yi@dliflc.edu, or Ms. Marzenna Krol, HPA, (831) 242-3655 or marzenna.krol@dliflc.edu.

Sincerely,



Phillip J. Deppen
Colonel, U.S. Army
Commandant

APPENDIX B

DLIFLC ODA VALIDATION Process

The following categories are included part of the ODA design and are part of the Diagnostic Profile reports:

1. Content
 - a. Main ideas
 - b. Supporting ideas
 2. Linguistic questions
 - a. Vocabulary
 - i. Foreign Language Objectives (FLO) Topics
 - ii. ODA Subtopics
 - b. Structure
 - i. Language-specific features
 - c. Discourse
 - i. Language-specific features
- Listening includes an additional section:
- d. Speech Processing
 - i. Delivery—authentic vs. modified speech
 - ii. Vocabulary—oral vs. transcribed

(DLIFLC ODA CONOPS 2014, p. 5).

Each ODA grouping labeled “testlet” contains reading or listening stimuli and items specifically designed to measure core content through main idea and supporting idea skills, and linguistic items measuring lexicon, structure, and discourse following specific ILR guidelines for each level.

The test taker receives a set of three “testlets” during a test session. Per completion of a testlet grouping the system evaluates whether a more difficult or a less difficult testlet grouping is administered.

Each stimulus in the “testlet” includes a main question, one or two supporting questions depending on the testlet ILR level, five to seven contextual vocabulary items also depending on the ILR level assigned and one Structure item. Discourse items are not included in Level 1 but are included in testlets for levels 1+ to 3 and are designed according to the corresponding ILR difficulty. The testlet grouping evaluation contributes to the computer adaptive capabilities of the ODA (DLIFLC Online Diagnostic Assessment Program Review, 2015).

Stimuli Selection

Part of the process for validating the ODA requires that the selection of stimuli follow very specific criteria in accordance to the ILR requirements for each performance level. Item development does not start until stimuli have been rated by ODA experts and stimuli have been adjusted to meet ILR level requirements.

Stimuli are selected based on their varied distribution across several Foreign Language Objective topics and their subject appropriateness specific to a given ILR level.

A checklist with stimuli criteria to rate and approve stimuli prior item development includes:

- 1) ILR intended level
- 2) Specified linguistic requirements for the intended level
- 3) Topic and target language requirements for the intended level
- 4) Specific content requirements for the intended level
- 5) Review of stimuli to avoid prior knowledge information
- 6) Review of stimuli to avoid subject matter that may be outdated over time
- 7) Stimuli review for cultural appropriateness and cultural representation across target language's regions
- 8) Stimuli review for appropriateness in genre representation across different type of paper-based and electronic type of publications
- 9) Review of stimuli for suitable language use and length specific for targeted ILR level

(DLIFLC, 2015d).

Testlet Design

Once stimuli are approved, item development starts. For the item development, a set of four to six items is required for each stimulus. Per ODA specifications, there are two types of items: content-based items and linguistic items. Content-based items are designed to measure the understanding of main ideas and supporting ideas of different types of texts, details, ideas, and arguments. Linguistic items are designed to measure the understanding of sentence structure, vocabulary and phrases that could contribute to the reading comprehension, and discourse or connection between ideas. Linguistic items are classified under Linguistic, Lexical, and Discourse. (ODA Reading Diagnostic Profile DLIFLC, 2015; DLIFLC, 2015d).

The testlet design configuration is as follows:

Section 1: Content-based items

Main idea type question

Supporting idea type question

Supporting idea type question

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structural item

Discourse item

Item distribution. The ODA item distribution per testlet has been designed to meet adaptive test requirements and target difficulty. The item distribution per level is as follows:

ODA Test Design for Level 1

Section 1: Content-based items

Main idea type question

Supporting idea type question

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structure item

ODA Test Design for Level 1+

Section 1: Content-based items

Main idea type question

Supporting idea type question

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structural item

Discourse item

ODA Test Design for Levels 2, 2+ and 3

Section 1: Content-based items

Main idea type question

Supporting idea type questions (two items)

Section 2: Linguistic items

Vocabulary (lexical) items (five to seven items)

Structural item

Discourse item

(ODA Diagnostic Profile, 2015; ODA website, 2015; ODA Program Review, 2015).

Item requirements.

The ODA follows very specific guidelines for the development of items once stimuli have been approved. It uses an authoring system known as ODA Generator for the item development which provides the shell for the consistent development and management of items and later selection of testlets in order to meet ODA criteria based on ILR requirements. Items developed include open-ended and multiple-choice items.

The items developed to measure ODA objectives have specialized item formats. For example, main idea and supporting idea type of items are measured through open-ended item formats. Lexical items and structural type of items are measured through multiple-choice and open-ended item formats. While the lexical type items use a

distinctive open-ended design, the structural and the discourse items are developed using a varied of multiple-choice and open-ended format design. Below are some examples of ODA item formats.

Example of item format for Content-based item, Reading.



Example of item format for Lexical item, Reading.



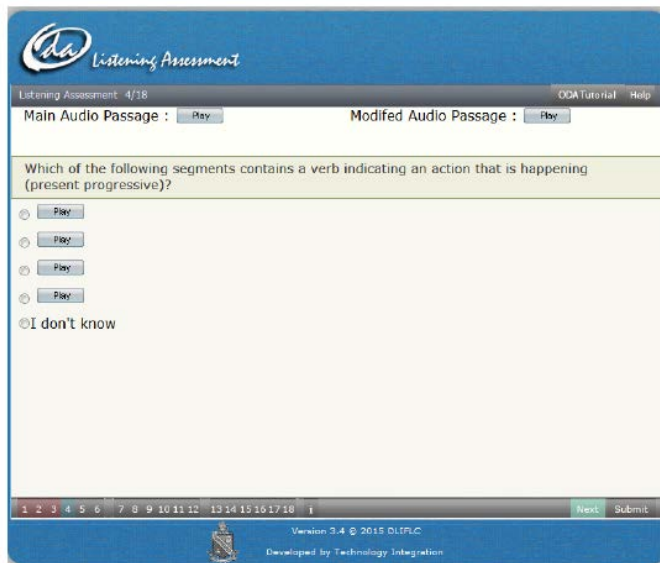
Example of item format for Structure items, Reading.

The screenshot shows a web-based Reading Assessment interface. At the top, there is a logo for 'CDA Reading Assessment' and a progress indicator 'Reading Assessment: 14/15'. The main content area contains a paragraph of Arabic text. A specific word in the text is highlighted. Below the text, there is a question in English: 'Identify what or who (subject) is performing the highlighted word.' Below the question is a text input field labeled 'Selected Word/Phrase:' and a 'Paste' button. At the bottom, there is a navigation bar with numbered buttons (1-15) and 'Next' and 'Submit' buttons. The footer indicates 'Version 3.0 © 2009 DLIFLC'.

Example of item format for Lexical items, Listening.

The screenshot shows a web-based Listening Assessment interface. At the top, there is a logo for 'CDA Listening Assessment' and a progress indicator 'Listening Assessment: 6/18'. Below the logo, there is a button labeled 'Listen to main passage' with a 'Play' icon. The main instruction reads: 'Give the English equivalent of the following words as used in the passage.' Below this instruction is a table with three columns: 'Audio', 'Item Transcribed', and 'English Meaning'. The 'Audio' column contains 'Play' buttons. The 'Item Transcribed' column contains the word 'display' repeated six times. The 'English Meaning' column contains empty text input fields. At the bottom, there is a navigation bar with numbered buttons (1-18) and 'Next' and 'Submit' buttons. The footer indicates 'Version 3.4 © 2013 DLIFLC' and 'Developed by Technology Integration'.

Example of item format for Structure items, Listening.



(DLIFLC ODA Program Review, 2015)

The ODA follows quality control procedures common in the assessment development industry for formative and summative assessment development. The ODA development and review cycle ensures the quality of stimuli, questions, and item development criteria for multiple-choice and open-ended items through a standardized item development cycle that includes strict stimuli review and approval prior to item development, item and testlet review, validation, revision, and monitoring.

As part of the review cycle, subject matter experts require peer reviews as well as senior reviews. Because the test items require English stems to elicit English responses, items also go through an English editing process.

Because of the high level of granularity required for the ODA, each item needs to have what is known as language metadata tags. These are identifiers used through the authoring system to track the item information required for the development of the narrative descriptions related to the skills measured for each individual reading and

listening item. These metadata tags are particularly necessary to identify structure and discourse type items and are also helpful to tag the narrative descriptions required for the individualized ODA diagnostic profiles. Also needed for the ODA diagnostic profiles is a diagnostic profile matrix. The diagnostic profile matrix needs to be updated, particularly for structure and discourse items once they are completed, and needs to be revised and re-edited once the testlets are selected in order for the narrative descriptions of the diagnostic profile matrix to provide clear diagnostic profile statements specific to the specific items developed.

The diagnostic profile statements need to provide meaningful information that is clear and comprehensive for test takers, so that users know exactly what are their areas of strength and weakness and how to make informed decisions about their next step in their learning process. After the profile statements from the diagnostic profile matrix are further revised and updated they become part of the ODA metadata and through the tagging system can be linked to the testlets for use by the ODA system. After all structure and discourse items in the selected testlets are tagged, the ODA assessment system is ready for the next validation cycle.

Testlet iteration.

Once items are approved and placed into testlets, they go through a cycle known as testlet iteration. This process requires a minimum of three testlets per level for levels 1 to 3 in order to fulfill the computer adaptive requirements for upward or downward performance level mobility. After all testlets are developed and accurately reviewed to measure the corresponding levels intended to measure, the adaptive features can also be

tested. Sets of three testlets are needed for upward and downward mobility in order to verify the accurate proficiency level of test takers.

Therefore, an ODA iteration requires a minimum of three testlets for each level and a total of six testlets for levels 1 to 3. This procedure ensures meeting the adaptive requirements of the ODA as well as the quality standards specific to formative assessments such as the ODA.

Per ODA Program Review (2015), below is minimum number of testlets needed to meet ODA computer adaptive requirements:

Level	Number of Testlets
1	6
1+	12
2	9
2+	6
3	6
Total	39

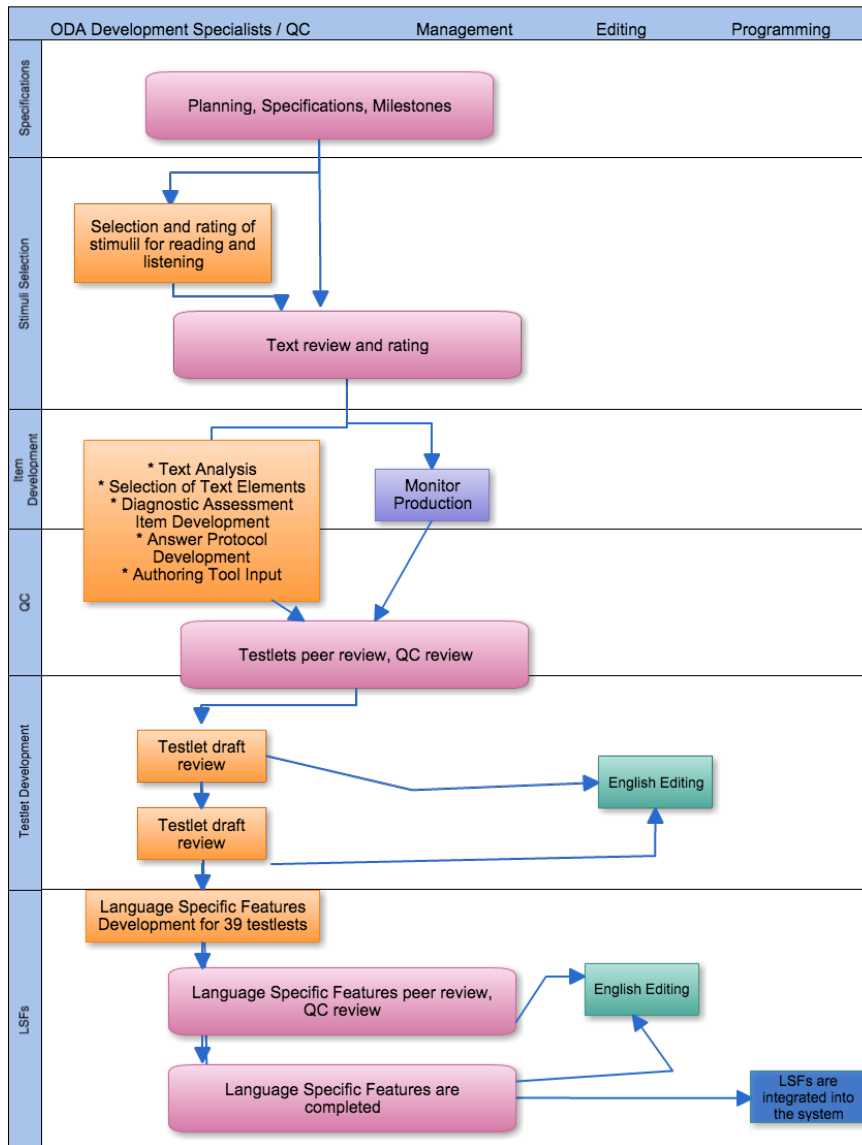
Once a group of testlets has the minimum number of testlet groupings to meet the ODA adaptive requirements, the ODA assessment system is ready to go to the next cycle of development and validation which include: a) the completion of the Diagnostic Profile Master, b) the testing and validation process, and c) monitoring of ODA fully operational items.

The workflow of the ODA is similar to workflows used in the assessment industry to ensure a reliable high quality development and production process. The difference

noticed is that the ODA workflow ensures its review and quality assurance through the use of automated checklists, which further ensure the quality and item development accountability. Subject matter experts, reviewers and managers need to physically click on every item review criteria's checklist and include written feedback in order to validate each item and its anatomic parts.

Below is a workflow showing the development and review of the ODA for the first phase of the development and validation process.

ODA Workflow First Phase



(DLIFLC ODA Program Review, 2015)

After all structure and discourse items in the selected testlets are tagged, and after the profile statements from the diagnostic profile matrix are further revised and updated, the ODA assessment system is ready for an automated review known as testlet checker, TCH.

The TCH is made with HTML scripting codes and Dynamic HTML as well as other scripting languages. The TCH ensures that items and testlets follow the technical specifications, naming conventions, and standards required, and identifies possible errors that may alter the effective flow of the ODA system. Per the ODA Program Review (2015) the TCH for the ODA was designed to: 1) use raw data to create xml files for testlet uploading; 2) check for naming conventions, audio files bit rate values, and possible human errors that may prevent the generation of site script information, grading, testing, or individual diagnostic profile output; and 3) automated verification of testlet distribution per level.

A series of reports are generated at each step in the TCH verification process, which include information about the type of error, and provide identifier information to locate the error in an item, testlet, or file data.

A review cycle that includes updating ODA input per TCH verification is implemented and a second TCH is performed. The xml testlet files produced by the TCH are uploaded into two secure server databases: one database for the reading content area, and another database for the language content area.

The ODA server databases for reading and listening are comprised of two segments: a client segment and a server segment. The xml loading process is made

through a series of scripts and server technology scripting language labeled as a “Testlet Loader.” These scripts aid the database loading process. Two different types of scripts are used to meet the specific requirements of the database: the client segment of the database uses HTML, Jscript, CSS, and JQuery scripts; and the server segment uses Jscript and Active Server Page Technology (DLIFLC ODA Program Review, 2015).

The “Testlet Loader” helps organize the xml testlets and link them to the corresponding tables and auxiliary (AUX) tables and segments of the database.

After testlets have been uploaded through the “Testlet Loader” process, the ODA is ready for the next validation cycles, which include in-house testing, beta testing, field testing, and what is known as “debugging.”

Per this cycle, different ODA stakeholders participate in the field-testing process, which include in-house developers, students, language schools, military bases, and DLIFLC language training detachments. The field-testing process includes checking the performance of the site as well as the item testlets. Per DLIFLC ODA Program Review (2015), the validation cycles after the ODA testlets are completed are as follows:

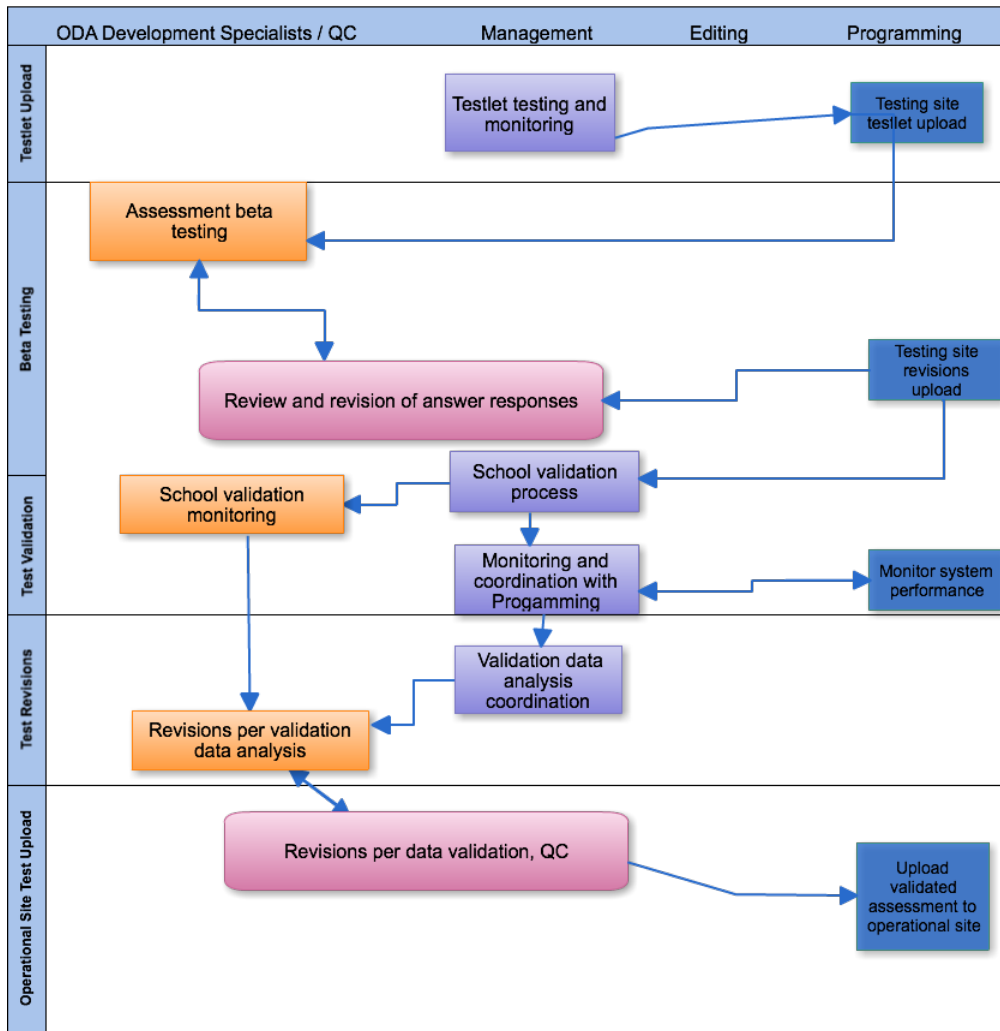
- 2) Item testlets are made available through an Internet testing site for testing.
- 3) Through the test taking process the system is “debugged” to ensure that interface works as expected.
- 4) Developers and reviewers take the test in its pre-operational form through the Internet site.
- 5) Revisions are made based on input from developers and reviewers.
- 6) Testing is performed with native speakers to review appropriateness of test at the higher levels, particularly Level 3.

- 7) Revisions are made.
- 8) Items are validated through the administration of the testlets to groups of students with different language ability levels and at different stages in the school semester to verify testlet levels and item discrimination.
- 9) Revisions are made.
- 10) Items are made operational through the ODA official Internet site.
- 11) Items are monitored to verify that they measure the target level, and are able to produce discriminating output between levels according to ILR criteria.
 - a. Items are verified to ensure that they lend to the targeted student performance outputs.
 - b. Testlets are verified to ensure that they produce the expected floor and ceiling output per testlet ILR level design.
 - c. Level testlets are validated to make sure that, for example, a Level 1+ student performs as expected on a Level 1 testlet but has difficulty at a Level 2 testlet, while a Level 2 student performs as expected on a Level 2 testlet, but has difficulty with a Level 3 testlet.
- 12) Items are also monitored to ensure that they lend to the expected open-ended item responses, the answers have the expected complexity and completeness and all possible correct responses are taken into account.

(DLIFLC ODA Program Review, 2015)

ODA Field Testing and Validation Cycle

ODA Workflow Second Phase



(DLIFLC ODA Program Review, 2015).

Client Segment of the ODA Databases

One essential component of the ODA server databases for reading and listening is the client segment of the ODA server. This client interface segment allows test takers access to the ODA assessment. The client segment uses CSharp Web service technologies, which include Microsoft NET systems. These technologies allow for the ODA to be available through tablets, smart phones, laptops, and desktop computers. The Web service technologies connect to the ODA server databases to support test

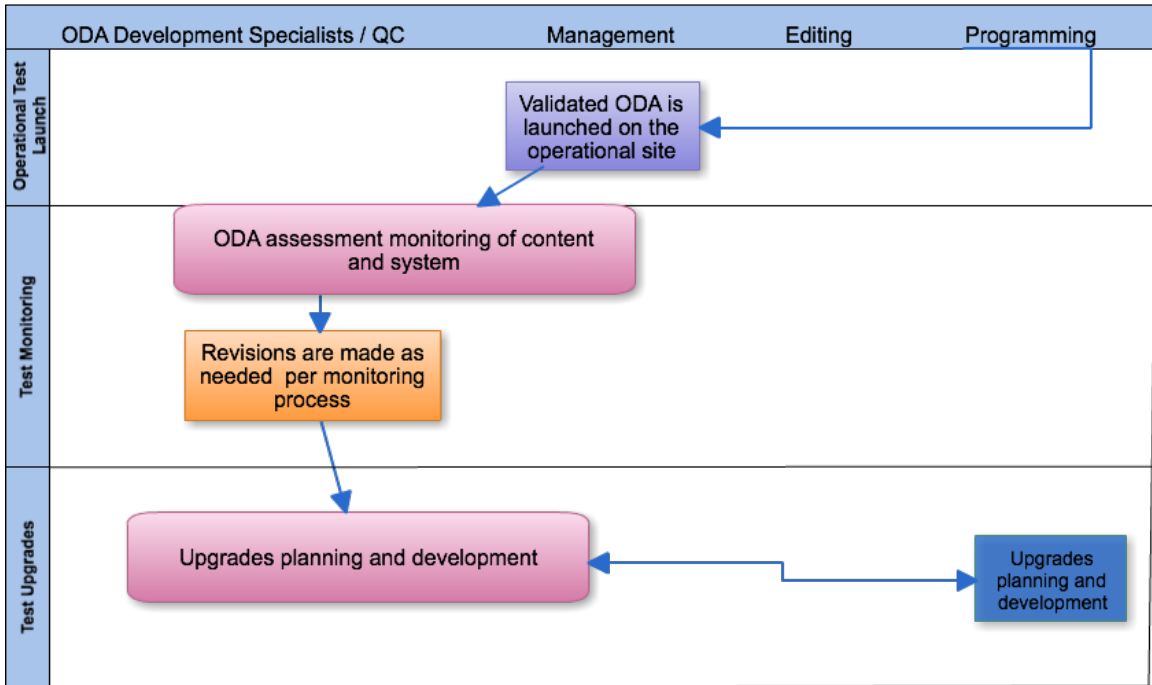
administration and produce diagnostic profiles. The Web service technologies have been updated over the years to ensure non-interrupted test taking and the most efficient asynchronous communication possible through Fail-safe Web applications that have the capability to provide service to over a minute in cases of network disconnect.

Another essential element that makes the ODA effective and increases its validity is the incremental integration of testlets over time as well as its technical capability to monitor the ODA results to make timely updates to the ODA assessment instrument once it is fully functional. This monitoring and updating of the ODA helps developers remove unexpected outliers, unforeseen discrepancies, or unidentified content issues found by users and include a user's survey. The test taker's response data and survey comments go back to the developers and managers for monitoring. The reprinting process, for paper-based assessments, could be very costly. For an online test, it could further strengthen the quality of its diagnostic assessment and diagnostic profile. In this context, overseeing and reviewing the ODA's assessment performance results once the ODA has become operational is an important step in the development and maintenance cycle. Therefore, the next steps in the validation process of the ODA are unique and relevant to well-designed formative diagnostic assessments.

Once the ODA items and testlets become operational, the ODA system is then monitored through a database. This database includes an automated feature labeled as "Item-User Correlation" that helps identify the level of discrimination between items and testlets across levels as well as the validation of all possible correct answers for open-ended items. Through this monitoring process, some items may be replaced or updated because its content may have become outdated, societal and cultural exposure to certain

content may over time elicit prior knowledge responses, items may not provide the expected outcomes, or there may be a need to develop new content on a specific area or skill where gaps might have been identified.

ODA Workflow Third Phase



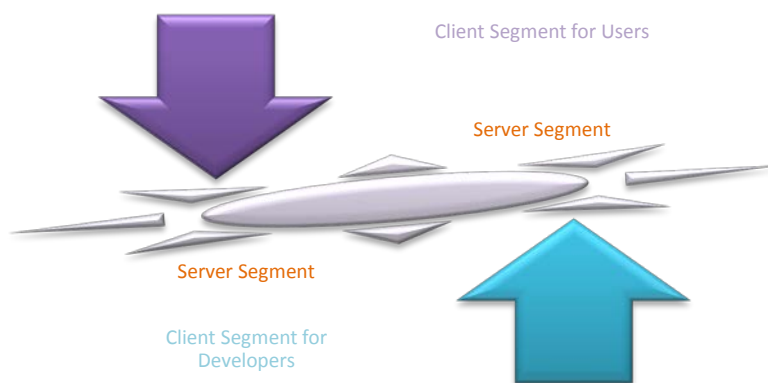
(DLIFLC ODA Program Review, 2015).

The standardized procedures and technical features available for database search, monitoring, verification and review of the ODA database server after items are made operational include:

- Main panel information
- Keyword review
- Tag review
- Answers from users
- Item to item correlation
- Item to user correlation

- Item performance
- Question review by primary tag
- Passage review by proficiency level
- Testlet inventory review
- Error list review
- Review of data by data type
- ODA rate of usage review
- Testlet rate of usage review
- Testlet uploading into the database rate
- Review of ODA User's Survey

The ODA database could be visualized by its segments: a server side which is the backbone of the system; a client side with a log-in access for users to provide input; and a log-in side for developers, to analyze the input. The client side was designed so that developers can monitor item performance, testlet data, and item and testlet correlation among other things. The server side connects the input from the test takers and the developers. It authenticates and stores scoring data, and allows for the delivery of score and item response information to the developers for item analysis verification.



While simplistic, the image below shows a representation of the ODA Database for each content area (one database for reading and one database for listening for each of the foreign languages available). The pink shape in between represents the server segment that connects the input from the users and the test developers and is designed to evaluate diagnostic assessment data and provide assessment data input to the developers for monitoring through the developers client segment of the database.

In this context, an essential component of the ODA server is the client segment for users. Per ODA Program Review (2015) the process flow for the ODA user's segment is as follows:

- 1) User's login for registration for password retrieval provides script information to the server. It allows for the user's segment of the database to verify and add new information.

- 2) Test taking process starts after successful login, test taker's selection of language, and self-selected starting performance level.

- 3) Generation of assessment through an algorithm is issued sending the information to the server client script.

- 4) Client side script interprets test taker's information and assembles online testing session with all possible assessment level testlet operations which include:

- a. Test takers' answering of three-passage testlets.
- b. Answers to passage testlets are received by the server.
- c. Server evaluates answers analyzing key words related to open-ended items and a grade is sent to the client side of the server.
- d. Evaluation and student responses are sent to the client side for developer's side of the server for future monitoring of answers, performing of statistics and other types of analysis including quality control procedures.

5) Upon final testlet assessment administration, two evaluations are determined: one for the current performance level and another for the target level required for the test taker to master the next level of proficiency.

6) The Server Side receives a script with the information and issues a diagnostic profile based on the user's specific assessment.

7) Test taker views an individual Diagnostic Profile upon completion of the ODA instrument.

8) Through the Profile Creation feature, the test taker receives an Individual Diagnostic Profile via e-mail, which can also be sent to other stakeholders.

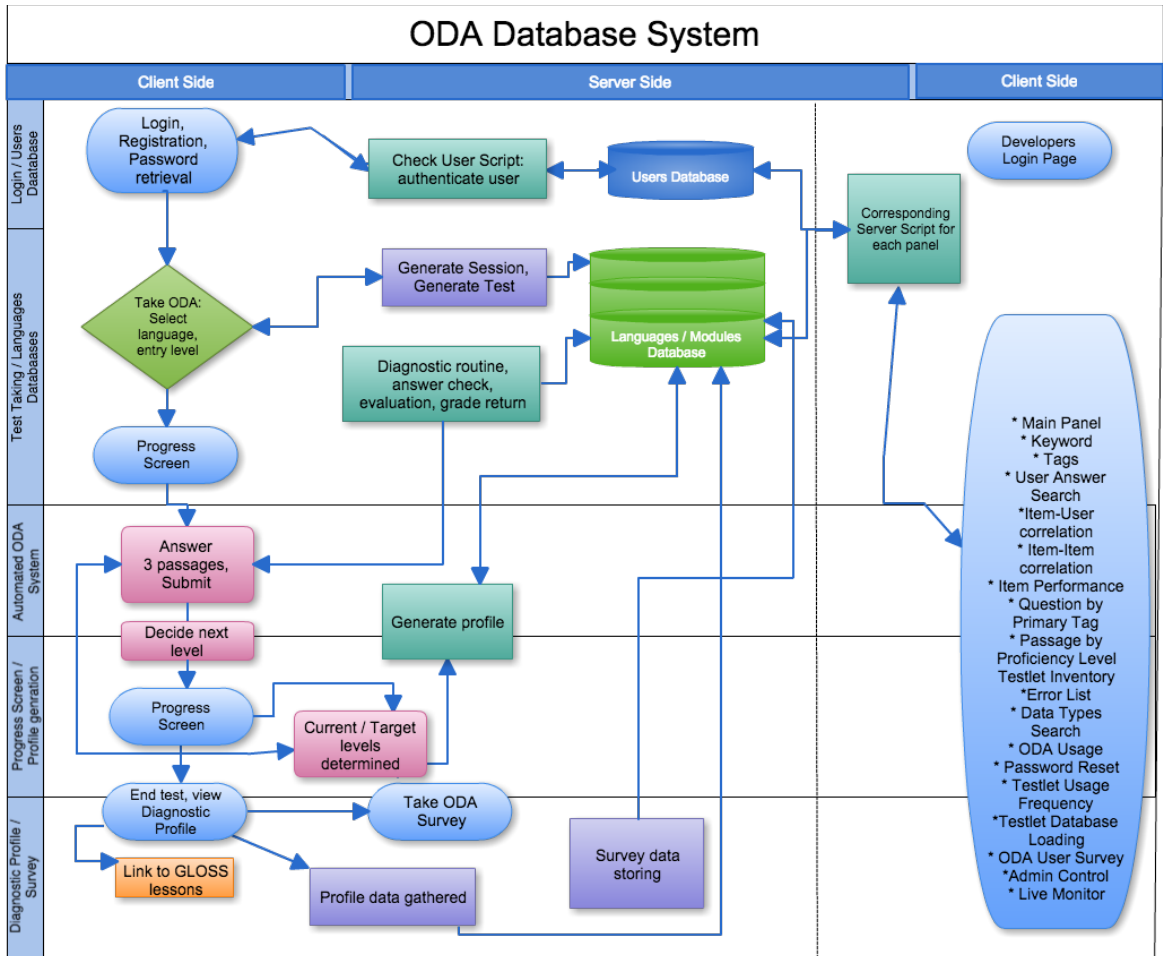
9) A feature to answer the user's survey is issued for test takers to respond to questions regarding the test taking experience, and specific questions regarding usability, assessments, diagnostic profile and provide any comments.

10) A script is issued on the server side of the database to store the assessment responses for future statistical diagnosis and item monitoring.

11) As of 2015, the ODA has a new feature, a Progress Screen. This screen is the result of feedback from users and its purpose is to provide information to test takers of the progress of the assessment, helping the user know how far he is in the completion of the assessment as well as the performance level obtained at specific points of the development progress. Because there is not a pre-established number of items delivered for any given ODA assessment, the Progress Screen feature is set through an algorithm that tracks the assessment progress of a given test taking session.

12) A current level and target level is identified by the ODA assessment.

The image below shows a more detailed representation of the ODA Database for each content area (one database for reading and one database for listening for each of the foreign languages available):



(DLIFLC ODA Program Review, 2015).

APPENDIX B

Example of ODA Diagnostic Profile

4/4/2017

ODAS -- Diagnostic Profile

Listening Assessment

Language: Spanish
Date of the Diagnostic Session: 4/4/2017
Time Spent for the Session: 115 Minutes
Name: [REDACTED]

Based on your performance in this ODA session, your ILR proficiency level estimate is **2** (Current Level).

Your goal is to work toward proficiency level **2+** (Target Level).

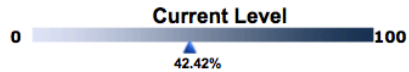
Note: The primary purpose of ODA is to provide you with formative feedback - feedback to help you in the learning process. The ILR level estimate you are given here is intended to function as a reference for charting your progress toward higher proficiency. You may or may not receive the same level at an official test.

The goal is to work incrementally toward your target proficiency level, by learning more about the content areas and the lexical, syntactical, and discourse aspects that you have not yet mastered. The following is a detailed diagnostic feedback on your performance.

Content Questions*



Linguistic Questions**



*Content Questions are all the questions about the meaning of a text, events, details, ideas and arguments.

**Linguistic Questions are those about key vocabulary, sentence structure and relations between ideas.

1 Performance Report - Current Level 2:

Content Questions

	Correct / Total
Main Ideas	2/3 (66.7%)
understand speech about common topics, well-known current events and everyday descriptions and narrations	
Supporting Ideas	4/6 (66.7%)
understand factual content	
understand important factual details	

Linguistic Questions

Vocabulary	6/21 (28.6%)
understand everyday topics, common personal accounts, or well-known current events	
Subject Area Breakdown	
Technology: discoveries, inventions and research	3/7
Society: social issues (e.g. poverty, crime...)	2/7
Geography: natural elements, weather and natural disasters	1/7
Structure	2/3 (66.7%)
understand narrations about current, past and future events	

Structural Feature Breakdown

understand the subject of a sentence when it is not explicitly stated and when it refers to a previously mentioned noun. example: el chofer perdió control del vehículo y chocó. (the driver lost control of the vehicle and [he/the driver] crashed.) 1/1

understand basic time frames in common utterances: past (simple and continuous [imperfect]), present (including present progressive), future, present perfect (has done), and passive voice (was killed). 1/2

Discourse 2/3 (66.7%)

understand basic elements of cohesion (e.g., pronouns, verb inflections)

Discourse Feature Breakdown

understand the way common cohesive devices of addition, such as “además” (besides) and “así como” (as well as), are used to connect ideas in common utterances. 2/2

understand the way direct object pronouns such as “lo” (it, he) and “los” (them) are used in common utterances. 0/1

Speech Processing***

Delivery	Original Version	Modified Version
	Correct / Total	Correct / Attempts
understand speech delivered at a normal rate	6/9	0/1

Vocabulary	Audio Forms	Transcribed Forms
	Correct / Total	Correct / Attempts
understand vocabulary items in their audio forms or with transcription	6/21	6/13

***This section addresses factors that can affect the comprehension of authentic speech, such as speed, clarity, or accent. Depending on the level of the passage, it may also assess the extent to which you needed modified speech (vs. authentic speech) or transcribed vocabulary.

The information under "Modified Version" and "Transcribed Forms" in this section reports on your need for listening help. The information in this column is only applicable if you required modified speech (vs. authentic speech) or transcribed vocabulary.

2 Performance Report - Target Level 2+:

Content Questions

	Correct / Total
Main Ideas	1/6 (16.7%)
understand most common factual topics	
Supporting Ideas	1/12 (8.3%)
understand some discussions on specialized topics	
understand some implied meaning	

Linguistic Questions

Vocabulary	1/42 (2.4%)
understand general vocabulary with occasional limitations	
Subject Area Breakdown	
Security: security issues (e.g. terrorism, trafficking)	0/7
Economy: economic issues	0/7
Society: social issues (e.g. poverty, crime...)	0/14
Science: health and medicine	1/7
Culture: famous people	0/7
Structure	3/6 (50%)

Structural Feature Breakdown

recognize the way speakers use the imperfect subjunctive to express contrary-to-fact or hypothetical conditions. example: si los electores realmente reflexionaran antes de votar, tendríamos mejores legisladores. (if the voters really reflected before voting, we would have better legislators.) 0/1

understand the way speakers use the conditional mood to make speculative and hypothetical remarks. example: tendrían que cambiar todo para que eso funcione. (they would have to change everything in order for that to work.) 1/2

understand the way speakers use various verb tenses to describe actions that happened in the past. example: “iban manejando a 90 millas por hora. (they were driving 90 miles an hour.) “las autoridades han arrestado todos los sospechosos hasta el momento.” (the authorities have arrested all the suspects up to this point.) habían sido detenidos por robo de identidad (“they had been detained because of identity theft.) 1/2

recognize the way speakers use the present and past subjunctive to comment on an action. example: es lógico que haya alguien que quiera ocupar mi puesto. (it is logical that there is someone who wants to take my position.) 1/1

Discourse

3/6
(50%)

understand some complex relations between utterances: connections and references

Discourse Feature Breakdown

understand the way speakers use direct object pronouns such as “lo” (it, he) and “los” (them) to refer to previously mentioned nouns. 0/1

understand the way speakers use idiomatic expressions to discuss abstract topics and express opinions. example: por mucho que le demos vueltas al asunto, el tema nos divide. (no matter how much go around and around on the issue, the topic divides us.) 0/1

understand the way speakers use some cohesive devices of contrast, such as “aunque” (even though) and “a pesar de” (in spite of), to contrast ideas in complex utterances. 2/3

understand the way speakers use rhetorical questions to make evaluative and/or 1/1

Structural Feature Breakdown

recognize the way speakers use the imperfect subjunctive to express contrary-to-fact or hypothetical conditions. example: si los electores realmente reflexionaran antes de votar, tendríamos mejores legisladores. (if the voters really reflected before voting, we would have better legislators.)	0/1
understand the way speakers use the conditional mood to make speculative and hypothetical remarks. example: tendrían que cambiar todo para que eso funcione. (they would have to change everything in order for that to work.)	1/2
understand the way speakers use various verb tenses to describe actions that happened in the past. example: “iban manejando a 90 millas por hora. (they were driving 90 miles an hour.) “las autoridades han arrestado todos los sospechos hasta el momento.” (the authorities have arrested all the suspects up to this point.) habían sido detenidos por robo de identidad (“they had been detained because of identity theft.)	1/2
recognize the way speakers use the present and past subjunctive to comment on an action. example: es lógico que haya alguien que quiera ocupar mi puesto. (it is logical that there is someone who wants to take my position.)	1/1

APPENDIX C

Listening Interagency Language Roundtable Descriptors

Preface

The following proficiency level descriptions characterize comprehension of the spoken language. Each of the six "base levels" (coded 00, 10, 20, 30, 40, and 50) implies control of any previous "base levels" functions and accuracy. The "plus level" designation (coded 06, 16, 26, etc.) will be assigned when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level." The "plus level" descriptions are therefore supplementary to the "base level" descriptions. A skill level is assigned to a person through an authorized language examination. Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively. Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details. Unless otherwise specified, the term "native listener" refers to native speakers and listeners of a standard dialect. "Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education. However, in cultures where formal higher education is common, the language-use abilities of persons who have had such education is considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language.

Listening 0 (No Proficiency): No practical understanding of the spoken language. Understanding is limited to occasional isolated words with essentially no ability to comprehend communication. (Has been coded L-0 in some nonautomated applications. [Data Code 00])

Listening 0+ (Memorized Proficiency): Sufficient comprehension to understand a number of memorized utterances in areas of immediate needs. Slight increase in utterance length understood but requires frequent long pauses between understood phrases and repeated requests on the listener's part for repetition. Understands with reasonable accuracy only when this involves short memorized utterances or formulae. Utterances understood are relatively short in length. Misunderstandings arise due to ignoring or inaccurately hearing sounds or word endings (both inflectional and non-inflectional), distorting the original meaning. Can understand only with difficulty even such people as teachers who are used to speaking with non-native speakers. Can understand best those statements where context strongly supports the utterance's meaning. Gets some main ideas. (Has been coded L-0+ in some nonautomated applications.) [Data Code 06]

Listening 1 (Elementary Proficiency): Sufficient comprehension to understand utterances about basic survival needs and minimum courtesy and travel requirements in areas of immediate need or on very familiar topics, can understand simple questions and answers, simple statements and very simple face-to-face conversations in a standard dialect. These must often be delivered more clearly than normal at a rate slower than normal with frequent repetitions or paraphrase (that is, by a native used to dealing with foreigners). Once learned, these sentences can be varied for similar level vocabulary and grammar and still be understood. In the majority of utterances, misunderstandings arise

due to overlooked or misunderstood syntax and other grammatical clues. Comprehension vocabulary inadequate to understand anything but the most elementary needs. Strong interference from the candidate's native language occurs. Little precision in the information understood owing to the tentative state of passive grammar and lack of vocabulary. Comprehension areas include basic needs such as: meals, lodging, transportation, time and simple directions (including both route instructions and orders from customs officials, policemen, etc.). Understands main ideas. (Has been coded L-1 in some nonautomated applications.) [Data Code 10]

Listening 1+ (Elementary Proficiency, Plus): Sufficient comprehension to understand short conversations about all survival needs and limited social demands. Developing flexibility evident in understanding a range of circumstances beyond immediate survival needs. Shows spontaneity in understanding by speed, although consistency of understanding is uneven. Limited vocabulary range necessitates repetition for understanding. Understands more common time forms and most question forms, some word order patterns, but miscommunication still occurs with more complex patterns. Cannot sustain understanding of coherent structures in longer utterances or in unfamiliar situations. Understanding of descriptions and the giving of precise information is limited. Aware of basic cohesive features (e.g., pronouns, verb inflections) but many are unreliably understood, especially if less immediate in reference. Understanding is largely limited to a series of short, discrete utterances. Still has to ask for utterances to be repeated. Some ability to understand facts. (Has been coded L-1+ in some nonautomated applications.) [Data Code 16]

Listening 2 (Limited Working Proficiency): Sufficient comprehension to understand conversations on routine social demands and limited job requirements. Able to understand face-to-face speech in a standard dialect, delivered at a normal rate with some repetition and rewording, by a native speaker not used to dealing with foreigners, about everyday topics, common personal and family news, well-known current events and routine office matters through descriptions and narration about current, past and future events; can follow essential points of discussion or speech at an elementary level on topics in his/her special professional field. Only understands occasional words and phrases of statements made in unfavorable conditions, for example through loudspeakers outdoors. Understands factual content. Native language causes less interference in listening comprehension. Able to understand facts; i.e., the lines but not between or beyond the lines. (Has been coded L-2 in some nonautomated applications.) [Data Code 20]

Listening 2+ (Limited Working Proficiency, Plus): Sufficient comprehension to understand most routine social demands and most conversations on work requirements as well as some discussions on concrete topics related to particular interests and special fields of competence. Often shows remarkable ability and ease of understanding, but under tension or pressure may break down. Candidate may display weakness or deficiency due to inadequate vocabulary base or less than secure knowledge of grammar and syntax. Normally understands general vocabulary with some hesitant understanding of everyday vocabulary still evident. Can sometimes detect emotional overtones. Some ability to understand implications. (Has been Coded L-2+ in some nonautomated applications.) [Data Code 26]

Listening 3 (General Professional Proficiency): Able to understand the essentials of all speech in a standard dialect including technical discussions within a special field. Has effective understanding of face-to-face speech, delivered with normal clarity and speed in a standard dialect on general topics and areas of special interest; understands hypothesizing and supported opinions. Has broad enough vocabulary that rarely has to ask for paraphrasing or explanation. Can follow accurately the essentials of conversations between educated native speakers, reasonably clear telephone calls, radio broadcasts, news stories similar to wire service reports, oral reports, some oral technical reports and public addresses on non-technical subjects; can understand without difficulty all forms of standard speech concerning a special professional field. Does not understand native speakers if they speak very quickly or use some slang or dialect. Can often detect emotional overtones. Can understand implications. (Has been coded L-3 in some nonautomated applications.) [Data Code 30]

Listening 3+ (General Professional Proficiency, Plus): Comprehends most of the content and intent of a variety of forms and styles of speech pertinent to professional needs, as well as general topics and social conversation. Ability to comprehend many sociolinguistic and cultural references. However, may miss some subtleties and nuances. Increased ability to comprehend unusually complex structures in lengthy utterances and to comprehend many distinctions in language tailored for different audiences. Increased ability to understand native speakers talking quickly, using nonstandard dialect or slang; however, comprehension is not complete. Can discern some relationships among sophisticated listening materials in the context of broad experience. Can follow some unpredictable turns of thought readily, for example, in informal and formal speeches

covering editorial, conjectural and literary material in subject matter areas directed to the general listener. (Has been coded L-3+ in some nonautomated applications.) [Data Code 36]

Listening 4 (Advanced Professional Proficiency): Able to understand all forms and styles of speech pertinent to professional needs. Able to understand fully all speech with extensive and precise vocabulary, subtleties and nuances in all standard dialects on any subject relevant to professional needs within the range of his/her experience, including social conversations; all intelligible broadcasts and telephone calls; and many kinds of technical discussions and discourse. Understands language specifically tailored (including persuasion, representation, counseling and negotiating) to different audiences. Able to understand the essentials of speech in some non-standard dialects. Has difficulty in understanding extreme dialect and slang, also in understanding speech in unfavorable conditions, for example through bad loudspeakers outdoors. Can discern relationships among sophisticated listening materials in the context of broad experience. Can follow unpredictable turns of thought readily, for example, in informal and formal speeches covering editorial, conjectural and literary material in any subject matter directed to the general listener. (Has been coded L-4 in some nonautomated applications.) [Data Code 40]

Listening 4+ (Advanced Professional Proficiency, Plus): Increased ability to understand extremely difficult and abstract speech as well as ability to understand all forms and styles of speech pertinent to professional needs, including social conversations. Increased ability to comprehend native speakers using extreme nonstandard dialects and slang, as well as to understand speech in unfavorable conditions. Strong sensitivity to

sociolinguistic and cultural references. Accuracy is close to that of the well-educated native listener but still not equivalent. (Has been coded L-4+ in some nonautomated applications.) [Data Code 46]

Listening 5 (Functionally Native Proficiency): Comprehension equivalent to that of the well-educated native listener. Able to understand fully all forms and styles of speech intelligible to the well-educated native listener, including a number of regional and illiterate dialects, highly colloquial speech and conversations and discourse distorted by marked interference from other noise. Able to understand how natives think as they create discourse. Able to understand extremely difficult and abstract speech.

APPENDIX C

Reading Interagency Language Roundtable Descriptors

Preface

The following proficiency level descriptions characterize comprehension of the written language. Each of the six "base levels" implies control of any previous "base level's" functions and accuracy. The "plus level" designation will be assigned when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level." The "plus level" descriptions are therefore supplementary to the "base level" descriptions. A skill level is assigned to a person through an authorized language examination.

Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively. Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details.

Unless otherwise specified, the term "native reader" refers to native readers of a standard dialect. "Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education. However, in cultures where formal higher education is common, the language-use abilities of persons who have had such education is considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the

language. In the following descriptions a standard set of text-types is associated with each level. The text-type is generally characterized in each descriptive statement. The word "read," in the context of these proficiency descriptions, means that the person at a given skill level can thoroughly understand the communicative intent in the text-types described. In the usual case the reader could be expected to make a full representation, thorough summary, or translation of the text into English. Other useful operations can be performed on written texts that do not require the ability to "read" as defined above. Examples of such tasks which people of a given skill level may reasonably be expected to perform are provided, when appropriate, in the descriptions.

Reading 0 (No Proficiency): No practical ability to read the language. Consistently misunderstands or cannot comprehend at all.

Reading 0+ (Memorized Proficiency): Can recognize all the letters in the printed version of an alphabetic system and high-frequency elements of a syllabary or a character system. Able to read some or all of the following: numbers, isolated words and phrases, personal and place names, street signs, office and shop designations. The above often interpreted inaccurately. Unable to read connected prose.

Reading 1 (Elementary Proficiency): Sufficient comprehension to read very simple connected written material in a form equivalent to usual printing or typescript. Can read either representations of familiar formulaic verbal exchanges or simple language containing only the highest frequency structural patterns and vocabulary, including shared international vocabulary items and cognates (when appropriate). Able to read and understand known language elements that have been recombined in new ways to achieve different meanings at a similar level of simplicity. Texts may include descriptions of

persons, places or things: and explanations of geography and government such as those simplified for tourists. Some misunderstandings possible on simple texts. Can get some main ideas and locate prominent items of professional significance in more complex texts. Can identify general subject matter in some authentic texts.

Reading 1+ (Elementary Proficiency, Plus): Sufficient comprehension to understand simple discourse in printed form for informative social purposes. Can read material such as announcements of public events, simple prose containing biographical information or narration of events, and straightforward newspaper headlines. Can guess at unfamiliar vocabulary if highly contextualized, but with difficulty in unfamiliar contexts. Can get some main ideas and locate routine information of professional significance in more complex texts. Can follow essential points of written discussion at an elementary level on topics in his/her special professional field. In commonly taught languages, the individual may not control the structure well. For example, basic grammatical relations are often misinterpreted, and temporal reference may rely primarily on lexical items as time indicators. Has some difficulty with the cohesive factors in discourse, such as matching pronouns with referents. May have to read materials several times for understanding.

Reading 2 (Limited Working Proficiency): Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform

sorting and locating tasks with written texts that are well beyond their general proficiency level. The individual can read uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straightforward/high-frequency sentence patterns. The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. Is typically able to answer factual questions about authentic texts of the types described above.

Reading 2+ (Limited Working Proficiency, Plus): Sufficient comprehension to understand most factual material in non-technical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance and/or intentionally disguised meaning.

Reading 3 (General Professional Proficiency): Able to read within a normal range of

speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and "read between the lines," (that is, understand the writers' implicit intents in text of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms.

Reading 3+ (General Professional Proficiency, Plus): Can comprehend a variety of styles and forms pertinent to professional needs. Rarely misinterprets such texts or rarely experiences difficulty relating ideas or making inferences. Able to comprehend many sociolinguistic and cultural references. However, may miss some nuances and subtleties. Able to comprehend a considerable range of intentionally complex structures, low frequency idioms, and uncommon connotative intentions, however, accuracy is not complete. The individual is typically able to read with facility, understand, and appreciate contemporary expository, technical or literary texts which do not rely heavily on slang and unusual items.

Reading 4 (Advanced Professional Proficiency): Able to read fluently and accurately

all styles and forms of the language pertinent to professional needs. The individual's experience with the written language is extensive enough that he/she is able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references. Able to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment). Able to read and understand the intent of writers' use of nuance and subtlety. The individual can discern relationships among sophisticated written materials in the context of broad experience. Can follow unpredictable turns of thought readily in, for example, editorial, conjectural, and literary texts in any subject matter area directed to the general reader. Can read essentially all materials in his/her special field, including official and professional documents and correspondence. Recognizes all professionally relevant vocabulary known to the educated non-professional native, although may have some difficulty with slang. Can read reasonably legible handwriting without difficulty. Accuracy is often nearly that of a well-educated native reader.

Reading 4+ (Advanced Professional Proficiency, Plus): Nearly native ability to read and understand extremely difficult or abstract prose, a very wide variety of vocabulary, idioms, colloquialisms and slang. Strong sensitivity to and understanding of sociolinguistic and cultural references. Little difficulty in reading less than fully legible handwriting. Broad ability to "read beyond the lines" (that is, to understand the full ramifications of texts as they are situated in the wider cultural, political, or social environment) is nearly that of a well-read or well-educated native reader. Accuracy is close to that of the well-educated native reader, but not equivalent.

Reading 5 (Functionally Native Proficiency): Reading proficiency is functionally

equivalent to that of the well-educated native reader. Can read extremely difficult and abstract prose; for example, general legal and technical as well as highly colloquial writings. Able to read literary texts, typically including contemporary avant-garde prose, poetry and theatrical writing. Can read classical/archaic forms of literature with the same degree of facility as the well-educated, but non-specialist native. Reads and understands a wide variety of vocabulary and idioms, colloquialisms, slang, and pertinent cultural references. With varying degrees of difficulty, can read all kinds of handwritten documents. Accuracy of comprehension is equivalent to that of a well-educated native reader.