MEASURING HERITAGE LANGUAGE LEARNERS’ PROFICIENCY FOR RESEARCH PURPOSES: AN ARGUMENT-BASED VALIDITY STUDY OF THE KOREAN C-TEST

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Young-A Son, M.A.

Washington, DC
September 24, 2018
MEASURING HERITAGE LANGUAGE LEARNERS’ PROFICIENCY FOR RESEARCH PURPOSES: AN ARGUMENT-BASED VALIDITY STUDY OF THE KOREAN C-TEST

Young-A Son, M.A.

Thesis Advisor: Lourdes Ortega, Ph.D.

ABSTRACT

Heritage language learners (HLLs) have increasingly become a focus of interest in applied linguistics research (Kagan & Dillon, 2012), but the lack of consistent conceptualization of HL proficiency has hindered the systematic accumulation of research knowledge about HLLs (Son, 2017). Shortcut proficiency measures may be a way to address these shortcomings (Norris, 2018). C-tests have been found to be particularly promising in providing a quick measurement of language learners’ global proficiency (Eckes & Grotjahn, 2006), yet validation studies on this shortcut measure have focused on Foreign Language Learners (FLLs).

To address these critical gaps, this study developed a validity argument (Kane 2006, 2011, 2013) to evaluate the use of an innovative Korean C-test (Son et al., 2018) to assess Korean HLLs and FLLs for applied linguistics research purposes. Ninety-three Korean language learners, 41 HLLs and 52 FLLs, were assessed using five instruments: the Korean C-test, an Elicited Imitation Test (EIT, Kim et al., 2016), ACTFL Oral Proficiency Interview-computer (OPIc), ACTFL Writing Proficiency test (WPT), and a self-assessment questionnaire. The resulting data were then examined for five inferences— theoretical grounds, evaluation, generalization, explanation, and extrapolation—in terms of warrants, assumptions, backing, and rebuttals.
C-test items accurately and reliably distributed Korean learners into a wide range of proficiency levels (IRT person separation index=5.50 and Cronbach’s α=.94). Although C-test performance was closely related to the literacy-based WPT proficiency indicator (ρ=.87), it also correlated strongly with the oracy-based OPIc and EIT proficiency measures (ρ=.81 or higher). Multiple regression analyses showed that both speaking and writing proficiency could explain 80% of the C-test score variance with $R^2 = .80$, $F(2,90) = 181.83, p = .000$. Nevertheless, writing was more important than speaking proficiency in predicting C-test performance for HLLs, whereas both language skill predictors were equally important for FLLs. Furthermore, a hierarchical cluster analysis provided a bottom-up categorization of learners into HLLs and FLLs and questioned the expectation that HLLs are always linguistically different to FLLs. Overall, the evidence supported the use of the Korean C-test to assess both HLLs and FLLs across a range of proficiency levels for applied linguistic research purposes. The study concludes with suggestions for future research.
ACKNOWLEDGMENTS

This dissertation study would not have been possible without the love and support from multiple people and organizations. I would like to first thank the 10 Korean language programs, who believed in the importance of this research for advancing research on Korean language learners and opened their doors for me to conduct this research: Georgetown University (Prof. Min Koo Choi), George Washington University (Prof. Miok Pak), Georgia State University (Prof. Hakyoon Lee), Emory University (Prof. Bumyong Choi), Duke University (Prof. Hae-Young Kim and Prof. Eunyoung Kim), University of North Carolina Chapel Hill (Prof. Jia Lin and Prof. Eunji Lee), Pennsylvania State University (Prof. Sorin Huh), Harvard University (Prof. Hi-Sun Kim), Rutgers University (Prof. Young-Mee Yu Cho and Prof. Hee-Chung Chun), University of Hawaii at Manoa (Prof. Mee-Jeong Park).

I would also like to thank the Assessment and Evaluation Language Resource Center (AELRC) for making it possible to conduct a large-scale validation study and funding the administration of most of the instruments. It was only through the AELRC that the development of the Korean C-test was possible. Thank you to the Korean C-test development team: John D., Amy, and Eunyoung. I learned so much from our discussions and Zoom conversations. Many thanks to the National Federation of Modern Language Teachers’ Association (NFMLTA), the Educational Testing Service (ETS), the Graduate School of Arts and Sciences at Georgetown University, and ACTFL for funding part of this research and making it possible to include learners that are representative of a full range of proficiencies from schools across the United States.
I am also deeply indebted to my dissertation committee members who supported me every step of the way from the beginning, when my dissertation was just an idea, to the very end. Dr. Ortega, muchas gracias por todo el apoyo y sobre todo por siempre motivarme a seguir adelante. Dr. Norris, thank you for always challenging me to think outside of the box and for being the master mind behind this dissertation. Dr. Malone (Meg) thank you so very much for all the advice not only about my dissertation and academic life but also about life in general.

To my Georgetown friends, who played such an important role in the completion of my degree: Thank you! Yuka, John, Mari, Lauren, Eunji, Jongbong, Rachel, Amy, Mina, Brandon, Bart, Todd, Yiran, Sakol, Julia, Angela, Fatima thank you for being there for me and for making the last 5 years so memorable.

Finally, I would like to thank my family who was so patient during my whole academic career (MA included) and believed I could finish someday. 엄마, 아빠, 언니, 감사합니다! Most importantly, I would like to thank my husband, Jörn. During the whole process he was my life coach, travel agent, career coach, copy-editor, statistics-consultant, and I owe the completion of this life chapter to him most of all. Danke schön und ich liebe Dich!
# TABLE OF CONTENTS

Chapter 1: Introduction ........................................................................................................... 1

1.1. Heritage Language Assessment Research Challenges ................................................. 3

1.2. Purpose of the Dissertation Study .............................................................................. 8

1.3. Structure of the Dissertation ..................................................................................... 10

Chapter 2: Language Proficiency ....................................................................................... 12

2.1. Defining Second Language Proficiency ..................................................................... 12

2.1.1. Models of second language proficiency ............................................................... 12

2.1.2. Defining second language proficiency through expert opinions ......................... 23

2.2. Measuring Second Language Proficiency in Second Language Acquisition Research ........................................................................................................................ 33

2.3 Defining Heritage Language Learners (HLLs) ............................................................. 38

2.3.1. Methods for identifying heritage language learners ............................................. 43

2.4. Defining Heritage Language Proficiency .................................................................. 45

2.4.1. A general second language acquisition approach to defining heritage language proficiency ......................................................................................................................... 48

2.4.2. A holistic approach to defining heritage language proficiency ......................... 52

2.4.2.1. Linguistic profiles of Korean heritage language learners ............................... 55

2.5. Measuring Heritage Language Proficiency ............................................................... 61

2.5.1. Instructional level .................................................................................................. 62

2.5.2. Scores on standardized tests of language proficiency ........................................ 63

2.5.3. Self-assessment scores ........................................................................................ 65

2.5.4. Scores on locally-designed language proficiency tests .................................... 68

2.6. Summary .................................................................................................................... 70
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.</td>
<td>What are C-tests?</td>
<td>72</td>
</tr>
<tr>
<td>3.2.</td>
<td>Korean C-tests</td>
<td>75</td>
</tr>
<tr>
<td>3.2.1.</td>
<td>Previous Korean C-test developed by Lee-Ellis (2009)</td>
<td>75</td>
</tr>
<tr>
<td>3.2.2.</td>
<td>New Korean C-test developed by Son, Kim, Cho, and Davis (2018)</td>
<td>79</td>
</tr>
<tr>
<td>3.3.</td>
<td>C-test in the Heritage Language Literature</td>
<td>81</td>
</tr>
<tr>
<td>4.1.</td>
<td>Trinitary Model</td>
<td>85</td>
</tr>
<tr>
<td>4.2.</td>
<td>Redefinition of Construct Validity</td>
<td>87</td>
</tr>
<tr>
<td>4.3.</td>
<td>Unified Model of Validity</td>
<td>90</td>
</tr>
<tr>
<td>4.4.</td>
<td>Argument-Based Approach to Validity</td>
<td>94</td>
</tr>
<tr>
<td>4.4.1.</td>
<td>Underlying inference model</td>
<td>95</td>
</tr>
<tr>
<td>4.4.2.</td>
<td>Describing the argument structure</td>
<td>97</td>
</tr>
<tr>
<td>4.4.3.</td>
<td>Discussing the advantages of the model</td>
<td>102</td>
</tr>
<tr>
<td>5.1.</td>
<td>Korean C-test Context and Intended Uses</td>
<td>107</td>
</tr>
<tr>
<td>5.2.</td>
<td>Inferences, Warrants, and Underlying Assumptions, and Possible Rebuttals</td>
<td>109</td>
</tr>
<tr>
<td>5.3.</td>
<td>Research Questions</td>
<td>117</td>
</tr>
<tr>
<td>6.1.</td>
<td>Participants</td>
<td>120</td>
</tr>
<tr>
<td>6.2.</td>
<td>Instruments</td>
<td>124</td>
</tr>
<tr>
<td>6.2.1.</td>
<td>Korean C-test (Son, Kim, Cho, &amp; Davis, 2018)</td>
<td>125</td>
</tr>
<tr>
<td>6.2.2.</td>
<td>Korean elicited imitation test (Kim, Tracy-Ventura, &amp; Jung, 2016)</td>
<td>126</td>
</tr>
</tbody>
</table>
6.2.3. Korean ACTFL Oral Proficiency Interview-computer and Writing Proficiency Test

6.2.4. Background and self-assessment questionnaire

6.3. Data Collection Procedure

Chapter 7: Results

7.1. Outcomes Relevant to the Theoretical Grounds Inference

7.1.1. Theoretical characterizations of C-test constructs

7.1.2. Summary of theoretical grounds inference

7.2. Outcomes Relevant to the Evaluation Inference

7.2.1. Text selection

7.2.2. Statistical characteristics of the C-test items

7.2.3. Item response theory measures for C-test items

7.2.4. Summary of the evaluation inference

7.3. Outcomes Relevant to the Generalization Inference

7.3.1. Reliability measures of C-test items

7.3.2. Differential item functioning for heritage language learners and foreign language learners

7.3.3. Possible rebuttal: Inaccurate definition of heritage language learners

7.3.3.1. Cluster analysis methodology

7.3.3.2. Cluster analysis results

7.3.3.3. Descriptive statistics, reliability measures, and Differential Item Functioning (DIF) analysis on the new clusters

7.3.4. Summary of the outcomes relevant to the generalization inference

7.4. Outcomes Relevant to the Explanation Inference

7.4.1. Korean C-test item characteristics and item difficulty
8.3. Conclusion ................................................................................................................. 285
Appendix A: Korean C-test – Texts and Answers......................................................... 289
Appendix B: Elicited Imitation Test Rubric with Examples........................................ 292
Appendix C: Elicited Imitation Test Example Matrix .................................................. 294
Appendix D: Background and Self-Assessment Questionnaire ................................. 298
Appendix E: Item Characteristic Curves for Seven Items .......................................... 310
Appendix F: Seven-Item C-test Item Map ................................................................. 311
Appendix G: Cluster Analysis Dendrogram............................................................... 312
Appendix H: Descriptive Statistics for 10 Variables across Four Clusters............... 313
Appendix I: Numeric Values for the ACTFL Proficiency Levels (from Vande Berg,
Connor-Linton, & Paige, 2009) .............................................................................. 315
Appendix J: Boxplots of Korean C-test, ACTFL Oral Proficiency Interview-computer,
ACTFL Writing Proficiency Test, and Elicited Imitation Test ......................... 316
Appendix K: Residual Scatterplots ............................................................................. 317
Appendix L: Bivariate Correlation Scatterplots ......................................................... 318
References .................................................................................................................. 319
LIST OF FIGURES

Figure 1. Timeline of different models and theories of L2 proficiency .............................. 13
Figure 2. Hulstijn’s (2015) representation of the BLC-HLC theory and its components. 22
Figure 3. Valdés’ (2005) representation of L1/L2 users bilingual continuum ..................... 46
Figure 4. Messick’s (1989) matrix of the facets of validity .................................................. 92
Figure 5. Example of Toulmin’s (2003) model of inference .............................................. 96
Figure 6. The bridge analogy for the interpretative structure ............................................. 98
Figure 7. Validity argument structure for TOEFL based on Chapelle et al. (2010) .......... 100
Figure 8. Comparison between participants’ first and most comfortable language ...... 124
Figure 9. Data collection procedure .............................................................................. 135
Figure 10. Korean C-test score distribution (N = 93) ......................................................... 151
Figure 11. C-test mean scores and 95% confidence intervals per item ............................. 153
Figure 12. Seven-text C-test item map (N = 93) ............................................................... 156
Figure 13. Four-Text C-test Item Map (N = 88) ................................................................. 158
Figure 14. Differential item functioning for HLLs vs. FLLs ................................................. 167
Figure 15. Distance between agglomeration coefficients at each cluster stage .......... 172
Figure 16. ACTFL test scores comparison across four clusters ........................................ 174
Figure 17. Length of stay studying in Korea across four clusters ..................................... 174
Figure 18. Learners’ Korean language use with family members across four clusters . 175
Figure 19. Learners’ frequency of Korean language use across four clusters ............... 175
Figure 20. Differential item functioning for Cluster 1 vs. Cluster 2 ................................. 183
Figure 21. Word count by vocabulary level ...................................................................... 193
Figure 22. Item map for all micro-items in the C-test (N = 83) ........................................ 194
| Figure 23. | Comparison of micro-item difficulty between HLLs and FLLs | 198 |
| Figure 24. | Scatterplot of bivariate correlation between ACTFL WPT and C-test scores | 202 |
| Figure 25. | Distribution of ACTFL OPIc (left) and ACTL WPT (right) | 216 |
| Figure 26. | Scatterplot of bivariate correlation between C-test and ACTFL OPIc and WPT mean scores | 219 |
| Figure 27. | Error bar plot for C-test mean scores across eight ACTFL proficiency levels | 224 |
| Figure 28. | Error bar plots for C-test mean scores across three main ACTFL proficiency levels | 226 |
| Figure 29. | Steps representing the Korean C-test validity argument | 242 |
| Figure 30. | Representation of evaluation with three assumptions, backing, and rebuttal | 248 |
| Figure 31. | Representation of generalization with four assumptions, backing, and rebuttal | 256 |
| Figure 32. | Representation of explanation with three assumptions, backing, and rebuttal | 266 |
| Figure 33. | Representation of extrapolation with three assumptions, backing, and rebuttal | 276 |
LIST OF TABLES

Table 1. NCSSFL-ACTFL Can-Do Statements for Interpersonal Communication........ 26
Table 2. CEFR Global Scale .................................................................................. 30
Table 3. Characteristics of L1, L2, Heritage Language Acquisition .................... 47
Table 4. Summary of Inferences, Warrants, and Assumption in the Korean C-test
         Interpretive Argument............................................................................. 110
Table 5. Participant Demographics..................................................................... 122
Table 6. Scoring Method and Examples .............................................................. 126
Table 7. Tasks and Abilities of the WPT by Proficiency Level ............................. 131
Table 8. Theoretical Grounds Inference Warrants, Assumptions, and Research
         Questions.................................................................................................. 138
Table 9. Evaluation Inference Warrants, Assumptions, and Research Questions..... 146
Table 10. Item Fit Statistics for Seven Items (N = 35) .......................................... 148
Table 11. C-test Descriptive Statistics and Reliability Estimates........................ 150
Table 12. Kolmogorov-Smirnov Test of Normality ............................................. 152
Table 13. Descriptive Statistics for C-test Items .................................................. 153
Table 14. Summary Statistics for Person and Item Fit.......................... ................. 155
Table 15. Item Fit Statistics for Four Items (N = 88) ........................................... 157
Table 16. Generalization Inference Warrant, Assumptions, and Research Questions ... 162
Table 17. Reliability Measures for C-test when Including Seven and Four Items .... 163
Table 18. Reliability Measures for each Item on the C-test .................................. 164
Table 19. Comparison of HLLs and FLLs Performance on C-test Items .............. 166
Table 20. Variables Considered for Cluster Analysis.......................................... 170
Table 21. Crosstabulation of Learner Types and Cluster Groups.......................... 173
Table 22. Descriptive Statistics and Reliability Measures for Different Cluster Solutions

Table 23. Explanation Inference Warrant, Assumptions, and Research Questions

Table 24. Number of Clauses per Sentence across All C-test Passages

Table 25. Number of Content and Function Words across Texts

Table 26. Ten Most Difficult and Easiest Micro-Items

Table 27. ACTFL WPT Descriptive Statistics

Table 28. Multiple Regression Model Summary

Table 29. Regression Coefficients for the Criterion Variable: C-test Performance

Table 30. Multiple Regression Model Summary for the HLL and FLL Group

Table 31. Regression Coefficients for Each Learner Group

Table 32. Extrapolation Inference Warrant, Assumptions, and Research Questions

Table 33. Descriptive Statistics for ACTFL OPIc, WPT, OPIc & WPT Mean, and EIT

Table 34. Spearman Rank Order Correlation Coefficients between C-test and Other Criterion Measures

Table 35. Spearman Rank Order Correlation Coefficients between C-test and Other Criterion Measures by Learner Type

Table 36. Spearman Rank Order Correlation Coefficients between C-test and Other Criterion Measures of Intermediate-Level HLLs and FLLs

Table 37. Comparison of ACTFL Proficiency Numerical Scales

Table 38. Descriptive Statistics of the Korean C-test Scores across Eight ACTFL Proficiency Levels

Table 39. Descriptive Statistics of the Korean C-test Scores across ACTFL Main Proficiency Levels

Table 40. Pairwise Comparisons of C-test Performance between ACTFL Main Proficiency Levels
Table 41. Cross Tabulation of C-test Mean Scores and ACTFL Main Proficiency Levels ................................................................. 228

Table 42. Cross Tabulation of C-test Mean Scores and ACTFL Main Proficiency Levels ........................................................................................................ 229

Table 43. Descriptive Statistics: Self-Assessment of Four Korean Language Skills ..... 230

Table 44. Descriptive Statistics: Self-Assessment of Four Korean Language Skills for Each Learner Type ................................................................. 231

Table 45. Pearson Correlation Coefficients between C-test Performance and Four Korean Language Skills ........................................................................ 232

Table 46. Pearson Correlation Coefficients between C-test Performance and Four Language Skills by Learner Group ......................................................... 233

Table 47. Test Use Inference Warrant, Assumptions, and Research Questions .......... 238
Chapter 1

Introduction

Heritage language learners (HLLs) have become a major focus in research on language acquisition. This interest is partly due to an increasing influx of migrants to the US which, as multiple scholars (e.g., Beaudrie & Ducar, 2012; Leeman & King, 2015; Kagan & Dillon, 2012) note, has in turn meant that enrollment rates in some foreign language classrooms are soaring. Although the definition of HLL is still a contended topic, from the language education standpoint, they have generally been described as learners who have been raised in a home where a language different to the majority language, or the language of the host country, is spoken; who have a personal connection to this home language and its culture; and who have some degree of proficiency in that language (Fishman, 2001; Valdés, 2001; Van Deusen-Scholl, 2003). In terms of what sets them apart from foreign language learners (FLLs), most of the research (see Kondo-Brown, 2010; Montrul, 2010) has pointed out that there are important differences in the way they process and acquire languages. In particular, relative to FLLs, the skills and experiences of HLLs with using written language have been described as being limited, as opposed to their oral language skills which have been reported to be greater than typical FLLs. Thus, the literacy-oracy continuum has been an important and difficult area for research in the investigation of HLLs (Kagan & Dillon, 2004; Kondo-Brown, 2005; Montrul, 2009; Valdés, 2005). In turn, these differences have called for a careful investigation of HLLs’ unique characteristics and most importantly how these characteristics influence their instructional needs.
One of the main challenges of studying HLLs, however, has been defining who they are and what their proficiency is like (Kondo-Brown, 2010; Montrul, 2008; Valdés, 2001). Attempts to tackle this issue have taken primarily two different directions: (a) a general SLA approach to identifying differences between heritage and non-heritage language learners’ abilities (e.g., Kim, Montrul, & Yoon, 2009; Montrul, 2008, 2013; Montrul & Perpiñán, 2011; Pascual y Cabo, & Gómez, 2015; Silva-Corvalán, 2016), and (b) a holistic approach to defining HL proficiency for educational research (e.g., Fairclough, 2012; Kagan, 2005; Kagan & Dillon, 2008, 2012; Kondo-Brown, 2003, 2010; Lee & Shin, 2008; Malone, Peyton, & Kim, 2014; Sekerina, 2013). Research within the first category has intended to describe what heritage learners can do and how their early exposure to the heritage language (HL) may influence the way they process and learn the language. Studies within the second approach have focused on practical issues in the educational context, such as identifying and addressing HLLs’ academic needs including designing curricula and assessment tools specifically oriented towards educating HLLs.

Although both approaches to HLL research have made contributions to our understanding of HL learning and HL educational needs, there is still limited systematic accumulation of knowledge in either approach. This limitation in the research is largely due to a lack of consistency in how HL proficiency has been measured. The multiple and different approaches to measuring HL proficiency have hindered the comparability of results across studies. The definition of a low-proficiency level HLL in a study that assessed HL proficiency through a standardized test, for example, might be completely different from the same-level HLL in another study that assessed proficiency through a
self-assessment questionnaire. The limited number of and access to reliable assessment tools specifically designed to assess HLLs can be one of the main causes for the lack of agreement across scholarly work. In addition, there has been a lack of research focused on teasing apart differences in the performance of HLLs and FLLs on these assessments, which have generally been designed in the first place for the purposes of measuring FLLs. Moreover, there is still no clear understanding of the construct of language proficiency when it comes to HLLs.

In order to address the aforementioned challenges, this dissertation research attempted to make a contribution to the HLL research community, particularly in the area of researching and developing assessments of a less commonly taught heritage language (LCTHL), Korean, by investigating the validity of a shortcut technique for assessing global proficiency, namely the C-test. This type of shortcut measure has gained increasing popularity for its highly practical and effective way of measuring language learners’ global proficiency as well as reliably distributing them in a wide range of proficiency levels (e.g., Norris, 2018; Eckes & Grotjahn, 2006 on C-tests). Therefore, this tool was thought could provide a solution for the urgent need for reliable measures of general language proficiency for HL research purposes.

1.1. Heritage Language Assessment Research Challenges

In an interview with Guadalupe Valdés (Son, 2016), one of the earliest and most established experts in the field of Heritage Language Acquisition (HLA), she defined heritage language education as the “stepchild of foreign language education.” She explained that compared to FLLs, HLLs have not been considered legitimate or
mainstream language learners, and that for this reason there has been a general
misconception to instruct HLLs the same way as FLLs. Although, this statement was
intended to describe HL education, it also applies to the current state of HLL research,
particularly research oriented towards the assessment of HLLs. As with HL instruction,
HLLs have not been considered a “legitimate” group of language learners when it comes
to assessing their language abilities. There has been a highly pervasive belief that the
same methods and instruments used to measure FLLs’ language ability can also be
applied and used for assessing HLLs’ language abilities, without prior validation of this
claim. In addition, as the “stepchild” of foreign language education, some of the pitfalls
of SLA research have been in certain ways inherited by HLA research. One noteworthy
pitfall has been the lack of reliable tools that can assess language proficiency and that can
help interpret and generalize results across studies. In this section, I highlight some of
these persistent gaps in the literature of HL assessment (for a review see also Son, 2017).

First, it should come as no surprise that HL proficiency has been ill-defined and
taken for granted, as this is also the case of proficiency in general in the SLA literature
(Hulstijn, 2011). There remains a dearth of literature addressing the definition of HLLs’
language proficiency. It is still unclear what HLLs at an advanced-level of language
proficiency, for example, are able to do that is different from intermediate- or beginner-
level HLLs. Unlike the assessment of language proficiency for FLLs, there are no
specific proficiency guidelines that can provide benchmarks for assessing HLLs (Kagan
& Dillon, 2012; Malone et al., 2014). This lack of conceptualization of HL proficiency
has been an added complexity to the interpretation of research findings that involve
language proficiency as an independent or moderating variable.
Second, as a consequence of the lack of proper conceptualization of HL proficiency, there has been overall confusion in the way HL proficiency can be operationalized. When describing the L2 proficiency measures in the SLA research, Thomas (1994, 2006) suggested in both of her syntheses that there was little understanding about what proficiency is and how it can be measured. She identified four different ways in which language proficiency has been operationalized in the SLA research domain. These indicators fit well with the operationalizations found in the HLA research. A preliminary review of the HLL literature by the present author revealed four different indicators of proficiency levels, namely, (a) institutional level (e.g., “first year”, “second semester”), (b) scores of standardized proficiency tests (e.g., OPI, DELE), (c) self-assessment scores (e.g., via Likert scales), and (d) scores of locally-designed tests of proficiency (e.g., an existing placement test in a program, or a test designed for a specific study). As Norris and Ortega (2012) have pointed out, this variability of proficiency measures across studies challenges the systematic accumulation of knowledge and the comparability/generalizability of results. Two hypothetical studies (e.g., Study A and Study B) that explored the same linguistic phenomenon (e.g., gender agreement) for learners of the same HL group (e.g., Spanish HLLs) could find contradicting results. One study can claim, for example, that HL proficiency greatly influences the outcomes of a certain test, while the other can find no differences across proficiency groups. If Study A used institutional level as a measurement of proficiency and Study B used self-assessment questionnaires, it is unclear whether the results are actually contradicting or whether the differences in proficiency measurement were the determining factor for such contradicting findings. Accordingly, further research should focus on finding ways to
consolidate these different indicators of proficiency into a useful, reliable, and practical method, which would help improve the interpretation of results across different research inquiries.

Third, there is a general lack of practical and reliable instruments available to the research community to assess HLLs’ general proficiency. Hence, it is not uncommon for researchers to use the same measures of proficiency originally designed for assessing FLLs in order to assess HLLs. In most cases, there is no prior evaluation of the implications for using these tools for such purpose. For example, standardized tests, such as the OPI (see Sekerina, 2013 for more examples) or language proficiency scales, such as the ACTFL proficiency guidelines, are often used to provide an overall idea of HLLs’ proficiency and to classify them into different proficiency groups. Montrul (2008, 2009, 2016) has explained that due to the exposure and use of the HL during childhood, HLLs appear to be at times more similar to native speakers and at other times they align more with FLLs. Hence, it remains unclear whether the same tools used for assessing FLLs’ ability can be as useful for measuring HLLs. More scholarly work, then, should focus on the different ways in which assessment tools are able to elicit HLLs’ language ability, in particular, and more validation studies should be conducted to explore whether various measures are equally useful for assessing HLLs’ global proficiency.

In this regard, although efforts to address the need for reliable and practical instruments for assessing general proficiency in SLA research purposes have led to the development of shortcut measures, including the C-test (e.g., Norris, 2018), these measures have yet to be evaluated for their use to assess HLLs. On the one hand, a large body of academic work has found that these instruments are practical in providing a
quick measurement of language learners’ global proficiency (see Eckes & Grotjahn, 2006), which is useful for research purposes. Furthermore, this type of instruments has provided one solution to the above noted need for more generalizability of findings, as the scores of these tests can be used as an indicator of language proficiency across different studies. On the other hand, most of these measures have mainly targeted adult FLLs as their intended user. As a result, more research will need to focus on how these measures behave with different learner groups, such as FLLs and HLLs. That is, validation studies will need to carefully evaluate the extent to which they elicit similar or different language abilities from such distinct learner groups.

Finally, the lack of understanding of what validation is and what its process entails has undermined the importance of validating testing instruments for the use of assessing HLLs. Norris (2006a, 2016) has noted that there is a widespread confusion within the applied linguistics research about the meaning of terms such as “evaluation,” “assessment,” and “measurement.” These terms are incorrectly being used interchangeably. Along the same lines, validity, or the extent to which evidence supports the interpretation of test scores according to the intended users and uses of the test, has been misinterpreted as an inherent characteristic of a test. In other words, validation has been reduced to a process of investigating whether or not a test possesses this characteristic, rather than the evaluation of test result interpretations. This misconception of the term can be observed in claims such as “this is a valid test” or “the results showed the test is valid,” which can be often found in applied linguistics research. Perhaps, it is due to these misconceptions that research focused on validation has been lagging behind, especially in studies about HLLs. Therefore, more efforts should be focused on raising
awareness of what validation is and how evidence can be gathered to support the use of a given test.

1.2. Purpose of the Dissertation Study

Given the aforementioned critical gaps, the present dissertation study sought to examine the development of a Korean C-test (Son, Kim, Cho, & Davis, in press), and evaluated its use for assessing HLLs in addition to FLLs. More specifically, the study reports on a validation study that thoroughly evaluated the interpretation of test results for the purported intended use of providing a shortcut measure of Korean language learners’ general language proficiency for research purposes. It followed Kane’s (2006, 2011, 2013) argument-based approach to validity by (1) sketching out the interpretive argument chain based on the proposed interpretation and uses of the test, and (2) developing a validation argument that evaluates the coherence of such interpretations and provides relevant evidence to support them. Although the study examined the various components of the argument chain from theoretical grounds to the test utilization (see chapter 5), it primarily focused on the evaluation, generalization, explanation, and extrapolation inferences of the interpretive argument chain. Broadly speaking, these four inferences mainly explored the construct of the test and its applicability to different types of learner groups, in this case the FLLs and HLLs.

The C-test technique was of particular interest as it aligned with the need in the applied linguistics research for practical tools that can reliably assess general proficiency. Due to its easy administration and scoring method, it can be used as a quick measurement of language learners’ general proficiency for research purposes. As mentioned earlier,
this would in turn address the inconsistent use of indicators of proficiency that hindered the generalizability of results across studies. Furthermore, unlike other shortcut measure techniques (e.g., cloze tests, elicited imitation tests) that have been more frequently studied, and despite the results of most previous studies on C-tests supporting its use as a reliable measure of global proficiency (e.g., Babaii & Jalali Moghaddam, 2006; Eckes & Grotjahn, 2006; Klein-Braley, 1997; Raatz & Klein-Braley, 2002), there has been less scholarly attention exploring the design of such testing technique. In particular, limited research has focused on the use of C-tests for the assessment of LCTLs such as Korean. To my knowledge, Lee-Ellis (2009) is the only study that has reported on the design and implementation of the C-test for Korean language learners. Although the study provided substantial evidence for its reliability in discriminating a wide range of Korean learners’ proficiency, it did not address the possible differences in constructs being assessed based on different learner populations. Most importantly, it did not follow a systematic process of validation where different claims are stated and then backed by relevant evidence, keeping the intended use and users as the core of the validation.

Hence, Kane’s argument-based approach to validity is particularly useful for the purposes of this study because, unlike previous validity frameworks, it integrates multiple pieces of evidence into a single coherent model that can justify the test uses and score interpretations (Purpura, Brown, & Schoonen, 2015). As it follows a systematic process of analyzing claims, it also encourages the carefully examination of the evaluation questions that need to be answered first. Moreover, it provides a useful method of identifying the pieces of evidence required to answer these evaluation questions. Within this framework, the constructs of language proficiency that the test intends to assess need
to be clearly specified and evaluated, thus it directly addresses the issue of vagueness in construct definition. In the case of the present study, it supported a thorough analysis of what constructs were being assessed for each learner group, FLLs and HLLs, and whether these constructs of language proficiency were dependent on learner population. That is, it examined the extent to which the C-test behaves differently for each learner population. In addition, it facilitated the analyses of the types of language skills that the C-test can predict. On the whole, this study also sought to contribute to the list of the few validation studies (e.g., Drackert, 2016; Purpura, Brown, & Schoonen, 2015) that have applied this comprehensive framework for evaluating instruments in the SLA research, and thus set the standards for future validation studies following the same purposes.

1.3. Structure of the Dissertation

This dissertation followed a similar format of previous validation studies applying the Kane’s (2006) argument-based approach to validity (e.g., Chapelle et al., 2008, 2010; Drackert, 2016) and therefore it was guided by the interpretive argument chain from which the evaluation questions emerge. Before presenting the interpretive argument, the dissertation first begins with the description of the extant literature on the definition and operationalization of language proficiency in Chapter 2. This chapter presents a brief historical overview of the different theories and models of L2 proficiency. It also provides a glance to the definition of HL proficiency and how it has been conceptualized in the HLA research. Subsequently, Chapter 3 synthesizes the literature on C-tests and their role in assessing FL proficiency as well as HL proficiency. This chapter touches upon some of the caveats of the previous Korean C-test study by Lee-Ellis (2009) and
discusses the ways in which the new Korean C-test (Son et al., 2018) attempted to address those issues. Next, Chapter 4 reviews the literature on validity, highlighting the advantages and disadvantages of the different validity models in chronological order, leading up to the argument-based approach to validity. It contextualizes the latter model within the wider domain of validity research and points out the advantages that this model provides for the purposes of this dissertation study. After describing the relevant literature, the study follows a similar structure to Chapelle et al.’s (2008) validation studies. Thus, Chapter 5 outlines the interpretive argument chain for the Korean C-test and specifies the evaluation questions as well as the evidence needed to validate the argument. Then, Chapter 6 describes the methodology of data collection, including the process of administering four tests of Korean language proficiency, namely the Korean C-test, the ACTFL Oral Proficiency Interview-computer (OPIc), ACTFL Writing Proficiency Test (WPT), and the Korean Elicited Imitation Test (EIT), to gather data for empirical backing of the different assumptions made in the interpretive argument. Chapter 7 presents the outcomes relevant to each of the inference in the Korean C-test Interpretive argument chain. Finally, Chapter 8 offers a critical evaluation of the validity argument by examining the coherence of the argument and the appropriateness of the quantity and quality of the evidence provided. It also considers some of the ramifications of the study in terms of the future impact on scholarly work focused on Korean foreign and heritage language learning. This chapter also includes a description of the limitations of the study providing recommendations for the future research agenda in terms of how it can address the drawbacks of the present study.
Chapter 2

Language Proficiency

Defining and measuring second language proficiency has been an extremely challenging task for both foreign language educators and SLA researchers alike. From the foreign language pedagogy perspective, language educators have been presented with the difficult task of measuring learners’ progress in acquiring a foreign language, identifying their weaknesses and strengths, and making sure to meet with rigid accountability measures, all of which require a clear understanding of the components that constitute language proficiency. From the SLA research perspective, scholars have faced the challenging task of defining the construct of language proficiency at different language developmental stages and finding the right tools to measure this construct, in order to investigate what and how learners acquire language abilities. In the extant SLA literature, language proficiency has been included as an independent or moderating variable, making its definition and measurement a crucial component in the research design. This chapter describes how scholars in the language testing, and more generally, in the applied linguistics community have tried to describe second and heritage language proficiency and to implement ways to measure it.

2.1. Defining Second Language Proficiency

2.1.1. Models of second language proficiency

The definition of L2 proficiency and the articulation of its constructs has truly undergone a metamorphosis throughout time, which can be witnessed through the numerous approaches and models adopted to describe multiple skills and components of
language ability. Without a doubt, the evolution in the conceptualization of language proficiency has been largely influenced by theories of language acquisition as well as psychology (Hulstijn, 2010). As Purpura (2008) notes, there are at least two approaches for assessing language ability, namely, through theoretical models that illustrate the different components of language proficiency and through the consideration of the opinion of experts who have had extensive experience with language learners. In this section, some of the most influential models of L2 proficiency are synthesized in chronological order (see Figure 1).

Figure 1. Timeline of different models and theories of L2 proficiency

In regard to the first approach for assessing language abilities, numerous models of L2 proficiency have been developed by scholars to understand the complex theoretical underpinnings of what it means to be proficient in a language. Most of these models have been driven by the goal of ultimately being able to quantify language proficiency, either as a whole or as individual components. Historically, Lado (1961) was one of the first
scholars who proposed a comprehensive model of language proficiency that also accounted for the language skills that could be tested. In this model, language proficiency was described as being composed of four language elements: phonology, grammatical structure, lexicon, and cultural meanings. These elements, he explained, were integrated in four main skills: speaking, listening, reading, and writing. In other words, being proficient in speaking a language, for example, would involve the integration of several elements like phonology, lexicon, etc. Carroll (1968) further developed Lado’s model by clearly defining each of the four main skills and articulating how they could be measured. He advocated for a separate assessment of productive (i.e., speaking and writing) and receptive skills (listening and reading) using both discrete-point tasks (i.e., task with a single stimulus and a response) which assessed one discrete component of formal knowledge, and integrative tasks (i.e., tasks with multiple stimuli and responses) which tested the knowledge of multiple components of language. Since then, as Cumming (1996, 2008) remarks, there has been a continued prevalence of this four-skills model in language education, assessment, and policy.

By the end of the 1970s, Oller (1979) put forth a unitary model of competence that described language proficiency as consisting of one single global trait as opposed to four different skills, referred to as pragmatic expectancy grammar. In this model, two major contexts of language use could be identified, namely, linguistic context (i.e., verbal and gestural aspects of language) and extralinguistic context (i.e., objective and subjective aspects including events, people, relationships, perceptions, etc.). For Oller, language proficiency was defined in terms of the ability to synchronize these two contexts, that is, the linguistic elements, such as phonology, lexis, and morphosyntax,
with the extralinguistic contexts, or the ability to use appropriate language in a given context. He argued that this global trait could be measured through integrative tasks such as dictation, cloze tests, paraphrase recognition, and question answering, among others. He claimed that the successful completion of, for example a fill-in-the-gap test, required test-takers to integrate grammar, meaning, and pragmatic abilities. This model drew considerable criticism among many language acquisition and language testing researchers, as it was not supported by empirical research (e.g., Bachman & Palmer, 1982). Language proficiency was considered to be multicomponential and thus scholars did not agree with the idea of measuring it solely through integrative tasks, such as the ones mentioned by Oller.

Along these lines, another major criticism to Oller’s (1979) model was raised by James Cummins at the beginning of the 1980s. In response to negative views on bilingual education, Cummins (1980) advocated for a reconceptualization of bilingual proficiency. In his model, he made an important differentiation between two dimensions of language proficiency, namely the Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP). BICS in an L1 were defined as consisting of language skills that all normal individuals “manifest in everyday interpersonal communicative situations” (p. 84) regardless of whether they had received instruction or not. CALP, on the other hand, was characterized as the language abilities that learners acquire through formal instruction and thus are “strongly related to literacy and to other decontextualized verbal-academic tasks” (p. 86). He explained that Oller’s model only accounted for the latter dimension of proficiency and that it was unable to explain whether the global language proficiency trait underlay learners’ language ability
in different languages. According to Cummins, the CALP of the L1 and L2 were interdependent and “previous learning of literacy-related functions of language (in L1) [predicted] future learning of these functions (in L2)” (p. 86). Moreover, Cummins opposed the Separate Underlying Proficiency (SUP) model, which conceived bilinguals as having a brain with two separate language proficiencies, that of L1 and a completely separate system for L2. In his view, the Common Underlying Proficiency (CUP) model better explained the bilingual learner’s language system as it described CALP as underlying both L1 and L2, which can develop together if given adequate exposure. He argued that most of the misconceptions about the negative effects of bilingual education came from the SUP model and presented empirical research that directly contradicted it, calling for a redefinition of the bilingual language proficiency model.

Another criticism of Oller’s (1979) proposal was based on models that described language proficiency as being multicomponential. In the 1970s, Hymes (1971) presented a theory of language competence, which, as opposed to Chomsky’s (1965) views on linguistic competence and performance, highlighted the importance of language use and context. Hymes argued that the rules of language use were also important and that the rules of grammar would be useless without them. As a result, he proposed the notion of communicative competence, which included grammatical competence (i.e., the knowledge of grammar rules) and sociolinguistic competence (i.e., the knowledge of language use). This theory, although present from the beginning of the 1970s, became entrenched in the language testing community in the 1980s when, largely influenced by it, Canale and Swain (1980) presented a model of communicative competence. This model would later become greatly influential in communicative language teaching and
testing. Canale and Swain underscored the equal importance of both grammar knowledge and language use, stating that “[j]ust as Hymes (1972) was able to say that there are rules of grammar that would be useless without rules of language use, so we feel that there are rules of language use that would be useless without rules of grammar” (p. 5). They defined communicative competence as being multicomponential, consisting of (a) grammatical competence (i.e., knowledge of lexis and the rules of syntax, morphology, semantics, and phonology); (b) sociolinguistic competence (i.e., knowledge of sociocultural rules of language use and rules of discourse); and (c) strategic competence (i.e., knowledge of verbal and non-verbal communication strategies primarily used during communication breakdown). Nevertheless, this model was criticized for its limited explanation on how these different components were associated with each other, and for the lack of empirical evidence to support the whole model (Harley, Allen, Cummins, & Swain, 1990).

Following Canale and Swain (1980), Bachman (1990) and Bachman and Palmer (1996) proposed a model of communicative language ability (CLA) that underscored the importance of language function and use. In this model, language ability was defined largely by two categories, namely, language knowledge and strategic competence. While the latter component referred to metacognitive strategies and components, the former was further divided into organizational competence and pragmatic competence. Organizational competence consisted of language abilities related to grammatical competence (i.e., vocabulary, morphology, syntax, and phonology) and textual competence (i.e., cohesion, rhetorical organization). Pragmatic competence was composed of (a) functional competence (i.e., ability to produce and interpret appropriate
language functions), which was further divided into ideational, manipulative, heuristic, and imaginative macro-functions, and (b) sociolinguistic competence (i.e., sensitivity to the conventions of language use in a given context), which was further classified as the sensitivity to dialects and language varieties, registers, naturalness, and cultural references and figures of speech. As did previous models of language proficiency, this framework also received criticism. McNamara (1990), for example, problematized the degree of importance that could be assigned to each of the components, particularly for performance tests which are scored by raters. He explained that the scoring could be biased towards assigning more importance to a particular component (e.g., grammatical competence) over the other (e.g., pragmatic competence).

In the early 2000s, Purpura (2004) extended Bachman and Palmer’s (1996) model of language proficiency to include a distinction between grammatical form and meaning. Purpura claimed that the CLA model failed to articulate the difference between grammatical form and meaning. To illustrate this difference, he explained that a learner might have knowledge of grammatical forms (e.g., simple past tense, present perfect) but at the same time have little understanding of the difference in the notions of these forms (e.g., setting the scene, current relevance). Moreover, he pointed out that the CLA model did not expound on the way grammatical competence is used as a resource for conveying literal, intended, or implied meaning. Accordingly, he proposed a definition of language knowledge that differentiated grammatical form, grammatical meaning, and pragmatic meaning. He described knowledge of grammatical forms as the knowledge of forms on the phonological, lexical, morphosyntactic, cohesive, information management, and interactional levels. Knowledge of semantic meaning, on the other hand, referred to the
knowledge of the way different parts of an utterance are arranged to convey a literal meaning and the way in which the speaker uses these parts to convey an intended meaning (p. 63). Finally, knowledge of pragmatic meaning involved the ability to understand the implied meaning encoded in an utterance, that is, the meaning that is “beyond what is explicitly expressed by the grammatical forms and their intended meanings” (Pupura, 2008, p. 61).

Like Purpura (2004), numerous other researchers have provided multiple perspectives of and alternative components to the CLA model to help better explain the constructs of language proficiency (e.g., Hamp-Lyons & Henning, 1991; Hinkel, 2002; Sasaki, 2002). These different perspectives seem to agree on the multicomponential characteristic of language proficiency and on the way learners appear to develop abilities in each component separately. However, as Purpura (2008) notes, they have not reached a consensus on which components should be included in the model, how they are interrelated, or how they are acquired and organized in a learners’ mind. Additionally, Cumming (2008) has argued that these models tend to simplify the notions of literacy and oracy modes of communication. Generally, literacy skills have been assessed through the reading and writing, whereas oracy skills have been assessed through speaking and listening. These conceptualizations have neglected the assessment of multiliteracies (e.g., visual or media literacies) and the “notions of hybridity, contextualization, and multimodality” (p. 6).

Another caveat to these models is their inability to account for the different dimensions of language proficiency and for the relationship between L1 and L2 proficiency. Recently, Hulstijn (2011, 2015) presented a more comprehensive framework
of language proficiency that addressed both native speakers’ and learners’ language proficiencies, and expounded on cognitive competences. More concretely, Hulstijn (2011, p. 242) defined language proficiency as:

“the extent to which an individual possesses the linguistic cognition necessary to function in a given communicative situation, in a given modality (listening, speaking, reading, or writing). Linguistic cognition is the combination of the representation of linguistic information (knowledge of form-meaning mappings) and the ease with which linguistic information can be processed (skill). Form-meaning mappings pertain to both the literal and pragmatic meanings of forms (in decontextualized and socially-situated language use, respectively).”

Together with this more comprehensive conceptualization of language proficiency, Hulstijn (2011, 2015) provided a model of language proficiency that, based on research data, included two separate dimensions, namely, a basic vs. higher level cognitive dimension and a core vs. periphery dimension. The first dimension is similar to Cummins’ (1980) BICS and CALP model in that it separates two levels of cognition: basic level of cognition (BLC) and higher (or extended) level of cognition (HLC).

Hulstijn explained that the BLC refers to language skills that all normal individuals have in common. It involves (a) the implicit knowledge of phonology, morphology, and syntax; (b) the explicit knowledge of lexis; and (c) the automaticity of processing both types of knowledge. BLC is also described as being limited to speech as opposed to writing. The HLC is characterized as being an extension to the BLC and thus identical to it, with the exception that it pertains to the perception and production of utterances with low-frequency lexical items and morphosyntactic structures. Furthermore, in contrast to the BLC, the HLC is not limited to speech but also includes written language. This
distinction between the two levels of cognition has not been addressed by previous models and it appears to be crucial as it claims that L2 learners’ attainability of each level of cognition is separate. Hulstijn (2011) asserted that “while L2 learners can acquire HLC in their L2 as native speakers can (depending on their intellectual skills, education, professional careers and leisure-time activities), it remains an open question to what extent postpuberty L2 learners can fully acquire BLC in their L2” (p. 242). He also proposed two corollaries related to this idea: (1) while late L2 learners can attain BLC in areas of vocabulary and most grammar, it will generally be more difficult to do so in areas of phonology; and (2) late L2 learners can become as proficient in the HLC as L1 speakers “of the same intellectual, educational, professional and cultural profile, despite some deficiencies in their L2 BLC” (Hulstijn, 2015, p. 48).

The second dimension proposed by this model makes another important distinction, this time between what Hulstijn referred to as core and peripheral proficiencies. The core component of proficiency consists of linguistic knowledge of appropriately using linguistic forms (i.e., phonological, morphophonological, morphosyntactic, lexical and pragmatic forms) in a given communicative context, as well as the speed to process these linguistic forms. In contrast, the peripheral component includes less- or non-linguistic abilities, such as interactional abilities, strategic competence, metalinguistic knowledge, and discourse competence. Moreover, he indicated that both dimensions of language cognition (i.e., BLC vs. HLC and core vs. periphery) should be considered orthogonally, illustrating the relationship between the two dimensions conceptualized in both L1 and L2 development, in a bottom-up model, as in Figure 2. As observed through his representation of the BLC-HLC theory, Hulstijn
associated HLC with peripheral components and some core components separate from BLC. Meanwhile, the BLC was described as overlapping only with the core components of language abilities.

Figure 2. Hulstijn’s (2015) representation of the BLC-HLC theory and its components.

Hulstijn (2015, p. 46)

In sum, this section has provided a brief overview of some of the most influential perspectives and definitions of language proficiency. Overall, language proficiency, which was first defined in terms of four skills or one global trait, has developed over time to include multiple components that also recognize the importance of language use and contexts, as well as language abilities beyond those specific to the knowledge of linguistic forms. Most recently a model of language proficiency has included a more complex framework that includes several dimensions of language cognition,
differentiating levels of language ability that are acquired implicitly and those acquired only through formal education. Although these definitions have become increasingly more comprehensive, these theoretical models have not always been applied in the real-life educational contexts. On the one hand, the language acquisition theorists have mainly focused on providing plausible explanations of how learners acquire certain components of a language, and on the other hand, language educators have primarily been interested on the methods to make proficiency assessable. These two approaches to conceptualizing proficiency have often times been working separately, hindering the further understanding of how to operationalize language proficiency constructs and how to measure them.

2.1.2. Defining second language proficiency through expert opinions

Eliciting expert opinions constitutes the second major approach to defining and assessing language proficiency. This approach, as Purpura (2008) argues, has generally not been based on empirical research but rather on the accumulated experience of language educators who have been in direct contact with language learners and who have been attuned to their pedagogical needs. Moreover, this approach has been concerned more directly with addressing practical issues in language education like setting standards that can be used as the basis of curricula, syllabuses, assessments (i.e., formative or summative), textbooks, and that can provide language programs with evidence of accountability. These expert views have been the backbone of several language proficiency standards and guidelines, such as the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines (ACTFL, 2012a) and the Common European Framework of Reference (Council of Europe, 2001), both of which have had a
major impact on foreign language education, testing, and policy in the US, Europe, and beyond. Most importantly, these types of guidelines have provided benchmarks that define what learners are expected to know after a certain period of FL instruction.

The origin of the ACTFL proficiency guidelines can be traced back to the Interlanguage Roundtable (ILR) scale, which was developed in the 1950s by the Foreign Service Institute (FSI) in order to teach and assess FLs for different government agencies. In 1981, the FSI and the Educational Testing Service (ETS) collaborated in a project sponsored by the government to adapt the scales for academic use in schools and colleges, becoming what we know as the ACTFL proficiency guidelines (Liskin-Gasparro, 2003; Malone, 2003). The implementation of these proficiency guidelines in academic settings was motivated by efforts to address problems with the then prevalent foreign language requirement in higher-education or the “seat-time” standard (i.e., language programs required learners to enroll and pass two-years of foreign language courses). Norris and Pfeiffer (2003) pointed out that one of the main issues with this standard has been that FLLs were not achieving adequate levels of language skills by the end of the two years (i.e., the standard time to fulfill the FL requirement). Liskin-Gasparro (2003) also noted that before the use of ACTFL guidelines, there was a disconnect between results of empirical research elicited in a controlled setting and what language educators were experiencing first-hand. Consequently, as these guidelines seemingly addressed these issues, they rapidly gained popularity among the language education community.

The most recent version of the ACTFL proficiency guidelines defines language proficiency in terms of five main levels, namely, novice, intermediate, advanced,
superior, and distinguished, over four modalities, that is, speaking, listening, reading, and writing. Each major level is further subdivided into a low, mid, and high levels. This framework of proficiency represents a language development continuum that ranges from having language ability comparable to a “highly articulate, well-educated language user to a level of little or no functional ability” (ACTFL, 2012a, p. 3). In addition, learner performances are assessed on the basis of three main criteria: content, function, and accuracy. The National Council of State Supervisors (NCSSFL) together with ACTFL provides a global description of learners’ language skills within each level of proficiency through holistic can-do statements in interpersonal communication, presentational speaking, presentational writing, interpretive listening, and interpretive reading. Table 1 presents the global can-do statements for interpersonal communication across all six levels and nine sublevels of proficiency (for the complete can-do statement global scale see NCSSFL-ACTFL Can-Do Global Can-Do Benchmarks, n.d.).
<table>
<thead>
<tr>
<th>Benchmarks</th>
<th>Can-Do Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Novice</strong></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>I can communicate on some very familiar topics using single words and phrases that I have practiced and memorized.</td>
</tr>
<tr>
<td>Mid</td>
<td>I can communicate on very familiar topics using a variety of words and phrases that I have practiced and memorized.</td>
</tr>
<tr>
<td>High</td>
<td>I can communicate and exchange information about familiar topics using phrases and simple sentences, sometimes supported by memorized language. I can usually handle short social interactions in everyday situations by asking and answering simple questions.</td>
</tr>
<tr>
<td><strong>Intermediate</strong></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>I can participate in conversations on a number of familiar topics using simple sentences. I can handle short social interactions in everyday situations by asking and answering simple questions.</td>
</tr>
<tr>
<td>Mid</td>
<td>I can participate in conversations on familiar topics using sentences and series of sentences. I can handle short social interactions in everyday situations by asking and answering a variety of questions. I can usually say what I want to say about myself and my everyday life.</td>
</tr>
<tr>
<td>High</td>
<td>I can participate with ease and confidence in conversations on familiar topics. I can usually talk about events and experiences in various time frames. I can usually describe people, places, and things. I can handle social interactions in everyday situations, sometimes even when there is an unexpected complication.</td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>I can participate in conversations about familiar topics that go beyond my everyday life. I can talk in an organized way and with some detail about events and experiences in various time frames. I can describe people, places, and things in an organized way and with some detail. I can handle a familiar situation with an unexpected complication.</td>
</tr>
<tr>
<td>Mid</td>
<td>I can express myself fully not only on familiar topics but also on some concrete social, academic, and professional topics. I can talk in detail and in an organized way about events and experiences in various time frames. I can confidently handle routine situations with an unexpected complication. I can share my point of view in discussions on some complex issues.</td>
</tr>
<tr>
<td>High</td>
<td>I can express myself freely and spontaneously, and for the most part accurately, on concrete topics and on most complex issues. I can usually support my opinion and develop hypotheses on topics of particular interest or personal expertise.</td>
</tr>
<tr>
<td><strong>Superior</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I can communicate with ease, accuracy, and fluency. I can participate fully and effectively in discussions on a variety of topics in formal and informal settings. I can discuss at length complex issues by structuring arguments and developing hypotheses.</td>
</tr>
<tr>
<td><strong>Distinguished</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I can communicate reflectively on a wide range of global issues and highly abstract concepts in a culturally sophisticated manner.</td>
</tr>
</tbody>
</table>
As represented throughout Table 1, the distinction between proficiency levels is dependent on (a) the range of topics, that is, whether it is personal or interpersonal, familiar or unfamiliar, formal or informal, abstract or concrete; (b) complexity of language produced, that is, whether it is short or extended, complex or high-frequency structures/words; (c) fluency of the performance; and (d) accuracy of the performance.

Numerous scholars have pointed out multiple caveats to the use of these guidelines to assess FL proficiency, including setting up a standard on the basis of a ‘native speaker’ (e.g., Barnwell, 1989; Lantolf & Frawley, 1985; Valdés, 1989), limited applicability to all FL programs from different institutions (Norris & Pfiffer, 2003), and lack of empirical evidence to substantiate their rationale (e.g., Dandonoli & Henning, 1990; Fulcher, 1996; Valdés, Haro, & Echevarriarza, 1992), among others. Nevertheless, due to its ready-availability, practicality, and applicability to FL pedagogy, the Guidelines have been highly pervasive in FL instruction in higher education in the US. In addition, more scholarly work has increasingly been oriented towards investigating what these guidelines actually represent in terms of discrete linguistic features. The Linguistic Correlates of Proficiency (LCP) project, for instance, is a government-funded initiative that has set forth to examine what linguistic features can be identified in each of the levels on the ILR scale. The goal of this project has been “to identify linguistic features implicated in progress on the ILR scale—specifically, from ILR 2 to 2+, and from 2+ to 3 —and the project ultimately aimed to improve the teaching, learning, and testing of high-level listening and speaking abilities” (Long, Gor, & Jackson, 2012, p. 108). This research studies have investigated multiple languages such as Russian (Long et al., 2012), Korean (Lee, Moon, & Long, 2009; see section 2.4.2.1), Persian, and Chinese (Pelz et al.,
Overall, these studies have found that control of most of the linguistic features in phonology, morphology, syntax, lexis, and collocations distinguishes between ILR proficiency levels 2 and 3. However, the control of only a subset of linguistic features was able to distinguish between levels 2+ and 3, as well as 2 to 2+. In other words, the ILR scale was able to distinguish different levels of proficiency according to the control of linguistic features. However, specific differences in linguistic ability were not as evident for the some of the sublevels.

Another global proficiency scale widely used in FL education has been the CEFR. The use of these standards has greatly impacted policy and education of FLs primarily in Europe (see Byram & Lynne, 2012). In fact, the intended uses articulated by the Council of Europe (2001) include providing “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc” (p.1), planning of language learning programs, language certification, and self-directed learning (p. 6). Similar to the ACTFL guidelines, the CEFR defines language proficiency on the basis of six different levels of proficiency and describes the language abilities in each level according to can-do statements. North (2000, 2007, 2014) explained that the scale and descriptors were originally based on the first seven levels that David Wilkins presented in the 1977 Ludwighaven Symposium. In 1993, the Swiss research project was developed with the objective of calibrating and finding empirical support for the use of these scales and descriptors. The qualitative phase of the study was conducted first to analyze the clearness and viability of an initial pool of about 2000 descriptors, some from existing language assessment scales and others created for the study. This research phase was conducted through a series of 32 workshops with teachers in the course of two years. In
these workshops, teachers were asked to offer their expert opinions about the descriptors and also sort them for the same category into different levels. The best descriptors were selected for the next quantitative phase of the study, where teachers were asked to respond to a 50-item questionnaire about the descriptors. They were required to assess the difficulty of using such descriptors in order to calibrate the scales. In addition, learners were asked to try out the descriptors and no differences in the interpretation of the descriptors were found between teachers and learners.

As Hulstijn (2011, 2014) and Little (2006) have pointed out, the CEFR can be described in terms of two dimensions, namely, a horizontal dimension consisting of language activities and the quantity of skills of functions, topics, contexts that learners can control for communication, and a vertical dimension composed of ascending levels of proficiency, which are based on the quality with which the communication is effective. Table 2 presents the global scale of the Common Reference levels with their respective descriptors.
Table 2

**CEFR Global Scale.** *(European Council, 2001, p. 24)*

<table>
<thead>
<tr>
<th>Levels</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic User</strong></td>
<td></td>
</tr>
<tr>
<td><strong>A1 – Breakthrough</strong></td>
<td>Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.</td>
</tr>
<tr>
<td><strong>A2 – Waystage</strong></td>
<td>Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.</td>
</tr>
<tr>
<td><strong>Independent User</strong></td>
<td></td>
</tr>
<tr>
<td><strong>B1 – Threshold</strong></td>
<td>Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.</td>
</tr>
<tr>
<td><strong>B2 – Vantage</strong></td>
<td>Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.</td>
</tr>
<tr>
<td><strong>Proficient Speaker</strong></td>
<td></td>
</tr>
<tr>
<td><strong>C1 – Effective Operational Proficiency</strong></td>
<td>Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.</td>
</tr>
<tr>
<td><strong>C2 – Mastery</strong></td>
<td>Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.</td>
</tr>
</tbody>
</table>
As observed in Table 2, the progression of can-do statements throughout the CEFR levels share numerous characteristics with the ACTFL proficiency guidelines, particularly in terms of their reference to the difference in the ability to produce language on a range of topics, as well as the complexity, accuracy and fluency of the language produced. However, the CEFR additionally makes specific references to the variety of tasks that learners can perform (e.g., introduce oneself, describe events, hopes and dreams, or summarize information).

Overall, the CEFR has majorly influenced the FL education in Europe and increasingly around the world (Byram & Parmenter, 2012). Not surprisingly its impact on language instruction and policy has stimulated several productive discussions, as the one observed in the 2007 issue of Perspectives in the Modern Language Journal and the General Language Proficiency Symposium at the Language Testing Forum (LTF), which have mainly focused on identifying the advantages and disadvantages of the CEFR (see Harsch, 2014 for a full review of the LTF discussions). On the one hand, the CEFR has been praised for its emphasis on plurilingualism, which as Byrnes (2007) has pointed out, is different from multilingualism in that it values the communication in “any of the languages available to a particular user—not only major languages, even less only English as a lingua franca” (p. 642) recognizing the importance of a plurilingual environment. On the other hand, the CEFR has also received strong criticism for its lack of basis in SLA theory and in empirical research (Alderson, 2007; Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu, 2004; Cumming, 2009; Fulcher, 2004; Hulstijn, Schoonen, de Jong, Steinel, & Florijn, 2012; Spolsky, 2008). In two position papers, Hulstijn (2007, 2014) articulated several limitations of the CEFR by critically analyzing
the structure and description of this framework. One of the biggest concerns was the limited empirical research that supports the progression of the levels in the CEFR. Hulstijn (2007, 2011, 2014) described the CEFR as having a static component, which pertains to language proficiency, and a dynamic one, which refers to language development. However, he noted that the framework lacks specificity in terms of whether all levels can be developmental. For example, it is still unclear whether the successful performance at a certain level (e.g., B1) is possible with the linguistic abilities typically present in lower levels (e.g., A2 or A1). Furthermore, he problematized the wording of the descriptors of the CEFR, particularly for higher levels (i.e., B2, C1, C2), as they imply the requirement of intellectual skills in addition to language proficiency for successful performance. He mentioned, for example, the descriptors for level B2: “Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail.” As observed through this example, the higher levels in the CEFR appear to make no differentiation between language proficiency and intellectual skills. Accordingly, Hulstijn (2007, 2014) concluded both of his position papers by calling for more empirical research that can address all the aforementioned issues and provide backing for the claims the CEFR makes.

To summarize, this section has provided an overview of two of the most commonly used scales of language proficiency worldwide, namely the ACTFL proficiency guidelines and the CEFR. Unlike the proficiency models described in the previous section, the conceptualization of L2 proficiency within these frameworks is more oriented towards practical issues in FL education. These scales have served as
benchmarks that define language proficiency within a developmental continuum. Therefore, they have been useful tools for informing language educators in the design of curricula, syllabuses, and instructional material, for standardizing the FL education across different language programs, and for keeping these programs accountable for improving learners’ language competence. As in the case of L2 proficiency models, however, these scales have also been criticized for lacking empirical support for their rationale. As a result, numerous scholars have recommended a research agenda that engages in more rigorous operationalization of constructs and that does not depend solely on impressionistic perspectives of experienced language educators. One additional consideration that has been often forgotten in the list of recommendations and that should be regarded as crucial, is the interpretation of the intended uses of these standards. As Norris (2008, 2016) has indicated, the question of whether these are effective tools to define and assess language proficiency should be answered on the basis of their intended uses.

2.2. Measuring Second Language Proficiency in Second Language Acquisition Research

Although the models and methods to assess L2 proficiency mentioned earlier have had a great influence in FL education and policy, they have not always been incorporated into SLA research methodology. As Hulstijn (2012), and before him Norris and Ortega (2003) have noted, the definition and assessment of language proficiency has been taken for granted in the majority of SLA research, even though language proficiency has played the role of moderating or independent variables. The following section
provides a review of studies that have synthesized the methods in which language proficiency has been assessed within studies in SLA or bilingualism.

In two highly influential syntheses, Margaret Thomas provided a summary of the different methodologies used to measure L2 proficiency in SLA research. In Thomas (1994), she reviewed a total of 157 SLA studies from 1988 to 1992 appearing in four renowned journals in the field, namely, *applied linguistics (AL)*, *Language Learning (LL)*, *Second Language Research (SLR)*, and *Studies in Second Language Acquisition (SSLA)*. The results of this synthesis indicated that there were primarily four ways in which language proficiency was assessed: (a) impressionistic judgment (i.e., researchers’ or other people’s evaluation of the participants’ L2 proficiency, unsupported by empirical data); (b) institutional status (i.e., information about learners’ instructional level); (c) research-internal or in-house measures of proficiency (i.e., locally-designed test of language proficiency); and (d) standardized test scores (i.e., self-reported scores of tests such as TOEFL or the Michigan Test of English Language Proficiency). The most popular method was by far institutional status followed by scores of standardized tests and impressionistic data, and the least popular method was the development of in-house tests. Thomas underscored the challenges that these four methods for assessing language proficiency pose on the comparability of results across studies and thus ultimately also on the generalizability of findings in the SLA field.

More than one decade after her first synthesis, Thomas (2006) reexamined the trends in the measurement of L2 proficiency in SLA research by reviewing 211 articles from 2000-2004 in the same academic journals. The results of the new synthesis revealed that not much had changed over one decade. She found a minimal decrease of 7% in the
use of institutional status and a 5% increase in the in-house assessment method. This new synthesis, however, indicated that there was a positive change in the reporting of more detailed data concerning proficiency and statistical analyses that estimated the validity of participant groupings. In addition, it also revealed a tendency to use more than one method of proficiency assessment. When comparing both syntheses, Thomas concluded that although there appeared to be a movement towards reporting more information about language proficiency, within the strongly experimental subdomain of SLA/UG research, there was also a trend to devalue the assessment of L2 proficiency, regarding it as “unnecessary, unreliable, or unrevealing” (p. 294).

In another synthesis study, Tremblay (2011) found similar trends of L2 proficiency measurement in SLA research. After surveying 144 studies published from 2000-2008 in three academic journals, SLR, SSLA and French Language Studies (FLS), she found the same trends as those described in Thomas (1994). Although Tremblay reported more than one third of studies measuring proficiency through independent tests, which was more than reported by Thomas (1994), the most popular method was again instructional level, followed by self-reported scores of standardized proficiency tests (mostly from TOEFL), and the length of residence in an L2-speaking country or region, which was what Thomas described as impressionistic data. Most importantly, like Thomas, Tremblay also emphasized the importance of designing reliable/valid tools that can make language proficiency comparable across studies. In addition, she specifically pointed out that more research on independent tests of language assessment could help improve the generalizability of results across studies.
In a more narrowly-focused review study, Hulstijn (2012) explored the trends of language proficiency measurement solely for published research where the language proficiency was assessed as an independent or moderating factor, and thus was a necessary requirement for the study. In this review, all the 14 volumes of the academic journal *Bilingualism: Language and Cognition* from the years 1998-2011 were analyzed to identify relevant empirical studies, and a total of 140 studies were determined to fit the criteria. The results of the synthesis showed similar trends to those found in the aforementioned syntheses by Thomas and Tremblay. First, Hulsjitn (2012) also indicated an increase in the interest of assessing proficiency throughout time. While only 19% \((N = 37)\) of studies during 1998-2004 measured language proficiency, 54% \((N = 103)\) of studies did so during 2005-2011. Second, like Thomas (2006) and Tremblay (2011), this synthesis also indicated an increased tendency to use more than one method to assess proficiency for screening participants (e.g., language history, self-assessment, and cloze tests). Among these screening methods, the most common one was objective tests, such as standardized tests certified by testing institutes (e.g., TOEFL, IELTS) or in-house tests (e.g., translation test, cloze test or c-test), and self-assessment tests (e.g., Bachman & Palmer’s 1989 self-assessment). In terms of statistical analysis, ANOVAs were the most frequently used statistical tool to find differences between groups, although the within group comparisons were not as commonly explored. Hulstijn concluded his synthesis by presenting some pressing issues in research on SLA and bilingualism, which included: (a) incomplete reporting (i.e., when significant results are found, variance is not reported); (b) circular reasoning (i.e., measuring the structure under investigation as part of the construct of language proficiency); (c) comparison of language proficiency across
different languages (i.e., different languages have different grammatical and lexical systems challenging their comparability); (d) disregard for the differences between BLC and HLC when assessing language proficiency; and (e) limited empirical research that defines levels of proficiency (see also Hulstijn, 2010).

As observed through these comprehensive syntheses, although language proficiency has played a major role in SLA and bilingualism research, only limited attention has been given to the definition of its constructs, the validation of the interpretations of these constructs, and the design of reliable tools for measuring them. Norris and Ortega (2003, 2012) pointed out that depending on the different epistemologies within the SLA field, the definitions and thus the operationalization of language proficiency vary considerably. Benchmarks, for example, differ according to the subdomain of SLA. In formal linguistics, learners’ performance is compared to an L1-speaker-baseline, whereas in research where L2 learners are viewed as emerging bilinguals, the point of comparison is no longer an L1 speaker but rather a balanced bilingual. Moreover, for socially-minded SLA, the benchmark is defined as “the successful enculturation of the L2 practices without oppressive assimilation” (Norris & Ortega, 2012, p. 578). Nevertheless, these differences have often times not been taken into consideration when selecting or designing tests of language proficiency for research purposes. It is not uncommon to see researchers using instruments designed for educational purposes, for example, to assess proficiency for different research purposes, without prior evaluation of the implications that this decision might have on the results. In this regard, Hulstijn (2010) has advised that “for every individual SLA study, given its (1) goal, (2) research questions, and (3) theoretical embedding, the researcher has to
decide which construct of [language proficiency] LP, or which LP component(s) should feature as a variable (variables) and how it (they) should be measured” (p. 187), and that only after all these factors have been examined, can the decision for an appropriate test of proficiency be made.

Furthermore, while there has been a positive trend towards using and reporting one or several indicators of proficiency in a single study, one of the persistent points of concern has been the inconsistency of the methods for assessing L2 proficiency across studies, which has hindered the systematic accumulation of knowledge in the field (Norris & Ortega, 2003, 2012; Thomas, 1994, Tremblay, 2011). Accordingly, one additional consideration to be made, besides addressing the goals, research questions, and theoretical underpinnings, is that the tools to assess language proficiency should also make possible the comparability of empirical results across studies exploring similar research inquiries.

2.3. Defining Heritage Language Learners (HLLs)

For several decades now, HLLs have been a major concern in FL education and research. Some scholars have focused on identifying numerous differences between HLLs and traditional FLLs, in language exposure, learning development continuum, and sociocultural motivations, among other characteristics, and have called for research that investigates their unique characteristics as learners and that addresses their pedagogical needs. Before moving on to discussing the definitions of HL proficiency, it is important to describe who these learners are and what characteristics differentiate them from the
traditional FLLs. This section summarizes the multiple definitions of HLLs in the HLL literature and the different methods that have been used to identify them.

As Malone et al. (2014) among other scholars have noted, one of the main difficulties when studying HLLs has been defining who these learners are. There has been little consensus on what characteristics define a HLL. Due to differences in the amount and quality of experience that HLLs have with their HL, researchers have found the task of classifying these learners into a single group, while setting them apart from the FLL group, a very challenging one. For example, in terms of HL proficiency, HLLs may range from having little or no proficiency to high proficiency levels at the beginning of their language instruction (Polinsky & Kagan, 2007). Furthermore, HLLs may also differ in the extent to which they identify themselves with the heritage culture as well as the quantity and quality of interaction with HL-speaking relatives. At the same time, their motivations to learn and/or maintain the HL can be widely different (E. J. Kim, 2006; J. S. Lee, 2002; Shibata & Koshiyama, 2001). Kondo-Brown, moreover, has problematized the notion of a simple dichotomy of HLLs vs. FLLs, explaining that within the HLL group there are finer-grained differences in language skills and language use that need to be taken into consideration when trying to describe HLLs. In her 2005 study on Japanese HLLs, she found that there were at least two markedly different groups of HLLs, namely what she referred to as (a) “HLL identity group,” which was composed of HLLs who have ancestral ties to Japan and those who have at least one Japanese-speaking grandparent, and (b) “competent HLL group,” which was composed of HLLs with Japanese-speaking parents. Although both groups of HLLs shared a cultural connection to the HL, the first group was more similar in language skills (e.g., grammar, listening
and reading comprehension) and language use (e.g., speaking Japanese every day, watching TV) to the FLLs than to the second group.

An added complexity to the systematic accumulation of knowledge within this domain of research has also been the wide range of terms used around the world to refer to *heritage language*. Kagan and Dillon (2012) pointed out that in Europe, HLs have been sometimes referred to as minority or regional languages, which in the United States may not be considered HLs. Moreover, in Australia and in the United Kingdom the term *community language* has been preferred over *heritage language* (see Lo Bianco, 2008). The use of these different terms reflects the differing foci of the research in this domain. While in Europe, particularly in the Scandinavian countries, the focus has been on maintaining the first language, in Australia scholarly work has underscored the importance of investigating the family and community together with their role in HL maintenance (p. 491). In addition, the term *heritage language learner* has also been used interchangeably with many others such as native speaker, bilingual, home background, or incomplete acquirer. Draper and Hicks (2000) argued that although these terms have often been used to refer to the same type of learner population, they can have different connotations.

Without a doubt, Valdés’ (2001) definitions of HLL have been the most used and cited within the HLA literature in the United States. These definitions are based on two different perspectives. From the perspective of endangered languages and HL maintenance, HLLs are defined more broadly as “individuals [who] have a historical and personal connection to the language” (p. 38). From the perspective of language educators, HLLs are characterized as “student[s] who [are] raised in a home where a non-English
language is spoken, who speak or at least understand the language, and who [are] to some degree bilingual in that language and in English” (p. 38). Therefore, the first perspective has focused on the personal connection to the HL, where familial or ancestral roles are particularly important (Fishman, 2001; Van Deusen-Scholl, 2003), while the second one is more oriented towards proficiency in the HL. Along the same lines, Wiley (2001) made a differentiation between the interpretations of the term HLL from three contexts: the educational program, the community, and language use. The definitions from the first and second contexts align with Valdés’ definition from the educational and HL maintenance perspectives, respectively. Wiley, however, included an additional context, language use, to emphasize the importance of considering different language varieties and their impact on HL learning.

In addition, Hornberger and Wang (2008), underscored HLLs’ agency in defining themselves as HLLs. They asserted that HLLs are individuals “who exert their agency in determining if they are heritage language learners of that language” (p. 6). Thus, in their definition, in addition to their familial or ancestral ties, learners should identify themselves as such HLLs. Moreover, they made a differentiation between the terms heritage speaker and heritage language learners, indicating that not all HLLs are heritage speakers. Similarly, Polinsky and Kagan (2007) separated HLLs into two categories, (a) those who identify themselves strongly with their heritage culture but are not able to speak in or understand their HL (i.e., broad definition) and (b) those who have some skills in their HL because they have been exposed to it during their childhood but who have not yet achieved mastery in that language (i.e., narrow definition).
Furthermore, Kagan (2005) and Kagan and Dillon (2004, 2008) proposed a more fine-grained grouping of HLLs according to their experience with formal instruction in the heritage country. For instance, they explained that Russian HLLs can be classified into four different groups, namely, (1) those who completed high school in a Russian-speaking country; (2) those who attended middle school in a Russian-speaking country; (3) those who attended elementary school in a Russian-speaking country; and (4) those who were born in the US (p. 101). This classification is more in line with the second approach of defining and assessing language proficiency (see section 2.1.1 above) as it was based on data from experienced instructors and scholars as well as from HLLs’ OPI performances and self-perceived proficiencies. Results from biographical questionnaires showed that HLLs perceived their competence in completing high-proficiency level tasks according to these four groups. Overall, almost all of the HLLs from group 1 and about half from group 2 perceive their proficiencies in Russian to be good enough to complete high-level tasks (e.g., write term papers) while only a quarter of HLLs from group 3 and very few HLLs from group 4 reported being able to complete them.

Along similar lines, in a more comprehensive survey that included the responses of 1,732 students representing a total of 22 languages, Carreira and Kagan (2011) also found that HLLs’ proficiency and language use is closely related to their age of arrival to the host country, in this case the United States. More specifically, they found that those HLLs that arrived at a later age in the United States tended to use the HL with more frequency in a wide range of contexts and interactions than those who were born in the United States or arrived at an earlier age. In general, however, they reported that the vast majority of the HLLs (i.e., 70.2% of the respondents) had used exclusively the HL at
home until the age of five. In turn, this helps explain their perception of speaking and listening skills as being considerably stronger than their reading and writing skills.

The present dissertation, focusing specifically on HLLs of Korean, followed an educational perspective towards defining HLLs. Thus, this HLL definition was in line with Valdés’ (2001) definition from the educational perspective and it also takes into consideration Kagan and Polinsky’s (2007) narrow definition as well as Kagan’s (2005) description of HLLs in group 3 and 4. Accordingly, HLLs were defined as individuals who have been exposed to the HL at home during their childhood and who have, to a certain extent some capacity in speaking or comprehending the language but who may or may not be fully proficient speakers of the HL because English became their dominant language. It was also important for these learners to be able to identify themselves as heritage learners (at least as ethnically Korean) in keeping with Hornberger and Wang’s (2008) interpretation. When describing the current state of Korean HLLs in the US, Shin and Lee (2013) indicated that the number of non-traditional HLLs has increased, including international adoptees, children from international marriages, or third- and fourth-generation Korean American children whose parents have distanced themselves from the heritage culture and lost their ability to communicate in the HL. These non-traditional KHLLs have usually been described as learners who have not been exposed to the HL at home. For the purposes of this study, however, only HLLs who have been or were exposed to the HL at home either by parents or other family members and who identified themselves as belonging to this group were considered Korean HLLs.

2.3.1. Methods for identifying heritage language learners

Another challenge that researchers have faced when trying to define HLLs, has been the design or selection of tools or methods that can accurately and reliably elicit the
information needed to classify learners as HLLs or FLLs. Studies that recruit participants from a language program composed of a separate track for HLLs and FLLs, have often used already categorized learners to group them into different learner types. This is similar to the “institutional status” strategy mentioned above only applied to HLL-identification rather than to proficiency. Overall, these studies have not provided much detail on how these learners were initially classified into these different groups. Furthermore, there have also been studies that use interviews to elicit information about learners’ linguistic backgrounds. However, due to time constraints in collecting and analyzing such extensive interview data, only a few studies (e.g., Beaudrie & Ducar, 2005; Oh & Au, 2008) have employed this methodology for the sole purpose of identifying HLLs.

Perhaps, then, the most common instrument used in the HLA research to identify HLLs has been background questionnaires. These questionnaires usually consist of items that elicit learners’ linguistic background information. The most common questions can be classified into two categories: (a) place of birth and age of arrival to the host country, which are often open-ended questions directed to both learners and their parents; and (b) language contact and/or use, which includes questions about learners’ use of the target language with different family members as well as within multiple contexts (e.g., school, watching TV, listening to the radio).

The number and type of items included in the questionnaires have ultimately depended on the purpose of the study for which they are being used. On the one hand, questionnaires that are used as an initial stage to separate HLLs and FLLs for purposes of placement testing, have tended to be shorter and composed of discrete items such as
yes/no statements or Likert-scale items (e.g., Beaudrie & Ducar, 2012; Burgo, 2013; Fairclough, Belpoliti, & Bermejo, 2010). Beaudrie and Ducar (2012), for example, used 14 yes/no statements, four of which were distractors and 10 of which explored learners’ childhood and current target language use/contact. In this particular case, if learners answered affirmatively to three or more statements, they were classified as HLLs. On the other hand, for research oriented towards investigating the factors that predict HL maintenance, the questionnaires have been characterized as being lengthier and apart from the three most common types of questions mentioned above, they have also included items examining HLL’s ethnic identity as well as motivation and attitudes towards the heritage culture and language (see Alarcón, 2010; Jensen & Llosa, 2007; E. J. Kim, 2006; Kondo-Brown, 2005; J. S. Lee, 2002). In keeping with the definition of HLL used for the purposes of this dissertation, a questionnaire, adapted from empirically tested questionnaires (e.g., E.J. Kim, 2006; Marian, Blumenfeld, & Kaushanskaya, 2007; Montrul, 2012; Torres, 2012), will be designed and will focus mainly on HLLs’ HL exposure, use, and contact during their childhood and current daily life. Detailed information about this questionnaire is provided in Chapter 6.

2.4. Defining Heritage Language Proficiency

Similar to the definition of L2 proficiency, there has been limited understanding of what constitutes the construct of HL proficiency. It is unclear what concrete aspects of HLLs’ language proficiency are unique to this greatly heterogeneous learner group and what characteristics set them apart from FLLs and native speakers (NSs). More scholarly work, however, has increasingly been oriented towards a coherent and holistic definition
of HL proficiency. As a first step, some scholars have proposed different models to describe HL proficiency, situating HLLs within a bilingual continuum. Valdés (2001, 2005), for example, provided a representation of HLLs’ language proficiency based on their strengths in their L1 and L2 (see Figure 3). In this model, HLLs were mapped out in a continuum where individuals whose HL is the dominant language (e.g., Language A) were placed on one extreme, those whose majority language is the dominant language (e.g., Language B) were placed on the other extreme, and those with similar proficiency in both of the languages were in the middle of the continuum (e.g., AB). She argued that HLLs are distinct from each other in the way they present different strengths in both of their languages, and that these strengths together with their preference in using one or the other language fluctuates depending on the context and nature of the interaction. Like Cummins (1980), she rejected the idea of defining a bilingual as two monolinguals in one. This representation of bilinguals has been particularly helpful in describing HLLs because it is inclusive of various types of HLLs, from those whose dominant language is the HL through those whose dominant language is the majority language (i.e., the host country language). Furthermore, Valdés (2001) explained that bilingualism might change over time and that bilinguals who were dominant in one language might change to being dominant in the other language, which is a common occurrence among many HLLs.

Figure 3. Valdés’ (2005, p. 414) representation of L1/L2 users bilingual continuum
In regard to the characteristics of HLA, other scholars have also proposed a different kind of continuum, where NSs and FLLs are in the extremes and HLLs are positioned between these two extremes. Within this model, HLA was compared to L1 and L2 acquisition. Montrul (2008, 2009, 2016) summarized these differences in the three types of language acquisition in a table (Table 3) that focused on (a) exposure, (b) input, (c) outcome, and (d) motivational and affective factors.

Table 3

*Characteristics of L1, L2 and Heritage Language Acquisition. (Montrul, 2009, p. 188)*

<table>
<thead>
<tr>
<th>L1 Acquisition</th>
<th>L2 Acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early exposure to the language</td>
<td>Late exposure to the language</td>
</tr>
<tr>
<td>Naturalistic setting (aural input)</td>
<td>Instructed and/or naturalistic setting (aural and written input)</td>
</tr>
<tr>
<td>Control of features of language acquired very early in life (phonology, some vocabulary, some linguistic structures)</td>
<td>Grammar may be incomplete (no chance to develop other structures and vocabulary)</td>
</tr>
<tr>
<td>Developmental errors</td>
<td>Developmental and transfer errors</td>
</tr>
<tr>
<td>Outcome is successful and complete</td>
<td>Outcome is variable proficiency. It is typically incomplete</td>
</tr>
<tr>
<td>Fossilization does not occur</td>
<td>Fossilization is typical</td>
</tr>
<tr>
<td>No clear role for motivation and affective factors to develop linguistic competence</td>
<td>Motivation and affective factors play a role in language development</td>
</tr>
<tr>
<td>More complex structures and vocabulary developed at school after age 5, when metalinguistic skills develop.</td>
<td>Experience with literacy and formal instruction</td>
</tr>
</tbody>
</table>

The shaded cells in Table 3 represent the factors that characterize HLA. As can be observed through this table, certain characteristics of L1 and L2 acquisition overlap with HLA. Most relevant to this dissertation are the characteristics of the input that HLLs receive and their implications for oracy and literacy skills. In general, HLLs, like L1 learners, mostly receive aural input within a naturalistic setting very early in life. For this
reason, they have control over some aspects of language in phonology, vocabulary, and grammar that are usually acquired by monolingual children early in life. It is also important to highlight that the last row in Table 3, which is related to the development of literacy skills, is not shaded. Although access to and experiences with the heritage language and culture are widely different among HLLs, it is often the case that HLLs do not receive formal instruction in their HL (Montrul, 2016). This limited access to formal instruction affects HLLs’ acquisition of literacy skills. Therefore, they are different to L1 acquirers in that they are not able to learn the more complex structures and vocabulary acquired at school after the age of five, and at the same time those HLLs who have not sought formal instruction are different to FLLs in that they have underdeveloped literacy skills.

Much as with defining and operationalizing L2 proficiency, scholarly work dedicated to defining HL proficiency has also taken two different approaches: a cognitive approach, which takes the form of empirical studies exploring the difference between HLLs and FLLs in the use of discrete linguistic forms, and a holistic approach, which is characterized as being more oriented towards addressing pedagogical needs and defining proficiency in broader concepts of language competence. The following two sections offer a synthesis of the studies taking both approaches.

2.4.1. A general second language acquisition approach to defining heritage language proficiency

A large body of literature, following a general SLA approach, has set out to explore the extent to which the differences in the timing, mode, and quality of input that HLLs and FLLs receive, can influence their linguistic outcomes in phonology, grammar,
and vocabulary. In terms of phonetic perception and production, multiple scholars have indicated that due to the early exposure to aural input, HLLs have an advantage over FLLs and some, especially those with advanced-levels of HL proficiency, are similar to NSs. Most of these studies have assessed phonological abilities through NSs’ ratings of native-like perception as well as measures of voice-onset-time (VOT) for specific phonemes (Au, Knightly, Jun, & Oh, 2002; Godson, 2004; Gor, 2014; Lein, Kupisch, & de Veijer, 2015; Oh & Au, 2008; Oh, Jun, Knightly, & Au, 2003). All in all, these studies have characterized HLLs as having a native-like accent.

With respect to grammar, particularly in morphosyntax, the picture has not been as clear. While there is some evidence that suggests that HLLs are more similar to FLLs than to NSs in the way they accurately judge grammaticality of structures such as sentence agreement (e.g., Au et al., 2002) or case particles in relative clauses (e.g., O’Grady, Lee, & Cho, 2001), research focused specifically on advanced-level HLLs, has indicated that HLLs consistently outperform FLLs from the same level (S. Y. Lee et al., 2009), but still lag behind the NSs as, for example, in the case of verbs of motion in Russian (e.g., Gor, Cook, Malyushenkova, & Vdovina, 2009). Moreover, Laleko and Polinsky (2013) highlighted the importance of considering the distinct characteristics of HLLs from different HLs when interpreting the results. In their study, they examined the performance of Japanese and Korean HLLs on grammaticality judgment tests (GJT) assessing Topic and Nominative markers. They found that Korean HLLs performed more similarly to the NS group than the Korean FLLs did, whereas the Japanese HLLs were significantly different from the baseline NS group. The results of these studies have
underscored the crucial role that proficiency and HL play on the results of empirical studies.

In addition, a closer examination of HLLs’ linguistic systems has revealed a more complex interaction between linguistic abilities and time of acquisition. The scholarly work of Silvina Montrul and colleagues, for example, suggested that HLLs outperform FLLs in complex structures that are acquired before the development of metalinguistic skills (Montrul, 2008, 2009, 2010; Montrul & Perpiñán, 2011; Montrul, Foote, & Perpiñán, 2008). In Montrul and Perpiñán (2011), HLLs’ use of Tense-Aspect and Mood (TAM) morphology was examined through a series of tasks, including two written morphology recognition tasks and two sentence conjunction judgment tasks. They found that while HLLs were better than FLLs in tasks that elicited knowledge in tense-aspect, which is an early acquired structure, they were not better in mood, which is a later acquired structure. Moreover, Kim, Montrul, and Yoon (2009) found differences between early bilinguals and late bilinguals of Korean in their grammaticality judgment of Korean anaphors observed through the reflexives –caki, –casin, and –caki casin. The former bilingual group was described as HLLs who were born in the US and were exposed to both English and Korean simultaneously from birth but whose dominant language they claimed had gradually shifted to English. The latter bilingual group was described as HLLs who were born in Korea and thus were exposed only to Korean as their L1 from birth and then immigrated to the US at a later stage of childhood or early adolescence. The results of the study suggested that late bilinguals performed similarly to the NS baseline on the GJTs whereas the early bilinguals, who had acquired the HL early in life, showed incomplete acquisition of the Korean anaphor system. Apart from incomplete
acquisition, researchers also explained these results in terms of dominant language transfer. That is, they claimed that early bilinguals had transferred their knowledge of anaphors from their dominant language (i.e., English) into Korean. This claim, however, should be interpreted with caution, as it was not clear how researchers had established dominance of a language. Although it was reported that language proficiency was measured by a cloze test, no separate measurement of dominance appears to have been used. It is important to note that dominance and proficiency are two different constructs and that having a high proficiency in a language does not imply having high dominance in that language (Birdsong, 2014).

Even though these studies have shed light on some important aspects of HL proficiency, the great variability across studies, regarding the HL of focus, the specific structures under investigation, the tasks used to elicit language, and the definitions of HLLs, has prevented a broader interpretation of results, particularly for articulating the constructs of HL proficiency. Unlike models of L2 proficiency, there is as yet no comprehensive model that can explain what components of language (i.e., grammar, meaning, pragmatics) constitute HL proficiency, how these components interact, and most importantly how they can be measured. Designing these types of models becomes more challenging without verifying and consolidating results across studies. Perhaps, then, one of the first steps to address this issue should be the design and implementation of reliable instruments that can serve to establish baselines for both HLLs and FLLs as an initial stage of empirical studies, particularly for cross-sectional studies. There is a critical need to design instruments that are able to assess HLLs’ literacy and oracy skills in order
to have a better perspective of the interaction between both skills and what learners can do at different stages of their language development.

2.4.2. A holistic approach to defining heritage language proficiency

From a pedagogical perspective, scholars have adopted a more holistic approach to defining HL proficiency. Within this approach, research focuses more on describing what HLLs from different levels of proficiency can functionally do and how these data can inform language educators in the design of curricula and tests for educational purposes. Thus, unlike studies from the cognitive approach, which discuss the differences between learner types by exploring discrete linguistic features in areas such as phonology and morphosyntax, studies within the holistic approach examine these differences in terms of broader concepts of language competence, such as listening, reading, speaking, and writing. In addition, this type of research already assumes that there are indeed differences between HLLs and FLLs or NSs in their language development as well as their pedagogical needs. Therefore, discovering whether or not there are differences between learner groups is no longer the main goal in this type of inquiry, and instead attention is focused on factors, both intrinsic and extrinsic, that make each HLL different in their linguistic abilities.

Following this holistic approach, Kagan and Dillon (2012) provide a description of the typical profile of HLLs’ proficiency. This approach is similar to the second approach of defining L2 proficiency described in section 2.1.2, where language proficiency is defined based on expert opinions from highly experienced language educators. Therefore, according to HLL-experts and the results of an extensive self-assessment questionnaire conducted by the National Heritage Language Resource Center
(NHLRC survey, 2009), Kagan and Dillon (2012) noted that “listening is perceived to be and is in fact HLLs’ strongest skill. It is typically followed by speaking, then reading and writing” (p. 498). In addition, the authors indicate that in terms of vocabulary, HLLs can have a larger repertoire than FLLs, but it is limited to informal and interpersonal registers. Similarly, HLLs’ writing performances lack formal registers and complex structure. The unbalanced development of HLLs in oracy and literacy skills has been documented in multiple studies (e.g., Friedman & Kagan, 2008; Kondo-Brown, 2003, 2004) and it has been described as one major point of concern in HL instruction (Kondo-Brown, 2010). Overall, this description is in line with Montrul (2008) in that it suggests that HLLs are in general more proficient in oracy skills (i.e., listening and speaking) but relatively less proficient in literacy skills (i.e., reading and writing), due to the nature of HL input they are exposed to during their language development. In addition, this is also relevantly explained by Hulstijn’s language proficiency model. HLLs appear to show some command in language skills within what Hulstijn (2011) has described as the BLC, as they seem to show implicit knowledge of phonology, morphology, and syntax, as well as some explicit knowledge of lexis. Nevertheless, they still do not show full command of language skills in terms of the HLC, as they show difficulty in the perception and production of utterances with low-frequency lexical items and morphosyntactic structures.

HLLs’ self-reported competence in their HL is also a useful source of information that has helped describe the construct of HL proficiency. Overall, this information has been elicited through self-assessment questionnaires composed of Likert-scale type items that ask HLLs to rate their language ability in four broad categories: reading, listening,
speaking, and writing. In line with the opinion of experienced language instructors reported in Kagan and Dillon (2012), Jensen and Llosa (2007) found that HLLs from four different HL backgrounds, namely Korean, Russian, Thai, and Vietnamese, also ranked their HL abilities in the same order, with listening being the strongest skill followed by speaking, reading, and the weakest skill being writing. This difference in self-perceived proficiency across different language skills was consistent with the results of other HLLs’ studies using self-assessment questionnaires (Beaudrie & Ducar, 2005; Kang & Kim, 2012; Kondo, 1997). Overall, HLLs appear to perceive their oracy skills higher than their literacy skills.

Because no specific standards have yet been articulated for HLLs, HL proficiency has been defined most often through standards originally designed for FLLs. This is the case of the ACTFL proficiency guidelines. Kagan and Friedman (2003), for example, described Russian HLLs, who have had little or no exposure to formal education in the HL (i.e., Group 3 and 4 in Kagan & Dillon, 2008), usually exhibiting characteristics of learners from the Intermediate-Low to Advanced levels of oral proficiency according these proficiency guidelines (ACTFL, 1999). That is, these learners had a range of abilities from being able to perform simple everyday transactions through predictable, concrete, and familiar information to communicating effectively about abstract topics and solving situations with unexpected complications. Polinsky and Kagan (2007) also offered a rough description of what heritage speakers’ proficiency looked like in terms of the ACTFL proficiency guidelines and the Interlanguage Roundtable (ILR). They described speakers who were recent immigrants and whose

---

1 Kagan and Dillon (2012, p. 500) briefly mention that ACTFL is developing HLL-specific guidelines.
dominant language was the HL (i.e., acrolectal speakers), as exhibiting language abilities of the Advanced-High or Superior levels in the ACTFL guidelines or levels 2+ to 5 on the ILR. Furthermore, first or second generation HLLs who had different degrees of proficiency in their HL and majority language (i.e., mesolectal speakers), were referred as having Intermediate-High to Advanced-Mid proficiency levels on ACTFL and 1+ to 2 on the ILR. Finally, HLLs from the third or fourth generation whose dominant language was the majority language (i.e., basilectal speakers), were classified as having an Intermediate-Low to Intermediate-Mid on the ACTFL or level 1 on ILR equivalent language proficiency (p. 386). The use of these types of proficiency guidelines (i.e., ACTFL or ILR), however, is still a contentious topic (see next section for a complete discussion on this issue). For this reason, scholars have been very careful not to over-interpret these descriptions of proficiency.

2.4.2.1. **Linguistic profiles of Korean heritage language learners**

The use of proficiency standards originally created for FLLs, such as the ACTFL guidelines and ILR scale, have also been used to profile Korean language learners, including HLLs. Y.G. Lee, Kim, Kong, and Hong (2005), for example, sought to profile advanced Korean language learners (ranging from levels intermediate high (IH) to advanced high (AH) in the ACTFL guidelines) based on learners’ performances on the OPI and on a written test composed of multiple-choice and fill-in-the-blank items assessing knowledge of sentence structure (i.e., particles, connectives, and complex predicates), and lexis and collocations (i.e., collacational pairs, idiomatic expressions, and mimetics). In terms of oral proficiency, the researchers examined speech samples using CAF measures. Unsurprisingly, they found that the more proficient learners produced
more syntactically and morphosyntactically complex, as well as more accurate oral narratives. Nevertheless, when comparing the OPI performances to written test performances, which measured more fine-grained linguistic skills, researchers noted that the levels in the OPI did not match the accuracy scores in the written test. Analyses of learners’ scores on the written test divided the group of advanced learners into three groups, namely least advanced, more advanced, and most advanced. Not all learners scoring at the level of AH in the OPI scored most advanced in the written test, suggesting a gap between Korean learners’ oral and written skills.

The scores of the written test also indicated that overall, regardless of proficiency level, learners scored lower on tasks involving idiomatic expressions and collocational pairs. Interestingly, when analyzing HLLs and FLLs separately, differences in the relationship of scores on type of tasks were found, particularly in regard to the tasks assessing connectives. While HLLs scores on connective tasks were either the highest or the second highest, FLLs found connectives to be the most challenging after idiomatic expressions. Both learner groups found idiomatic expressions to be the most challenging. All in all, HLLs outperformed FLLs in all types of linguistic features. Due to the greatly unbalanced number of heritage (N=23) and FLLs (N=4), however, these results should be interpreted with caution.

In addition, a follow-up study including revised items in the written test and focusing on nine HLLs, found that even within a single advanced-level proficiency group there were differences in performance on certain linguistic features. The scores on honorifics, for example, were found to be different for all levels. The least advanced
proficiency group found these items to be among the most challenging, while the most advanced group found them one of the least challenging.

Along similar lines, as one of the LCP projects (see section 2.1.2), S. Y. Lee, et al. (2009) investigated the relationship between Korean language learners’ grammatical competence and overall proficiency as measured by the ILR scale. The researchers explored learners’ performance on eight types of tasks (i.e., GJT, acceptability judgment task, lexical decision task, picture-word discrimination, phoneme discrimination, phoneme monitoring, picture-sentence matching, and accent detection) that assessed their knowledge of receptive language skills in the areas of phonology (e.g., discriminating voiceless alveolar stops), morphology (e.g., particle stacking), syntax (e.g., tense dependency), lexis (e.g., compound nouns), collocation (e.g., light verbs idioms), and accent detection (e.g., Korean dialects). The tasks were administered to 52 language learners (32 HLLs and 12 FLLs) from a wide range of proficiencies (ILR 1+ to 4) as measured through their OPI performance.

Correlation analyses showed overall medium to strong positive correlations between the task mean scores and the ILR scale, ranging from 0.59 to 0.97, indicating that the language skills described in the ILR scale can be deconstructed by concrete tasks. The general trend in learner performance suggested that knowledge of linguistic features such as tense dependency, past tense in relative clauses, negation, locative verbs, honorifics, light verbs, morphology, compound nouns, phoneme differentiation (for picture- word discrimination and minimal pair tasks), reflexive –caki, increased following the same developmental trend described by the ILR scale. In contrast, other features such
as conjunction, particle stacking, numerals, *wear*-verbs, syllable structure, overall vocabulary, /p/ phoneme, and accent detection did not follow the same trend.

Researchers considered the differences in performance between HLLs and FLLs as one of the main causes of the overall low correlations between the knowledge of certain linguistic features and the OPI scores. They explained that HLLs outperformed FLLs in (a) GJT(s, specifically items assessing relative clauses, locative verbs, particle stacking, numerals, wear-verbs, and conjunction; (b) acceptability judgment tasks, for items assessing honorifics and light verb idioms; and (c) picture-sentence matching task for assessing the reflexive –caki. They argued that the poor performance of FLLs in these particular tasks may have been the cause of lower correlations between the accuracy mean scores OPI scores. In addition, FLLs were found to do better in lexical decision tasks, particularly items assessing syllable structure, morphological rules and compound nouns. In this case, the researchers explained that the poor performance of the HLLs may have resulted in lower correlation coefficients. As there was an unbalanced sample size for each proficiency level and since there were groups composed of as little as one participant, they recognize that they are unable to make stronger claims about these results.

To recapitulate, both studies described above shed light on some of the linguistic features that Korean learners, both FLLs and HLLs, present at certain levels of speaking proficiency as described by the ILR scales. Some linguistic features, such as past tense in relative clauses, tense dependency, honorifics among many others were good indicators of proficiency. That is, the more proficient learners were (as measured by the OPI scores), the more knowledge in these specific features they were able to show. The result
of other features, like idiomatic expressions, accent differentiation, phoneme monitoring were generally inconclusive. Overall, these linguistic features did not seem to fit with the same trend of language development as the ILR. Y.G. Lee et al. (2005) explained some of the differences between OPI scores and written test scores as a result of differences in modality. In other words, having a high level of proficiency in speaking did not translate directly to being proficient in all types of written tasks. S.Y. Lee et al. (2009) provided an alternative explanation, stating that the large differences in performances between HLLs and FLLs could have affected the results.

There are, however, some caveats to take into consideration when interpreting the results of these studies. First, due to the small sample sizes per proficiency group, the results of the studies should not be over-interpreted. The unbalanced number of participants per proficiency group and per learner type hindered the true comparability of abilities between groups, particularly when these relationships were analyzed conducting correlational analyses. Second, the studies were unable to directly address whether the ILR is an appropriate scale to assess Korean language learners in general and separately by learner type (HLLs vs. FLLs). Although the differences in modality (speaking vs. writing), as well as learner type might have been the main factors in the weak relationship between OPI scores and accuracy rate in other tests, another plausible explanation can be that the inappropriateness of ILR scale to represent a general population of language learners, those that are different from government employees. Further analyses should tease apart these confounding factors in order to investigate the extent to which the differences in modality and learner type are actually affecting the assessment of learners’ proficiency.
In sum, studies following a more holistic definition of HL proficiency shed light on some of the important issues concerning HL competence. These studies underscore the unbalance in HLLs’ proficiency skills in terms of oracy and literacy. In general, HLLs possess strong skills in listening and speaking but weak skills in reading and writing. Results from both expert opinions and HLL self-rated proficiency support this view and point out one of the main concerns in HL instruction.

Although the holistic approach to defining HL proficiency is useful for addressing pedagogical needs, there is still lack of research that teases apart the descriptions of language competence in different levels of proficiency. More concretely, there is still a need to define and differentiate what HLLs can do at different language developmental stages. Although some studies, like the aforementioned Korean linguistic correlate studies, have presented some evidence of certain linguistic features that help explain Korean language learners’ progression of language skills according to proficiency levels, there is still much room for improvement. Certain analyses have been challenged by the sampling of learners, especially in terms of the sample size per proficiency level group and per learner type. In regard to the analyses of results, the comparisons between proficiency standards (i.e., ILR scale) and other measures (e.g., written tests, GJT, lexical decision tasks) should move beyond simple correlational analyses. Most importantly, the standards and tools that were originally designed for FLLs should be properly validated for the purposes of assessing a distinct learner group, the HLL group. Similar to the research following a general SLA approach, inquiry on HL education also shows much variability across studies in terms of the instruments used to measure HL
proficiency, challenging the generalizability of results. Accordingly, this research domain is also in need of tools that can reliably measure both HLLs’ literacy and oracy skills.

2.5. Measuring Heritage Language Proficiency

Much as with the research on defining HL proficiency, only limited scholarly attention has been paid to the operationalization and measurement of HL proficiency. In particular, there has been little understanding of what entails assessing the reliability of HL assessment instruments and evaluating the validity of the interpretations of test results. Often times, these concepts have been oversimplified or ignored, and the intended uses of the tests have not been taken into consideration when evaluating the appropriateness of their use in assessing HLLs. Many researchers, for example, have noted that in research that makes comparisons between HLLs and FLLs, there is need for a common instrument that can measure both learner groups (Alarcón, 2011; Montrul, 2016), yet not much attention has been paid to designing and validating such tools that can indeed assess both learner groups. In general, it appears that practicality has been the most important criterion when choosing an indicator of HL proficiency. Thomas (1994) explained that in addition to reliability and validity, practicality has been one important concern when it comes to assessing L2 proficiency. In fact, Bachman and Palmer (1996, 2010) as well as Bachman (2005), included practicality as one of the six qualities of test usefulness. Nevertheless, in HLL research this specific concern may have overshadowed reliability and validity. The majority of research in the domain has preferred the use of readily available or easily accessible indicators of HL proficiency. In fact, similar to the trends of L2 proficiency measurement observed in the synthesis studies by Thomas
(1994, 2006), Tremblay (2011), and Hulstijn (2012), the four most common indicators of HL proficiency in HLL research include, (a) instructional level; (b) scores on standardized tests; (c) self-assessment scores; and (d) locally-designed test scores.

2.5.1. Instructional level

For the most part, studies related to HL learning provide some type of information regarding HLLs’ instructional level (e.g., whether they belong to the first semester or fourth academic year) in the section that describes learners’ demographic information. At times, this information has been taken as the sole criterion to classify HLLs into different language proficiency levels. For example, HLLs from the first or second semester would be considered as having a beginning-level language proficiency, whereas third or fourth semester and fifth and sixth semester HLLs would be referred to as having intermediate- and advanced-levels of language proficiency, respectively. Moreover, as mentioned earlier, for programs composed of separate tracks for FLLs and HLLs, the differentiation between HLL- and FLL groups have usually been based on the enrollment of the learners in these tracks. In other words, learners enrolled in HLL-specific track have been considered HLLs, and those enrolled in the other track have been categorized as FLLs. More often than not, the original methods adopted to separate learners into separate tracks have not been reported. Despite the rigorous placement criteria that HL or FL programs might have indeed followed, this type of proficiency indicator and HLL-identifier is problematic, as institutions may vary in curricula, materials, time of instruction, and types of learners, among many other programmatic characteristics. Thus, this variability across language programs may render data on instructional level
unreliable for the comparability of results across studies that consider this information a proxy for HL proficiency or HLL identification (Norris & Ortega, 2012).

2.5.2. Scores on standardized tests of language proficiency

Scores on standardized language proficiency tests have been another common HL proficiency indicator. One of the most frequently used proficiency tests has been the ACTFL Oral Proficiency Interview (OPI) (Gor et al., 2009; S. Y. Lee, 2012; Y. G. Lee et al., 2005; S. Y. Lee et al., 2009; Long et al., 2012) or any variation of this test, including the computer-based OPI, the OPIc (Martin, Swender, & Rivera-Martinez, 2013; Swender, Martin, Rivera-Martinez, & Kagan, 2014) and the Simulated Oral Proficiency Interview (SOPI; Kagan & Friedman, 2003). These tests are standardized tests of oral proficiency, where learners are asked to interact with a certified ACTFL interviewer, a computer avatar, or a digital recording. Certified raters carry out the rating of speech samples according to the ACTFL guidelines (see ACTFL, 2012a). As described in the previous section, several studies have described HLLs’ linguistic abilities in terms of these proficiency guidelines, as they are highly prevalent and influential in the foreign language education community. In point of fact, Kondo-Brown (2010) noted that these proficiency guidelines are “likely to continue influencing the FL and HL education in the United States” (p. 27). Another standardized test that has been commonly used to measure Spanish HLLs’ language proficiency is the Diploma de Español como Lengua Extranjera (DELE) or some adaptation of this test (e.g., Montrul, 2004, 2005; Montrul et al., 2008; Montrul & Foote, 2014; Montrul & Ionin, 2012; Pascual y Cabo, & Gómez, 2015; Torres & Sanz, 2015). DELE is a standardized test that certifies individuals as having mastery in the Spanish language. It follows the Common European Framework of
Reference (see Council of Europe, 2014) and the exams are developed by the Universidad de Salamanca (DELE, 2006; see also Garvida, 2009). Other types of standardized tests observed in the HL literature include the Test of Proficiency in Korean (TOPIK; NIIED, 2014), which is a test developed by the Korea Institute of Curriculum and Evaluation (KICE) (E. J. Kim, 2006).

Multiple advantages have been associated with the use of these standardized tests in the HL literature. In the first place, these are tests that are readily available and well known among SLA researchers and language education community. Therefore, researchers can either administer the test or ask their participants to self-report their scores on these tests if they have already taken them. Additionally, all of the abovementioned tests and guidelines also include the assessment of HLLs or of learners from different backgrounds as one of their test uses. The TOPIK, for example, is described as a test for both Korean FLLs and overseas Koreans (i.e., HLLs). As a result, these tests have been considered a useful tool for measuring both HLLs and FLLs, especially for research focused on the comparison of these learner groups.

Nevertheless, the use of these standardized proficiency tests for assessing HLLs has also been a largely debated topic. In the case of the ACTFL OPI or the ACTFL proficiency guidelines, numerous scholars (e.g., Fairclough, 2006; Draper & Hicks, 2000; Valdés, 1989) have argued against their use as these tests have been initially designed for FLLs, and thus have followed a different type of language development continuum. Furthermore, as Valdés (1989) has pointed out, these guidelines do not take into consideration the different language varieties that HLLs might use. By contrast, others (e.g., Alonso, 1997; Kagan, 2005; Kagan & Friedman, 2003; Martin et al., 2013;
Swender et al., 2014) have provided arguments in support of the use of this test and guidelines for measuring HL proficiency. They have asserted that the guidelines follow different criteria of assessment, including function, content, and accuracy and therefore can be used for assessing both learner types. Moreover, if the main goal is to assess how functional HLLs are in communicating in a language, rather than to compare between the performances of learners and NSs, then there is no need for separate assessment criteria for HLLs and FLLs. In addition, in terms of language varieties, Kagan and Friedman (2003) have indicated that HLLs of languages other than Spanish, such as Russian, mostly use and are exposed to the standard variety of the language.

In a similar vein, Carreira and Potowski (2011) have criticized the use of DELE for assessing HLLs, as it typically involves the assessment of vocabulary and morphology. They have questioned the appropriateness of this tool for matching proficiency levels of HLLs and FLLs. Without an additional measurement of speaking performance, they discouraged the use of such tests for assessing HLLs. Another disadvantage worth noting is that the majority of these standardized tests have not been made publicly available and as a result can be very costly for some research (Lee-Ellis, 2009).

2.5.3. Self-assessment scores

Self-assessment questionnaires have been another useful tool to estimate how proficient HLLs are when performing different tasks in multiple contexts. They have predominantly been used in HL research that focuses on HLLs’ ethnic identity, motivations, and attitudes towards the HL. These questionnaires have been often used in combination with other indicators of proficiency like instructional level and/or scores on
other tests. They have been characterized as being different across studies in their design, more specifically, in length, type of item, and content. Some studies, for example, have included only four to five broad can-do statements regarding HLLs’ ability to understand, speak, read, and write in the HL on a Likert-scale from 1 to 5 (e.g., Laleko & Polinsky, 2013), none to fluent (e.g., Beaudrie & Ducar, 2005), or from native-like to non-native like (e.g., Alarcón, 2010; Montrul & Bowles, 2010). Others, on the other hand, have included an extended number of detailed questions, including HLLs’ HL use as well as their ability to perform specific tasks such as counting to 10, giving directions on the street, describing the US education system (Kondo-Brown, 2005, p. 579), or reading newspapers, writing about themselves, and talking on the phone (E. J. Kim, 2006, pp. 197-199). In addition, while most of the studies have designed tests consisting of Likert-scale type items (e.g., Alarcón, 2010; Kang & Kim, 2012), there have been a few studies that also included open-ended questions requesting HLLs to provide lengthier explanations of their abilities in certain contexts (e.g., Jensen & Llosa, 2007). The Likert-scale type items have usually been interpreted by either counting the number of learners who answered a certain scale-point (e.g., 60% of students responded being able to count to 10) or as scores of proficiency. The former method has been particularly useful for studies that have been interested in finding patterns across different types of HLLs (i.e., low- or high-proficiency HLLs).

Overall, self-assessments can provide meaningful insights into HLLs’ HL proficiency, and more importantly, they are uniquely advantageous as they can elicit information on HLLs’ own perception of their linguistic abilities. However, similar to the issues with using instructional level as an indicator of proficiency, the differences in the
design of self-assessment questionnaires across studies can impede the generalizability of results and a more comprehensive interpretation of them. Furthermore, there is still a lack of understanding of the ability of HLLs to appropriately and accurately self-assess their proficiency. Studies exploring this issue have yielded mixed results. Some studies (e.g., Martin et al., 2013; Swender et al., 2014) found that Russian and Spanish HLLs have the tendency to overestimate their speaking ability, particularly those with higher-levels of proficiency (i.e., learners who self-reported having advanced levels of oral proficiency had an actual competence of an intermediate level). On the other hand, Donovan, Di Silvio, Malone and de la Torre (2012), for instance, showed that Mandarin HLLs with actual advanced levels of proficiency accurately self-assessed their proficiency, but beginning-level learners tended to underestimate their speaking proficiency in OPIs. Interestingly, Kang and Kim (2012) suggested that the under- and over-estimation of HL proficiency can be connected to HLLs’ ethnic identification to the heritage culture. Their results indicated that Korean HLLs who strongly identified with the heritage culture, measured through J. S. Lee’s (2002) ethnic identity scale, tended to overestimate their speaking ability, while those who had a weak ethnic identity underestimated their proficiency in the same skill. However, for the writing skill self-ratings, no connection was found with ethnic identity measures. In sum, HLLs’ ability to self-report their language competence appears to be at times unpredictable, and numerous variables may influence the outcomes of these self-assessments. Therefore, these self-ratings require careful interpretation when being used as indicators of HL proficiency.
2.5.4. Scores on locally-designed language proficiency tests

Locally-designed language proficiency tests have been an additional type of instrument used to assess HL proficiency sometimes observed in the HLL literature. As Thomas (1994) described in her synthesis on L2 proficiency measurements, these tests are either specifically designed and implemented for the purposes of a specific study, or are readily-available tests used for program-specific purposes, like placement or diagnostic tests. As these tests are individually designed with a different underlying research purpose, they vary considerably in the type and number of task or items they are composed of and the language competence they intend to measure.

The majority of the locally-designed tests in the HL literature have been described as consisting of multiple-choice items that usually measure grammar and vocabulary (Alarcón, 2010), in addition to receptive language skills, such as listening and reading comprehension (Kondo-Brown, 2005). Nevertheless, there have been a few other local tests that assess language production, as in the case of performance-based tests, that measure speaking and writing (Kang & Kim, 2012) and cloze tests, which provides a more global measure of language proficiency (Kim et al., 2009; Lee-Ellis, 2011; Montrul & Bowles, 2010). Performance-based tests might have been less commonly used relative to the other types of local tests because, as opposed to dichotomously scored multiple-choice items and cloze test items, they require a rigorous process of training raters to score on a specific rating scale and rubric. In Kang and Kim (2012), for example, NS raters scored HLLs’ performances in terms of the amount and quality of learners’ output in spoken and written samples through a six-point scale.
Although not designed specifically with one research study in mind, program-specific tests can also be considered locally-designed tests as they have generally been developed for a specific local institution. On the one hand, the use of these tests to measure HL proficiency can be beneficial in terms of their ready availability and easy access. In most cases, these tests are pilot-tested prior to use, providing a more accurate estimate of performance. Kondo-Brown (2005) used the University of Hawaii at Manoa (UHM) Japanese placement test, which she described as going through major revisions after one year of piloting. As a result, she reported high reliability coefficients and strong correlations among the subsections of the test (i.e., listening, grammar, and reading comprehension). On the other hand, the performance on these tests might not be reflective of a global language ability since they were not designed for that purpose but rather for purposes of diagnosis or placement, for example. In the case of curriculum-based placement tests (e.g., Beaudrie & Ducar, 2012; Potowski, Parada, & Morgan-Short, 2012; Sohn & Shin, 2007), their design articulates a specific HL or FL program curriculum. Therefore, these tests measure language proficiency following instructional levels of a specific curriculum. As the needs and aims of language programs vary across different institutions, using scores of placement tests as an indicator of HL proficiency, might lead to the same issues as using instructional levels. That is, the variability of proficiency indicators can undermine the generalizability and comparability of results across studies. Moreover, the majority of these tests have solely assessed receptive skills or knowledge of grammar and vocabulary, which would only account for part of HLLs’ linguistic abilities. Lastly, although some studies have reported reliability coefficients,
test validation procedures have often times been forgotten or confounded with pilot-testing on NSs.

On the whole, the aforementioned four indicators of HL proficiency appear to be commonly selected for most research purposes due to their practicality. The instruments used for obtaining such proficiency indicators are, for the most part, accessible, available, or easy to create, giving them a misguided sense of versatility. However, reliability of such measures and the appropriate validation of the interpretation of results are crucial, and more scholarly efforts should be directed to this endeavor. In addition, these instruments should be examined carefully in terms of what kind of information they truly provide. There is still little understanding of what constructs they are assessing, particularly with regard to literacy and oracy or Hulstijn’s (2011, 2015) dimensions of language proficiency (i.e., BLC vs. HLC). Finally, as described throughout this section, another major concern is using a language proficiency measure that is comparable across different studies and that thus contributes to the generalizability of results and a truly systematic accrual of knowledge. Scholarly attention should also be directed towards using tools that can yield such comparable results.

2.6. Summary

As presented throughout this chapter, while research in the domain of HLA has made important advancements that have furthered our understanding of who HLLs are and what types of language abilities they have, there remain several points of concern in the literature that require more scholarly attention. Principally, research following both a general SLA or a holistic approach has yet to concretely define the construct of HL
proficiency. One of the main causes for this gap in the literature is related to the limited generalizability of results across studies, which has been the result of the great methodological variability in the way HLLs are identified, classified, and assessed. The inconsistent use of indicators of HL proficiency has challenged the systematic accumulation of research findings in the domain. This limitation has in turn been related to the lack of reliable assessment tools designed for the purposes of measuring learners’ proficiency. Although some scholarly work has filled in the need of such instruments by using those originally designed for assessing FLLs, there has been limited effort in validating the use of such instruments for assessing HLLs, who are a very different target learner group.

Given these limitations, the present dissertation research specifically focused on examining the use of an instrument that can also assess HLLs’ general proficiency for research purposes. It was important that the proposed tool of assessment met the following requirements: (a) be quick in providing evidence of general language proficiency; (b) be readily-available for researchers conducting studies involving HL proficiency as an independent or moderating variable; and most importantly (c) have followed a rigorous validation framework that helps evaluate the evidence that supports (or rejects) the interpretation of results based on the purported uses of such instruments.
Chapter 3

C-tests

Thus far, some of the most critical gaps in the applied linguistics research, particularly in the HLA research, have been discussed and one major limitation has been identified as the lack of assessment tools that can reliably assess learners’ general language proficiency. C-tests can help address this critical gap for the following reasons. First, this shortcut measure is particularly useful due to their practicality and free/easy access, which means it can be used by any researcher. Second, unlike other indicators of language proficiency, such as instructional level or self-reported scores on standardized tests, which differ from study to study or are very costly, C-tests can provide benchmarks for comparability and generalizability across studies as they can be made available to any researcher. Third, these tests are very practical. A C-test composed of 5 texts can be administered in less than 30 minutes and scored in 1-2 minutes per text (Eckes & Grotjahn, 2006; Grotjahn, 2000; Klein-Braley, 1994; Raatz & Klein-Braley, 2002). Most importantly, numerous studies that focused on the design and implementation of such tools, have demonstrated their high-levels of internal consistency as well as high reliability for measuring learners’ overall proficiency (e.g., Babaii & Ansary, 2001; Eckes & Grotjahn, 2006; Harsch & Hartig, 2015). This chapter describes C-tests and surveys the studies that have explored the usefulness of such tool and how it has been incorporated in the HLL literature.

3.1. What Are C-tests?

C-tests belong to the same family of tests of reduced redundancy such as cloze tests and dictation, (Klein-Braley, 1997; Raatz & Klein-Braley, 2002; Oller, 1979). The
principle of reduced redundancy is based on the notion that people often encounter situations in real-life where they have to deduce the meaning of distorted or incomplete messages by using the redundant components surrounding the message. Reduced redundancy tests try to measure the ability of language users to do so, relying on the assumption that the more proficient a learner is, the easier it is for them to correctly complete or guess the missing components (Coleman, 1996; Eckes, 2011; Klein-Braley, 1997; Spolsky, 1971). Therefore, C-test have been claimed to be useful in providing a holistic estimate of language proficiency (Klein-Braley, 1997; Raatz & Klein-Braley, 2002). Although the question of what exactly is being measured by C-tests is still a contended topic (more on this in Chapter 7), most experts agree that a complex level of language processing is required to be able to successfully complete these tests (Grotjahn, Klein-Braley, & Raatz, 2002). Test-takers must understand the overall meaning and structure of the C-test passages in order to complete the deleted parts of the words, and thus both receptive and productive skills are being measured.

C-tests were created as a way to improve upon some of the limitations of the classical cloze test, such as the lack of random samples of the elements in the text, inconsistent difficulty rate of texts when deletion rates differed, poor reliability and validity coefficients when testing homogeneous learners, among many others (see Klein-Braley & Raatz, 1984). In C-tests, ‘noise’ in the form of incomplete words is introduced to a self-contained authentic text that is usually one paragraph in length or longer. Each paragraph commonly contains about 20 to 25 incomplete words. The general deletion rule consists of deleting every second half of the second word starting from the second sentence (Grotjahn, 2000; Grotjahn, Klein-Braley & Raatz, 2002). There are, however,
some exceptions to this rule. For example, generally monosyllabic words are not deleted and are skipped in the counting. Also, in the case of compound words, only the second half of the second component is deleted (e.g., greenhouse → greenhou__). The scoring of each item within a text is dichotomous (i.e., right or wrong) with the consideration that there can be multiple correct answers. Usually these answers are determined based on pilot testing the C-test with native speakers. As the successful completion of each item is dependent upon other items within the text, each self-contained text is considered a super-item and it is scored in a scale of 1 to 25 (Raatz & Klein-Braley, 2002).

Even though most of the C-test studies followed the standard deletion rule described above, several studies investigated the effects of different deletion rates in measuring general language ability (e.g., Baghaei, 2011a, b; Lin, Yuan, Feng, 2008). The studies by Baghaei indicated that while a higher deletion rate led to increased internal reliability as seen through the measures of Cronbach’s alpha and Rasch measurement of fit, the C-test as a whole still measured one single construct. Furthermore, different languages, in particular those with non-roman alphabets, have tested out different deletion methods (Norris, 2018). In Korean and Japanese, for example, the definition of a word differs from that of most Western languages. For instance, in these languages postpositional particles (e.g., nominative and accusative particles), which are bound morphemes, have sometimes been considered a word on their own, thereby affecting the way they are or are not deleted (Roos, 1994). Overall, the development of C-tests and the flexibility of the deletion rule has been dependent upon the unique characteristics of different languages as well as the intended uses of the test (e.g., research purposes, placement purposes).
3.2. Korean C-tests

3.2.1. Previous Korean C-test developed by Lee-Ellis (2009)

Relative to C-tests assessing languages like English (e.g., Baghaei, 2008; Babaii & Ansary, 2001; Babaii & Jalali Moghaddam, 2006; Connelly, 1997; Harsch & Hartig, 2015; Jafarpur, 1995), for example, there has been dearth of literature oriented towards the design of C-tests for assessing LCTLs, such as Korean. Lee-Ellis (2009) is perhaps the first study that has addressed the need for such a tool for assessing Korean language learners and thus will be discussed here in detail. In her study, Lee-Ellis thoroughly described the design and implementation of the Korean C-test providing statistical evidence to support the reliability of such tool. The design of the test was carried out following two steps: text selection and C-test deletion method implementation. For text selection, the researcher consulted a wide range of texts from authentic contexts, such as textbooks, newspapers, magazines, and also created some of these texts when necessary. A total of ten texts were created or selected. Next, expert opinion from a DLI-certified rater was collected in regard to the ranking based on the difficulty of the texts. The expert also provided guidance in terms of which texts seemed more appropriate for the purposes of developing a C-test. Five out of the ten texts were chosen as appropriate, from which four had been created by the researcher based on Korean as a Second Language (KSL) textbooks and one had been selected from a Korean grade school textbook.

During the second phase of C-test design, that is, the C-test deletion method implementation, Lee-Ellis considered the Korean language specific characteristics. For example, it was important to establish, that unlike languages with Roman alphabets, the
The unit of orthography in Korean would be a syllable and not a consonant or vowel. Therefore, the deletion rule was applied to syllables, rather than letters. It was also crucial to define the boundaries of a word. As Korean spacing does not coincide with the boundaries of words, particularly for postpositional particles (e.g., object marker –lul or subject marker –ga), Lee-Ellis decided to consider such particles as part of the words that they appeared with. As a result, the deletion rule consisted of deleting every second syllable and everything thereafter (i.e., including postpositional particles and verb inflections) of every second content word.

In order to examine the reliability of the C-test and its items, Lee-Ellis administered the Korean C-test to 37 participants, nine FLLs and 28 HLLs. The scores were analyzed using Rash measurement statistics. Results of the Partial Credit Model (PMC) of the Rasch Analysis, showed that the test was able to distribute test-takers into different proficiency levels with an IRT item separation reliability of 0.93 and a person separation reliability of 0.97. These findings pertained the analysis of superitems. In other words, as the completion of the deleted words in a C-test passage required an overall understanding of the whole passage, all the items were considered to be interdependent. Thus, analyzing each passage as a superitem was considered more accurate than examining each item separately. The final version of her test was composed of four texts, which varied in difficulty to cover a wide range of proficiencies represented through the descriptors of the ILR scale.

Due to its ready-availability and efficiency in providing a quick measurement of general language proficiency, some scholars (Ahn & Kim, 2016; Hwang & Lardiere, 2011; Mueller & Jiang, 2013) investigating the acquisition of Korean as a foreign
language have found this tool very useful. Hwang and Lardiere (2013), for example, used this C-test to distribute participants into four different proficiency levels in their study on the use of the Korean plural marker -tul. They reported a significant difference between all four proficiency groups on their performance in the C-test (Welch’s $F(3, 30.439) = 154.910, p = .000$). These studies are examples of research that have implemented the C-test as a measurement of language proficiency to categorize participants into proficiency groups and examine proficiency as a moderating factor, or as an initial measurement of proficiency prior to treatment.

Although practical, useful, and reliable, there are some aspects of Lee-Ellis’ (2009) C-test that need to be more carefully examined, particularly in regard to the interpretation of its results and its uses. First, it is important to note that this C-test was designed with the primary intended use of measuring general proficiency of Korean language learners. Nevertheless, some of the decisions made in the process of selecting texts might have challenged this original purpose. Firstly, the text selection for the C-test passages and their ranking of difficulty were based on the ILR scale. Although the scale has been used for large-scale assessment purposes, it is worth noting that it was created primarily for government employees. Therefore, the scale itself was designed for a specific learner in mind that may be fundamentally different to other types of learners, such as the regular FLLs from post-secondary institutions. The studies that have used this tool for research purposes have mainly investigated Korean learners in four-year post-secondary institutions.

Secondly, even though Lee-Ellis underscored the importance of using passages that represented authentic samples of texts, the initial pool of ten passages might have not
been sufficiently representative of a wide range of authentic contexts. Moreover, the final pool of passages were mostly created by the researcher based on Korean textbooks, which again questions the authenticity of the texts. A wider range of contexts as well as more authentic texts can work better particularly, since the main intended use pertains the assessment of general language proficiency.

Another caveat, and one that Lee-Ellis (2009) also pointed out in her study, is the heterogeneity of the learner population to whom the C-test was administered. As noted above, the participants in the study included both FLLs and HLLs, from which the latter learner population was more representative. The findings reported in the study made no reference to the possible differences between the performances of both learner groups. Overall, the C-test showed high levels of reliability to assess a wide range of proficiency levels. However, there remain questions about the learner type specific results. For example, it is unclear whether some of the misfitting test-takers found in the study were due to the inherent differences in learner characteristic between FLLs and HLLs. Likewise, as the data from both learner populations were analyzed as a whole, it is difficulty to tease apart the differences in the level of text difficulty which could be learner type-dependent.

3.2.2. New Korean C-test developed by Son, Kim, Cho, and Davis (2018)

Motivated by the aforementioned limitations of Lee-Ellis’ (2009) C-test, a new Korean C-test (Son et al., 2018) was designed and piloted to improve upon some of these issues. While there were indeed some similarities between both versions of the Korean C-test, such as the primary intended use, which is to provide a quick measurement of Korean language learners’ general proficiency, as well as to a certain extent the
implementation of the C-test deletion technique, there were also several differences. In particular, there were differences in text selection procedure, word deletion method, and target learner population.

In terms of the text selection, the new study carried out a more rigorous process of selecting and assessing difficulty of the passages. The authenticity of the texts was considered one of the most important characteristics of the test. Thus, the researchers deemed crucial (a) to use proficiency scale that targeted a more Korean general learner population; (b) to retrieve texts from authentic sources and from a wide variety of contexts, and (c) to collect expert opinions about the difficulty of the text. As a result, it was considered more appropriate to use the TOPIK proficiency scale descriptors for text selection rather than the ILR scale. Since the TOPIK is a test that targets a general population of Korean language learners from around the world (NIIED, 2014), its rating scale lent itself better for the purposes of this C-test. In addition, the initial pool of possible passages to be included in the C-test was 24. These texts were selected from a wide range of contexts from newspaper articles, internet blogs, advertisement, and textbooks. Out of these 24 texts, 15 were chosen by the C-test development team based on the authenticity of the text source (i.e., the team preferred to have texts that were from original sources rather than from textbooks), the structure of the text, and how it lend itself to the C-test deletion rules. Subsequently, a total of nine experienced Korean language instructors were recruited to respond to a questionnaire regarding their perception of the passages and the difficulty of the 15 passages. The experts were asked to rate and rank the difficulty of the passages and based on their expert opinion, the 15 texts were organized in 5 levels of difficulty and were piloted to 37 Korean NSs.
According to the results of the NS pilot, the text pool was further reduced to 10 texts to be administered to NNS.

In regard to the application of the C-test technique to the passages, it followed similar rules to Lee-Ellis (2009) with the difference that the postpositional particles were counted words on its own. The rationale behind this difference was the provision of more context from which the learners could guide themselves to complete the passages. As the postpositional particles in Korean carry the important function of, for example, indicating whether the preceding word is a subject or an object, it was noticed that the deletion of those elements in some sentences were prone to producing more alternative answers. As the new C-test followed the dichotomous scoring method (i.e., right or wrong, no partial credit), it was important to reduce the number of alternative answers. It is worth noting, however, that based on the responses from the NS pilot, there were some items for which an alternative answer was allowed. Another feature different to Lee-Ellis’ study was the use of a Korean language syntactic parser (Korea Univ. Natural Language Processing Lab, 2005). The parser categorized all units in a sentence into different parts of speech. Thus, it allowed researchers to be more consistent and objective about the decisions made in regard to defining what counted as a word.

The C-test composed of 10 passages was pilot tested on 38 Korean language learners, 16 FLLs and 12 HLLs, from different universities across the United States. While each of the passages in the C-test were considered polytomous items with the same structure and number of scoring points, the difficulty level of the items as well as the score range for each item differed, thus the test results were analyzed through a Partial Credit Model (PCM) of Rasch Analysis (Knoch & McNamara, 2015). After thorough
analyses of person fit statistics, four participants were found to misfit the model, and thus were excluded for further analysis. According to the item fit statistics there was only one misfitting item with a zSTD score of -2.2. Therefore, this text was one of the first candidates for elimination. For more efficiency and practicality, it was decided that the final pool of passages for the C-test should contain at least 5 texts. For this reason, following several iterations of item and person fit analyses a final pool of 5 texts, which best fit the model, were selected. The final version of the C-test showed a person separation span of 5.23 logits with a model reliability coefficient of 0.96. These results showed that the new Korean C-test was able to reliably distinguish at least five proficiency levels using at least five polytomous items.

3.3. C-test in the Heritage Language Literature

In the HLA research, cloze tests or cloze passages have been preferred over C-tests as one measurement of global proficiency (e.g., Kim et al., 2009; Montrul, 2004, 2005). Numerous Spanish HLL studies (e.g., Montrul & Bowles, 2009; Montrul & Perpiñán, 2011; Montrul et al., 2008; Torres & Sanz, 2015) have consistently used the same cloze test for comparability purposes. The test has been described as a 30-multiple-choice-item cloze test, which is part of the Diploma de Español como Language Extranjera (DELE). Generally, these studies have justified the use of this cloze test by reporting its popularity among prior studies exploring HLLs as well as some reports of high internal consistency of the items in the test (e.g., Montrul & Bowles, 2009 reported a Cronbach’s alpha of \( \alpha = 0.84 \)). Another reason for preferring this test format appears to be its ready availability and easy access.
Although many have suggested C-tests as a better tool for assessing global proficiency when compared to cloze tests (e.g., Grotjahn, 2000; Klein-Braley, 1997; Klein-Braley & Raatz, 1984; Raatz & Klein-Braley, 2002; cf. Hulstijn, 2010), in the heritage learner literature the use of C-tests has been seldom reported. In one example, Lee-Ellis (2011) used the C-test she had designed in Lee-Ellis (2009) to assess Korean HLLs’ language proficiency. This study is one of the very few studies that used a C-test to assess HL proficiency. In this study, she explored how Korean NSs and HLLs process relative clauses and, unlike previous studies on this topic, Lee-Ellis was able to explore proficiency effects on the performance of learners with relative clauses, as she had a separate measurement of HL proficiency. In another study, Ahn (2014) used the same C-test to measure the proficiency of Korean foreign and heritage language learners for the purposes of exploring their interpretation of reference in discourse. The C-test scores allowed the grouping of participants into three different proficiency groups. Ahn found no significant differences in proficiency between FLLs and HLLs as measured by the C-test ($t(31) = 1.14, p = 0.27$).

Similar to the application of cloze tests in the Spanish HL studies, there is need to use a separate measure of general proficiency for research on different HLs. One major concern is whether these tools which had generally been designed for FLLs are appropriate for assessing HLLs. As Norris and Ortega (2003) have advised, researchers need to be careful about the valid test fallacy. Even though some measures might seem to be “valid” in previous studies or contexts, they might not be as effective for the specific uses of other studies. It is important to emphasize that validity is not a property of the test itself but a process in which the intended uses and interpretation of test results are
evaluated (more on this in Chapter 4). In other words, the use of these tests and the interpretation of their results should be validated for the specific purposes of a given study. Therefore, larger-scale validation studies, such as the present dissertation study, are crucial for examining the appropriateness of such shortcut measures for assessing HLLs’ global language proficiency.

In sum, as documented by the studies reviewed throughout this chapter, this shortcut measure has yielded promising results when it comes to measuring learners’ general language abilities for research purposes. The C-test has provided a solution for the lack of reliable instrumentation for measuring proficiency. Nevertheless, its applicability for different types of learners as well as the validity of the interpretation of its results and uses remains unclear. A thorough evaluation guided by the purported intended uses should be conducted prior to their use and applicability. In particular, it is crucial to investigate the construct that the test is measuring and whether this construct is independent of the learner population.
Chapter 4

Validity

Similar to the models for defining language proficiency, the definition of validity has undergone multiple changes over time. From simply being defined as a property of a test that indicates whether the test is measuring what it is supposed to measure, validity has become a much more complex concept that underscores the importance of the interpretations of test results and the decisions made based on them. The most recent publication of the Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014) defines validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” and describes the process of validation as “accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations” (p. 11). Unlike the 1999 Standards (AERA, APA & NCME, 1999), the 2014 Standards makes it more explicitly clear that validity is not a property of the test itself but rather an evaluation of the interpretations for specified uses of the test. Thus, it adds that “statements about validity should refer to particular interpretations for specified uses” and that “it is incorrect to use the unqualified phrase ‘the validity of the test’” (p. 11). Furthermore, it also highlights the importance of cost-benefit considerations (i.e., weighing the negative with the positive consequences of a test) when justifying the use of a test, which were not evident in the previous Standards version.

This chapter provides a brief overview of the historical evolution of the conceptualization of validity and its influence in the language testing research community. This brief historical glance into the different frameworks of validity helps
situate Kane’s (2001, 2006) argument-based approach to validity while highlighting some of its unique advantages, particularly when it comes to validating the interpretations of the shortcut measure proposed in the current dissertation study.

4.1. Trinitary Model

In the early 1950s, validity was narrowly defined in terms of 4 types of validity, namely, content, predictive, concurrent, and construct validity. These types of validity were considered a set of tools that were useful to validate specific types of tests (Kane, 2001). As a result, content validity was defined as the evaluation of how well the test content represented real-life tasks, and was used as a tool to validate tests such as achievement tests, including the SAT and GRE, among others. This type of validation involved consulting with experts in the field about the relevance and representativeness of the test content to the real-life domain tasks and behaviors. Predictive validity consisted of the evaluation of the extent to which the performance of test-takers on one test could indicate future performances at a later time. This type of validation was used as a method of validating aptitude, placement, or selection tests. Concurrent validity referred to the estimation of the extent to which test scores from one test were related to the scores of another already ‘validated’ test that was meant to measure the same criterion. This type of validation was associated with the validation of short-cut approximations of tests. In the 1975 Standards, predictive and concurrent validity were joined into a criterion-related validity type. It was mainly operationalized as correlations between the new test and what was already considered a ‘valid measure.’ Finally, construct validity consisted of the evaluation of the degree to which the successful
completion of a test could be attributed to a test-taker’s set of skills or psychological constructs. These constructs were theoretically-defined and inferred from evidence of “interrelations of the test scores with other variables, internal test structure, observations of response processes, [and] the content of the test” (AERA, APA & NCME, 1999, p. 174). Therefore, as Messick (1989) noted, this type of validity assumed almost all of the other validity evidence types and became the center piece of all validation.

Accordingly, between the 1950s and the 1970s, validity was conceptualized in terms of these three types of validity (i.e., content validity, criterion-based validity, and construct validity), hence the term “trinitary model” of validity. During this time, the focus of validation studies was on answering the question of how well the test measured what it was supposed to measure, or in the words of Cureton (1950) “how well a test does the job it is employed to do” (p. 621). Validity was, therefore, examined primarily through quantitative analyses that would determine the extent to which the test was accurate in measuring a particular behavior of interest or language skill. As Norris (2008) has noted, limited attention was given to the interpretation of test results and the decisions made according to the intended uses of the test. Validity was a property of the test and something that could be established, and thus thinking in terms of ‘a valid test’ or ‘an already validated test’ was common practice.

The Trinitary Model strongly influenced the way scholars conceptualized and evaluated validity. All three validity types had the advantage of being specific in regard to the kind of evidence needed to establish validity, which made this model appear practical (Kane, 2006). Moreover, this validity model lent itself well to the strong focus on quantitative measurement at the time. Nevertheless, this model and its subcomponents
were also subject to strong criticism, primarily for a lack of consideration of the intended uses and users of the test and the evaluation of the decisions made based on test results (e.g., Kane, 2001, 2006). Goodwin and Leech (2003) have also noted that there was no clear cut difference between construct validity and the other types of validity, which created confusion when conducting validation studies. In addition, all types of validity were considered equally important without further consideration of test result interpretations. Norris (2008) has added that there was also limited specific information regarding the amount of evidence needed for validation, the individual or individuals responsible for carrying out the validation process, or the appropriate method(s) to adopt for validation inquiry, among several other related questions (p. 39). In addition, one of the disadvantages of content validity, in particular, was its high degree of subjectivity, as the validation method consisted of expert judgments. Messick (1989) also pointed out that this type of validity provided no direct evidence of validity, as it was based on a sample performance. The criterion-based validity was criticized for its flawed logic (i.e., Test A is valid because it highly correlates with Test B, which was proven to be valid through the high correlation coefficients with a previous valid measure, Test C…) and proneness to circularity and infinite regress (Kane 2001, 2006). Construct validity, on the other hand, was problematized for its overreliance on well-defined formal theories that defined the constructs (see next section for more information on the limitations of this type of validity).

4.2. Redefinition of Construct Validity

The Trinitary Model remained highly prevalent through the 1980s, greatly influencing the way tests were developed and validation studies were conducted. In spite
of its sustained popularity, the model was also constantly challenged by scholars due to
its narrow definition of validation that disregarded a true evaluation of the interpretations
of test results. That is, these types of validity and their operationalization mainly based on
correlations failed to address the real meaning of test results (e.g., what does a correlation
of .8 mean between test A and test B in terms of the skills or behaviors intended to be
measured?). As a result, this validity model underwent several changes particularly when
defining construct validity.

In a seminal article in the *Psychological Bulletin*, Cronbach and Meehl (1955)
defined construct validity as the process of empirically evaluating the relationship
between test-takers’ observable attributes and test constructs. Inquiries related to
construct validation focused on answering the question of “what constructs account for
variance in test performance?” (p. 282). Construct validity was presented as an alternative
type of validity that could be used when the attribute being investigated did not have a
criterion and was not operationalizable. They underscored the need to specify what
psychological constructs or attributes accounted for test-takers’ performance in any test.
They regarded the analyses of the factors that determine the behavior under investigation
as fundamental for the process of any test validation. Furthermore, they described the
limitation of presenting criterion-related evidence as the only evidence for validating a
test, affirming that a given criterion could be extremely fallible or that test developers
could not possibly “present predictive validities for all possible criteria” (p. 298). For
example, if a language learner scored very low on speaking test A, the tester could claim
that the learner had poor speaking skills as proven by the high correlation coefficients
between speaking test A and a well-established speaking test B. However, another
criterion, such as the language teacher’s observation of the learner’s highly-proficient speaking abilities in the classroom could disprove this claim. Test anxiety or reluctance to talk to strangers in a test setting could have been other factors that might have affected the results of the test.

Although Cronbach and Meehl did not intend for this model of construct validity to be the only model of test validity, validity theorists expressed preference for this model over the content and criterion-based validity types, and by the end of the 1970s it was considered the whole of test validity (Norris, 2008). Certainly, this model of validity presented several advantages over the other validity models. First, it emphasized the need to specify, examine, and evaluate the proposed test interpretations. Cronbach and Meehl (1955) asserted that “[t]he proper goals in reporting construct validation are to make clear (a) what interpretation is proposed, (b) how adequately the writer believes this interpretation is substantiated, and (c) what evidence and reasoning lead him to this belief” (p. 297, italicized in the original). Hence, according to this model, the test itself was not validated but rather the proposed interpretation of test scores. Second, by emphasizing the need for a clear statement of proposed interpretations, the model also allowed for challenges to the proposed interpretations. That is, it took into account alternative test interpretations, which was not possible within the other validity models. Third, it advocated for more rigorous and extended research, as opposed to the previous models that mostly required a single empirical study (Kane, 2013).

Notwithstanding its major improvements from previous validity types, the construct validity model had several drawbacks. As mentioned earlier, the major limitation of the model was its reliance on sound theory. On the one hand, under what Cronbach (1988)
referred to as “strong program of construct validity,” constructs were defined based on strong theory, and empirical data were gathered to determine whether this theory was sound. This was at times problematic especially for areas of study that lacked strong theoretical basis, as in the case of language proficiency (see Bachman, 2007). On the other hand, Kane (2001, 2013) noted that within the “weak program of construct validity,” the definition of what constitutes as evidence was too loose, and thus any kind of evidence related to test scores was considered to be relevant to construct validity. This was the result of the vagueness of the model in providing guidance in terms of how validation should be conducted, that is, what type of evidence to collect and how much of it was necessary. Another major weakness of this model was its limited applicability to validating educational assessments, in particular when trying to answer questions related to the implications of test use and test interpretations that were not theory-driven (Norris, 2008).

4.3. Unified Model of Validity

Due to its capacity to address some of the major concerns related to the trinitary validity model, Cronbach and Meehl’s (1955) and later Cronbach’s (1971) construct validity model continued on as a highly influential model of validity. During the 1960s and 1970s, several attempts were made to develop this model as a more inclusive and general framework of validity (Kane, 2013). Like other validity theorists, Messick (1980), asserted that construct validity was “the unifying concept that integrate[d] criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships” (p. 1015). By the mid-1980s, the
conceptualization of the construct model had shifted from being one type of validity to becoming a general approach to validity. This model encompassed various methods of theory testing together with all other types of validity (content and criterion-based) as types of evidence for construct validation (Kane, 2001, 2006).

As a result, based on the construct model, Messick (1989) proposed a different framework of test validation that unified all notions of test validity into a general framework and placed special importance on the social and ethical consequences of score interpretations implied by test use. In his highly influential article on test validation, he introduced validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” and continued on to specify that “validity is an inductive summary of both existing evidence for and the potential consequences of score interpretation and use” (p. 13). Thus, Messick (1975, 1980, 1989), like Cronbach and Meehl (1955) and Cronbach (1971, 1980) had done already, emphasized that what is validated is not the test instrument but rather the interpretation of test scores, and that social consequences of the test use and test interpretations are also a crucial element in the validation framework. Messick (1989) summarized his model in a two-by-two matrix (see Figure 4) that included the factors of validity that he proposed as necessary. Broadly speaking, he depicted this model as consisting of two major dimensions, namely (a) the source justification, which he explained as being based on either evidence or consequence, and (b) the function or outcome of the testing, which consisted of either test interpretation or test use. As presented in Figure 4, test interpretations were supported by construct
validation, whereas test uses were supported by construct validation as well as the evidence for the relevance and/or utility of the test for an intended purpose or specific setting. Moreover, Messick explained that the consequential basis for test interpretations depended on the value implications of the construct label, of the theories and ideologies of the test interpretations (p. 20). Lastly, the consequential basis of test use consisted of the evaluation of potential or actual social consequences of using a test. Accordingly, each of the facets in the matrix were to be systematically investigated, placing particular importance on construct validity, but at the same time taking into account implications and social consequences as facets that would make the validity framework more comprehensive.

<table>
<thead>
<tr>
<th>EVIDENTIAL BASIS</th>
<th>TEST INTERPRETATION</th>
<th>CONSEQUENTIAL BASIS</th>
<th>TEST USE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Construct validity</td>
<td>Value implications</td>
<td>Social consequences</td>
</tr>
<tr>
<td></td>
<td>Construct validity + Relevance/utility</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 4. Messick’s (1989, p. 20) matrix of the facets of validity*

The comprehensiveness of this unified model of validity and its applicability to educational measurement clearly appealed to the testing community. Unlike previous models, it placed explicit emphasis on proposed test interpretations and, to some extent, it took into account the intended uses of the test when related to the construct being interpreted. At the same time, as it underscored the role of test consequences, it made the validation process itself purposeful. In other words, validation was no longer a means to answer a simple empirical question that focused solely on the accuracy of measurements,
but rather an inquiry that also included considerations of the societal values of test interpretations. Hence, this model positively influenced the way test developers and test users viewed tests and test validation.

As with other validity models, however, several limitations can be associated to Messick’s (1989) unified model of validity. Even though it was all-encompassing in terms of the different notions of validity, it was also open-ended. As Kane (2013) has noted, one of the main drawbacks of the model was that it was unclear about “where to start and when to stop” (p. 450). That is, the model provided no clear guidelines as to how to carry out validation in practice, particularly in regard to “how much evidence was needed to adequately support a validity claim” (Kane, 2006, p. 22), or “which validity questions [were] essential to support a test use” (Sheppard, 1993, p. 427). The infinite scope of the model and the way it was structured—with social consequences as the last cell in the matrix to be considered for validation—was misleading in regard to the importance of test use and consequences, which many (e.g., Sheppard, 1993, 1997) argued should be a key component of test validation. In addition, the model was unable to address questions relevant to the validation of alternative assessments (i.e., performance or classroom-oriented assessments), as well as those concerning the consequential facet of validity, especially for educational assessments (Moss, 1992; Norris, 2008).

To recapitulate, on the one hand, Messick’s unified model was valuable for its ability to conceptually encompass all major facets of validation into a general framework. Building on the construct validity model, it also underscored the importance of societal values and consequences of test score interpretation and test uses. On the other hand, it lacked specificity in guiding the implementation of such validity framework, especially
for educational assessments. While it did propose incorporating the purported test interpretation and test uses into the validation process, it did not clarify the way the social consequences of test uses should be addressed. Furthermore, as Sheppard (1993) suggested, it regarded construct validity as the core component of test validation leaving the interpretation of test scores and their uses as a secondary component of the framework.

4.4. Argument-Based Approach to Validity

Building on previous validity models and improving upon some of their limitations, Kane (1992, 2001) proposed the argument-based approach to validity as an alternative validity model. This model of validation shifted the focus from the reliance on formal theories to specific statements of proposed interpretations and test uses (Chapelle et al., 2010; Kane, 2006). Unlike Messick’s (1989) unified framework of validity, the consequences of the test interpretation and test use were taken as crucial to the whole validation process (Bachman & Palmer, 2010). In other words, as Bachman (2005) noted, this model was useful in providing a “logical linkage between validity and test use” (p. 7). Moreover, this approach to validation sought to primarily address issues of practicality. As described in the previous section, the unified framework of validity incited discontent among the testing research community mainly due to its lack of specific guidelines to carry out the validation process. By contrast, Kane’s validity framework provided a coherent program that specified a step-by-step process of making claims and supporting them with appropriate evidence. The requirement of this model, to clearly articulate the purported assessment uses and interpretations allowed assessment
stakeholders (i.e., assessment users and assessment developers) to better understand the rationale behind test score interpretations, to carefully examine what evidence is necessary to support such interpretations, and to evaluate the decisions made based on assessment results.

The argument-based approach to validity can be broadly explained in terms of two different kinds of argument, namely, the interpretive argument (also referred to as interpretive/use argument or IUA in Kane, 2013) and the validity argument. Within the interpretive argument, the proposed interpretations and uses of an assessment are specified through “a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the assessment scores” (Kane, 2011, p. 8). In other words, the interpretive argument tries to elucidate the reasoning behind drawing conclusions and making decisions based on the assessment results. The validity argument, on the other hand, evaluates whether the interpretive argument is coherent and plausible in the way that it specifies and lays out inferences and assumptions. Kane (2006, 2011) provided three main criteria to evaluate the interpretive argument including clarity of the argument, coherence of the argument, and plausibility of inferences and assumptions.

4.4.1. Underlying inference model

In order to analyze the arguments within the validity model, Kane (2006, 2011, 2013) followed the suggestions of other measurement theorists, such as Mislevy (1996) and Mislevy, Steinberg, and Almond (2003), and advocated for the use of Toulmin’s (1958), later updated in Toulmin (2003), general framework of practical reasoning. In Toulmin’s framework, arguments are defined as being composed of data which lead to a
claim, and warrants and rebuttals that respectively support the claim or reject it. Figure 5 illustrates Toulmin’s model of inference with an example. As illustrated in the example, inferences within an argument start with data (e.g., Jörn goes to the coffee shop every day), which are later used to make claims (e.g., Jörn likes coffee). That is, the inferences are the process of going from observations to claims. This inference structure also includes qualifiers (e.g., presumably), which indicate the strength of the claims. The claims, in turn, have to be justified by warrants (e.g., People usually go to coffee shops because they like coffee), which are statements of “a law, generally held principle, rule of thumb, or established procedure” (Chapelle et al., 2010, p. 7). These warrants can be supported by evidence, which is what Toulmin referred to as backing (e.g., a hypothetical National Coffee Shop Association survey could show that 99% of coffee shop patrons like coffee). The evidence to support warrants could come from theory, previous research, or from research specifically oriented towards validation (Bachman, 2005). Moreover, claims can also be rejected through rebuttals (e.g., Jörn buys ginger tea not coffee), which refer to conditions where the warrant would not apply, and these rebuttals are in turn also backed by evidence (e.g., Jörn does not like coffee).

**Figure 5.** Example of Toulmin’s (2003) model of inference
In the same manner, this structure can be applied to every inference made within Kane’s interpretive argument chain. As will be explained in detailed in the next subsection, the interpretive chain is composed of several arguments which consist of observations leading to claims. Each claim serves as the grounds (or data) for the next claim in the chain (Chapelle et al., 2010), and each of these claims should be supported by evidence.

### 4.4.2. Describing the argument structure

As stated earlier, the argument-based model requires a clear articulation of the claims and interpretations related to the intended uses of the test. Kane (2006, 2011, 2013) proposed organizing these claims and interpretations in a sequenced network of inferences that would include the observations supported by warrants and/or rebuttals, which could lead to accepting or rejecting the claims. Kane, Crooks, and Cohen (1999) applied an analogy of a bridge (see Figure 6) to describe the argument structure, which consists of links or inferences that go from a test-taker’s performance in a test (i.e., observations) to the interpretations made from those test results (i.e., target scores). They noted that “a failure of any one of the inferences undermines the plausibility of the interpretive argument as a whole” (p. 9). In other words, if one of the bridges (i.e., inferences) would break down, then the whole argument of using test scores for a particular test use would become questionable.

As represented through the first three bridges in Figure 6, Kane et al (1999) explained that the crucial inferences in the interpretive chain are *scoring*, *generalization*, and *extrapolation*. Kane (2006) described the first inference, *scoring*, as the link between test-takers’ observed performance and the observed score. The assumptions for this
inference are based on the reliable scoring procedure, including considerations of the appropriateness, accuracy, and consistency of the scoring criteria as well as the recruiting and training of raters. Bachman (2005) also included the consideration of test administration conditions as another assumption for this inference. The second inference, *generalization*, is described as the link from obtained scores to the universe score (i.e., the score that represents the general population). The assumptions, in this case, rely on the collection of a representative sample from the universe of generalization and thus is mainly concerned with statistical sampling. Finally, the third inference, *extrapolation*, consists of the connection between universe score to the target score (i.e., performance in real-life task). This inference is based on the assumption that the universe score is related to the target score and hence the performance elicited through the assessment is representative of the performance in real life.

![Bridge analogy for the interpretive argument structure](image)

*Figure 6. The bridge analogy for the interpretive argument structure Adapted from Kane et al. (1999) and Bachman (2005)*

Although Kane et al. (1999) referred to the abovementioned three inferences as “critical to the chain of inferences from the observed performance to the expected performance” (p. 9), in Kane (2013), they are more loosely described as the most commonly found inferences in the interpretive argument with the possibility of including
“other kinds of inferences, depending on what is being claimed by the interpretation and use of the test scores” (p. 454). Several other types of inferences have been added into the argument structure. For example, as observed in Figure 6, utilization has been included as an additional inference of the interpretive chain. This inference has been incorporated by several validity theorists (e.g., Bachman, 2005; Chappelle et al., 2010; Purpura et al., 2015; Xi, 2008) as it connects the target scores to the decisions made based on these scores. That is, it links the interpretation of test scores to the test uses. It relies on several assumptions including (a) the assessment results (i.e., scores or other information) have been reported in ways that are useful for test users, (b) the decisions made based on test results are appropriate, and (c) there are no unintended negative consequences of the test (Xi, 2008, p. 188). In addition, Chapelle et al. (2008, 2010; see Figure 7) included six types of inferences, namely, domain description, evaluation, generalization, explanation, extrapolation, and utilization in their argument structure for validating the Test of English as a Foreign Language (TOEFL). They used a more inclusive term for the scoring inference, referring to it as evaluation. Furthermore, they added two more inference types: domain description, which links the target language use (TLU) domain to the observed test-taker’s performance, and explanation, which connects the expected scores to the underlying test constructs. The former inference type relies on the assumption that tasks in the TLU domain and the skills to complete such tasks successfully can be identified, and that the tasks are representative of the domain. The latter inference type rests on the assumption that the tasks included in the test together with the skills required to complete them conform with theoretical expectations. The resulting argument structure is represented in Figure 7.
Along similar lines, Purpura et al. (2015) proposed an interpretive argument structure that included the same six types of inference as those included in Chapelle et al. (2008, 2010). This work is particularly relevant for the purposes of the present dissertation study, as it presented an example of a validation argument for L2 research measures. It used Révész’s (2012) study on working memory and the effect of recasts to represent the way Kane’s (2006, 2011, 2013) validity model can be applied for validating uses and interpretations of different L2 measures in applied linguistics. In her study, Révész investigated (a) the influence of the type of outcome measure on the effect of recasts and (b) the relationship between working memory capacity and the effect of recast on learning. Her research design included three measurement tasks, namely a GJT, a written picture description task, and two oral production tasks. Purpura et al. (2015) walked readers through the earlier validation models (e.g., content validity, criterion
validity, construct validity, Messick’s unified model) when implemented for purposes of validating the three measures used in this specific study. Most importantly, these authors also elucidated the process of applying Kane’s (2006, 2011, 2013) model for the validation of the L2 measures. They gave a detailed explanation of each of the different inferences and claims within the proposed interpretive argument chain (see Figure 7 above). For example, for the generalization inference, which links test-takers’ observed scores to the expected scores (i.e., the scores that are assumed to be consistent across different measurement conditions), they suggested that the underlying claim was that the scores on a measure were consistent across test forms and parallel tasks, as well as across raters. Furthermore, they pointed out that Révész convincingly presented empirical backing for this inference through the analyses of tasks (e.g., task difficulty) using Multifacet Rasch Analysis (MFRA) and raters’ performance by examining the inter-coder agreement between two raters. The authors presented the different claims for all of the inferences specified in the interpretive chain, and explained how and what type of evidence Révész had provided in her study.

In sum, the argument structure can be characterized as a network of interconnected inferences which go from claims regarding the representation of the test and the TLU domain to the decisions made based on the test results. The chain-like characteristic of the model makes the analyses and backing of every inference crucial, as failure at any one of the inferences can compromise the whole validity argument. The type and number of inferences identified for a specific assessment depends on the different claims made about the interpretation of test results and its intended uses. These inferences in the argument chain should be clearly stated with appropriate and sufficient
empirical evidence backing each of them. Overall, Kane (1992) summarized the validation process in four steps, which also help define the argument structure: “(a) [decide] on the statements and decisions to be based on the test scores, (b) [specify] the inferences and assumptions leading from the test scores to these statements and decisions, (c) [identify] potential competing interpretations, and (d) [seek] evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments” (p. 527).

4.4.3. Discussing the advantages of the model

As noted earlier, the argument-based approach to validity has provided an alternative model to test validation that has improved upon some of the earlier frameworks, and thus multiple advantages can be associated with the model. Bachman (2005), for example, pointed out that the clear articulation of the validity argument encourages researchers to think more carefully about the different claims that they make and the types of evidence that are needed to support those claims. Therefore, this framework offered a practical solution for validating the use of shortcut measures for research purposes, such as the Korean C-test, which was investigated in this dissertation. As described in detail in the next chapter, the primary intended use of these types of tests is to provide a reliable and quick measure of learners’ global proficiency. Kane’s framework allowed the careful examination of this intended use and other claims associated with the interpretations of the results of both tests, particularly in relation to the administration of these tests to different types of learners (i.e., foreign or heritage language learners).
Another major advantage of this validation model, and the one most relevant to this dissertation study, was that it lent itself to validation of not only educational assessments but also other types of assessments that are commonly used for SLA research. Purpura et al.’s (2015) example of the applicability of Kane’s validity model to L2 measures demonstrated that “the principles underlying validation apply to all L2 measures in applied linguistics, and not just to those in language assessment” (p. 40; emphasis in the original). As the authors noted, by following this validation approach, which advocates a clear statement of claims and interpretations of test scores/measures, SLA researchers are able to reach a better understanding of the rationales behind the different interpretations and uses of the L2 measures.

Kane (2006) also argued that a key advantage of this validation framework is that it provides clear guidance in terms of “allocating research effort and in gauging progress in the validation effort” (p. 23). According to Kane, defining the main inferences and assumptions leads researchers to focus on the most relevant validation questions and thus on gathering evidence that supports these key inquiries. There are, however, some drawbacks pertaining this claim. Norris (2008), for example, pointed out that there is still lack of specificity in terms of what are the important inquiries researchers should primarily focus on. In addition, although the model indeed underscores the need of sufficient and appropriate data for backing claims, it is unclear about what types of and how much evidence is needed. All in all, although the framework is comprehensive, it can also (still) be overwhelming, particularly to educational researchers, applied linguists, and SLA scholars for whom the evaluation of the uses of the L2 measures or instruments is only one step in the process of addressing a bigger research inquiry.
In the case of the C-test explored in the current dissertation study, Kane’s validity model indeed helped narrow down and better understand what questions needed to be answered for validating its intended uses. For example, the *generalization* inference was crucial for the present study, as one of the key questions related to whether this shortcut measure (i.e., C-test) could also be used for HLLs (see next chapter). Therefore, the model highlighted the importance of collecting evidence to support (or reject) the claim that this shortcut measure can be used for measuring both FLLs and HLLs. In consideration of Norris’ observation, the validation study pursued here was not intended to be a unique, standalone study, rather one initial investigation of the validity of this measure of global proficiency for the purposes of measuring both FLLs’ and HLLs’ general language proficiency for L2 research purposes.

Additionally, Kane’s validity framework was also useful in terms of theory-building as “the interpretative argument allows for inclusion of a theory-defined construct” (Chapelle, 2011, p. 24). For this reason, it was especially advantageous for examining constructs related to HLL’s language proficiency. The *explanation* inference, which explored the connection between expected scores and the theoretical expectation of language proficiency constructs, encouraged the careful examination of the underlying construct of HLLs’ language proficiency and how this shortcut measure could assess such construct. Building the argument for this specific inference allowed the identification of the specific claims that these tests intend to make in terms of the construct they attempt to assess, as well as the kind of evidence that is necessary to support those claims.

Finally, the process of building an interpretive argument also facilitated the identification of possible limitations of the assessments. The piecemeal nature of the
validity structure made it possible to reveal the potential issues of a given claim or the unattainability of certain evidence for backing. By implementing this validation model to evaluate the intended uses of the C-test, it made it easier to determine some of the inferences or claims in need of more empirical backing to support, or perhaps reject, certain claims.

On the whole, the argument-based approach to validity provided a useful framework to analyze the validity of claims and interpretations of the shortcut measure presented in this dissertation, namely the Korean C-test. As Kane (2001) suggested, it was only through the clear articulation of purported interpretations of the results of the shortcut measure and its intended use that validation was possible. While building a convincing argument for the validation of the shortcut measure may be complex, it allowed the development of an interpretive argument, which sought to (a) determine what kinds of questions are crucial for the validity argument, and (b) elucidate what types of backing are needed. In addition, it also helped build a validity argument that assessed whether certain claims could be supported or rejected.
Chapter 5

Interpretive Argument for the Korean C-test

As described in the previous chapter, Kane’s (2006, 2011, 2013) validity framework follows two main arguments, namely the interpretive argument and the validity argument, both of which are guided by the primary intended uses of the test. Kane (2004) explained that the interpretive argument is developed first, preferably as part of the test design, and then the validity argument is built empirically based on the interpretive argument, when the test has become operational. More specifically, Kane pointed out that the interpretive argument lays out the different claims, warrants, and inferences of the test, and indicates what types of evidence are needed to back these claims. Meanwhile, the validity argument provides and weighs the evidence to evaluate the coherence of the argument. This chapter focuses on the interpretive argument and begins by providing a detailed description of the context and main intended use for the Korean C-test. Then, it proposes an interpretive chain consisting of six different inferences, namely theoretical grounds, evaluation, generalization, explanation, extrapolation, and utilization. Each of these inferences is described in terms of warrants followed by assumptions from which evaluation questions emerge. It also specifies the evidence that is required to back those assumptions and explores possible rebuttals to some of the warrants.

It is important to note that the validation research conducted for this study focuses on test evaluation rather than test development. Unlike a large body of literature using the argument-based approach to validation at the early stages of test development (e.g., Chapelle et al., 2008; Drackert, 2016), the present study evaluates an already existing
Korean C-test (Son et al., 2018). Therefore, it critically assesses the coherence of the interpretive argument by weighing the evidence collected and considering possible alternative interpretations of the test results as well as possible rebuttals that could weaken the validity argument.

5.1. Korean C-test Context and Intended Uses

Referring back to Chapter 3, the new Korean C-test (Son et al., 2018) was designed to serve as a tool that provides a quick assessment of Korean language learners’ proficiency for research purposes. It was developed as one of a series of C-tests assessing different languages (see Norris, 2018). Moreover, it sought to improve upon some aspects of the previous Korean C-test (Lee-Ellis, 2009; see section 3.3.2) The differences in the design of the new C-test aimed at improving: (a) text selection, by choosing passages from a wider range of contexts; (b) test suitability for the assessment of Korean FLLs, by using a more general proficiency framework (i.e., TOPIK rather than the ILR); and (c) test accuracy, by modifying the deletion technique and scoring method. Item Response Theory (IRT) was used to analyze the items and test-taker performances, and a Rasch Analysis was conducted to select the best fitting items to be included in the C-test. Five polytomous items were identified as the most accurate and reliable in assessing five different levels of proficiency. Nevertheless, as these items were identified based on results of a relatively small sample size \(N = 38\), two more passages were selected for the current large-scale validation study. Accordingly, the Korean C-test investigated in this dissertation study was composed of seven polytomous items (or super-items), each containing 25 micro-items or blanks.
As briefly described in previous chapters, the primary intended use of the Korean C-test is to provide an estimation of Korean language learners’ general language proficiency for research purposes. Therefore, C-test scores should be interpreted specifically for this intended use as opposed to measuring language proficiency in a domain-specific context. In addition, there are two key components within the intended use description, namely, “language learners” and “general language proficiency.” It is crucial to define both components in further detail in the context of research in SLA and HLA.

As for the first aspect of the intended use description, the main demographic generally investigated in SLA research is the adult language learner from post-secondary institutions (mainly four-year universities). Liu and Brown (2015), for example, found that in studies of written corrective feedback, 75% of the studies surveyed ($N = 42$) were dedicated to exploring adult FLLs. Plonsky (2017) noted that this trend is also observed in other domains in SLA, such as TBLT (Plonsky & Kim, 2015). In the same vein, HLA research appears to have followed a similar trend, mostly concentrating on the adult-language-learner population primarily from institutions of higher-education. As a result, the present dissertation explored the use of the new Korean C-test as a test that mainly targets language learners, both FLLs and HLLs, who are adult language learners in institutions of higher education. In order to avoid the possible influence of instructional context on test-takers’ performance scores, this study restricted the context to universities across the United States.

In regard to the definition of “general language proficiency,” as Chapelle et al. (2008) have indicated, there is indeed “no single best definition of language proficiency”
However, in line with Hulstijn (2011), one good approach to defining proficiency involves considering not only the learners’ language abilities (i.e., knowledge of grammar, vocabulary, form-meaning connections) and linguistic processes and strategies, but also the context in which these language abilities are used (e.g., everyday activities, academic settings) through different modalities (i.e., listening, reading, speaking, and writing). For the purposes of this study, which sought to address the need of researchers to have a quick measurement of Korean learners’ general language proficiency, general language proficiency can be conceptualized as the ability of learners to communicate functionally, using their knowledge of grammar, vocabulary, and the relationship between form and meaning, in different contexts and through multiple modalities.

5.2. Inferences, Warrants, and Underlying Assumptions, and Possible Rebuttals

Based on the intended use stated above, the interpretive argument for the Korean C-test included six inferences, namely theoretical grounds, evaluation, generalization, explanation, extrapolation, and utilization. Each of these inferences included warrants and underlying assumptions that needed to be supported by empirical backing. As explained in Chapter 4, the inferences in the interpretive argument are structured in the form of a chain or series of connecting bridges, thus relevant and sufficient evidence should be provided in order to continue from one inference to the next. This section presents each of the inferences and warrants, and describes the necessary empirical backing to support (or reject) the assumptions underlying the warrants. In addition, it explores possible rebuttals to some of the warrants that could weaken the inferential link between the intermediate conclusions. Table 4 summarizes the Korean C-test interpretive
argument. It presents each inference with its respective warrant together with the assumptions associated with it. The table format follows previous studies that applied Kane’s validity argument, such as Chapelle et al. (2008, pp. 19-21) or Voss (2012).

Table 4
Summary of Inferences, Warrants, and Assumption in the Korean C-test Interpretive Argument

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical Grounds</td>
<td>Observations of the performance on the C-test reflect language learners’ global language proficiency as characterized by a large body of literature from C-test experts</td>
<td>1. The C-test format has been subjected to sufficient research to provide a trustworthy foundation for believing that it is a good measure of global language proficiency.</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Observations of the performance on the Korean C-test can be evaluated to provide scores that reflect general ability of the Korean language</td>
<td>1. The C-test text selection procedure is appropriate for selecting texts that can distribute learners of varying degrees of language abilities. 2. The statistical characteristics of items in the C-test are appropriate for norm-referenced decisions. 3. Items on the C-test provide an accurate measurement for distinguishing test-takers with a wide range of language abilities.</td>
</tr>
<tr>
<td>Generalization</td>
<td>Observed scores are estimates of expected scores that consistently reflect Korean learners’ (both HLLs and FLLs) language proficiency across items on the C-test</td>
<td>1. The items on the C-test can provide reliable and consistent estimates of test-takers’ performance 2. The configuration of items on the C-test is the most appropriate for providing reliable estimates of test-takers’ performances. 3. The number of tasks on the C-test is sufficient to provide stable estimates of test-takers’ performances. 4. The observed scores on the C-test consistently reflect the language abilities of both Korean heritage and foreign language learners. 5. The C-test text selection and deletion method are sufficiently detailed and consistent to produce equivalent test forms.</td>
</tr>
</tbody>
</table>
As the name indicates, shortcut measures, such as the C-test, are assessment tools that seek to provide a “shortcut” to a measurement of global language proficiency. In other words, rather than using a long battery of tests, these measures include fewer items.

Table 4 (continuation)
Summary of Inferences, Warrants, and Assumption in the Korean C-test Interpretive Argument

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant</th>
<th>Assumptions</th>
</tr>
</thead>
</table>
| **Explanation** | Expected scores are attributed to a global construct of Korean language proficiency | 1. The difficulty of C-test items is systematically influenced by item characteristics regardless of the learner type.  
2. Performance on the Korean C-test relate to performance on other literacy-based measures of language proficiency as expected theoretically.  
3. Performance on the C-test can be predicted by speaking and writing proficiency. |
| **Extrapolation** | The construct of Korean global language proficiency as assessed by the Korean C-test is related to other criteria of language proficiency that are reflective of HLLs’ and FLLs’ general language proficiency | 1. Test-takers’ performance on the C-test has a positive relationship with their performance on other measures of language proficiency (i.e., speaking, writing, and oracy skills).  
2. C-test performance can be interpreted in terms of other criterion measures generally used in the Korean language learning context.  
3. C-test performance has a positive relationship with learners’ self-ratings of their Korean language proficiency. |
| **Utilization** | C-test performance results can be used as a measure of Korean learners’ general language proficiency for applied linguistics research that seeks to examine both Korean HLLs and FLLs. | 1. The C-test scores/results are easily interpretable by applied linguistics researchers.  
2. The C-test is useful for applied linguistics research where language proficiency is a predictor or moderating variable.  
3. The C-test is useful for research focusing on both HLLs and FLLs or on one of the learner groups.  
4. The C-test helps to better understand research findings across multiple studies on Korean language learners. |
that can quickly but reliably assess a more general construct of language proficiency. The holistic nature of the Korean C-test construct (and of other shortcut measures constructs) sets it apart from other types of tests, such as English for Specific Purposes (ESP), English for Academic Purpose (EAP), or achievement tests, as it is purposefully not a domain-referenced test. Norris (2018) cautions that C-test interpretations “should remain limited to general estimations, and not extend erroneously to assumptions about skill-, domain-, or task-specific abilities to use the language” (p. 8). For this reason, although uncommon in previous validity studies implementing Kane’s (2006, 2011, 2013) validity model, which usually include “domain description” as their first inference in the interpretive argument, the present study included theoretical grounds as its first inference. At this first stage of the interpretive argument it was considered important to expand on the theoretical perspective of what C-tests are supposed to measure (or not measure) before moving into more empirically-based inferences. This inference, then, was based on the warrant that performance on the C-test reflects language learners’ global language proficiency as characterized by a large body of literature from C-test experts. Thus, this warrant was associated with one underlying assumption:

1. The C-test format has been subjected to sufficient research to provide a trustworthy foundation for believing that it is a good measure of global language proficiency.

The empirical backing was derived from the collection of findings from previous studies and the theoretical characterizations of C-test constructs.

The second inference, evaluation, was based on the warrant that the observations of performance on the Korean C-test can be evaluated to provide scores that
reflect Korean language learners’ general language proficiency. In other words, this inference connected the observed performance to the observed scores, and as Clauser, Kane, and Swanson (2002) described, it was mainly concerned about the accuracy of the implementation of the test. This warrant was, then, based on several underlying assumptions, including:

1. The C-test text selection procedure is appropriate for selecting texts that can distribute learners of varying degrees of language abilities.

2. The statistical characteristics of items on the C-test are appropriate for norm-referenced decisions.

3. Items on the C-test provide an accurate measurement for distinguishing test-takers with a wide range of language abilities.

The data for backing such assumptions were gathered from evaluating the C-test development procedure, particularly in terms of text selection, and how this process facilitated the appropriate assessment of Korean learners from a wide range of language abilities. Moreover, descriptive statistics also served as empirical backing for the second assumption. Finally, for the third assumption, item analyses through Item Response Theory (IRT) provided information about how the items behaved in terms of difficulty and consistency, and thus their accuracy in distinguishing learners of varying language proficiency levels. One possible rebuttal to this warrant and its assumptions was that the learners recruited for this study did not represent a learner sample with a wide characteristics range of proficiency levels.

The third inference, **generalization**, connected observed scores on the Korean C-test with expected scores. More specifically, this inference was based on the warrant that
observed scores are estimates of expected scores that consistently reflect Korean learners’ (both HLLs and FLLs) language proficiency across items on the C-test. Most relevant to the present study, this warrant also included the reliability of tasks and measures when examining all learners as well as heritage and foreign language learners separately. The underlying assumptions included:

1. The items on the C-test can provide reliable and consistent estimates of test-takers’ performance.
2. The configuration of items on the C-test is the most appropriate for providing reliable estimates of test-takers’ performances.
3. The number of tasks on the C-test is sufficient to provide stable estimates of test-takers’ performances.
4. The observed scores on the C-test consistently reflect the language abilities of both Korean heritage and foreign language learners.
5. The C-test text selection and deletion method are sufficiently detailed and consistent to produce equivalent test forms.

The empirical backing came from reliability measures of the items and item/person fit analysis through IRT models. The IRT model was especially useful as it could map test-takers’ language abilities and item difficulties on a single continuous latent variable, and thus helped assess whether the configuration of items on the tests was appropriate and also whether there were items missing for a particular learner ability level. Moreover, a Differential Item Functioning (DIF) analysis examined whether the C-test items were functioning the same way for both HLLs and FLLs, or whether a noticeable bias could be detected for a certain learner group. A rebuttal for this warrant was that the a priori
definition of HLLs and FLLs was not appropriate and needed further analyses. Therefore, data collected from the background questionnaire were analyzed to explore other possibilities for grouping learners.

The fourth inference, explanation, was the link between expected scores and the theoretical construct of general language proficiency. This inference was based on the warrant that expected scores are attributed to a global construct of Korean language proficiency. Three underlying assumptions for this warrant included:

1. The difficulty of C-test items is systematically influenced by item characteristics regardless of the learner type.
2. Performance on the Korean C-test relates to performance on other literacy-based measures of language proficiency as expected theoretically.
3. Performance on the C-test can be predicted by speaking and writing proficiency.

Empirical backing consisted of examining the item characteristics that affect item difficulty, for example in terms of the lexical and grammatical features of the micro-items in each C-test passage. In addition, correlational analyses provided evidence of the relationship between C-test performance and other literacy-based measure, namely the ACTFL WPT results. Another piece of evidence came from the results of multiple regression analyses that explored the extent to which speaking and writing proficiency could help predict test-takers’ performance on the C-test. If the C-test is useful in providing a global measurement of language proficiency, then both speaking and writing abilities should be able to predict C-test performance. One possible rebuttal to this warrant was the use of ACTFL tests to assess HLLs’ speaking and writing proficiency. Some scholars have pointed out that, given that these tests have not been designed
specifically for testing HLLs’ language ability, they should not be used for this purpose. This rebuttal related to the controversy about the appropriate use of certain standardized tests, such as the ACTFL tests for testing HLLs (see Chapter 2 section 2.5.2).

The fifth inference in the Korean C-test interpretive argument, extrapolation, entailed the warrant that the construct of Korean global language proficiency as assessed by the Korean C-test is related to other criteria of language proficiency that are reflective of general language proficiency. Thus, the underlying assumptions of this warrant were:

1. Test-takers’ performance on the C-test has a positive relationship with their performance on other measures of language proficiency (i.e., speaking, writing, and oracy skills).

2. C-test performance can be interpreted in terms of other criterion measures generally used in the Korean language learning context.

3. C-test performance has a positive relationship with learners’ self-ratings of their Korean language proficiency.

Empirical data to support these assumptions came from criterion-related validity studies, which explored the relationship between C-test scores and the scores obtained from multiple criterion measures, such as standardized tests of speaking and writing, as well as a measure of oracy (see next chapter for detail information on these measures). In addition, C-test performance was compared to performances on ACTFL tests in order to facilitate the interpretation of C-test scores. Finally, another source of empirical backing for this inference derived from correlational analysis between self-assessment scores and overall C-test scores.
The final inference, **utilization**, related to the connection between test scores and the intended uses of the test. Thus, it was based on the warrant that the C-test results can be used as a measure of Korean learners’ general language proficiency for applied linguistics research that seeks to examine both Korean HLLs and FLLs. It relied on several assumptions including:

1. The C-test scores/results are easily interpretable by applied linguistics researchers.
2. The C-test is useful for applied linguistics research where language proficiency is a predictor or moderating variable.
3. The C-test is useful for research focusing on both HLLs and FLLs or on one of the learner groups.
4. The C-test helps to better understand research findings across multiple studies on Korean language learners.

As the Korean C-test has yet to be deployed operationally for research on Korean language learners, the collection of empirical backing for this inference fell outside of the scope of the present dissertation study. For future studies, however, it would be useful to collect data that support such assumptions by gathering information on the efforts undertaken to make the Korean C-test publicly available and its results easily interpretable for any researcher who wants to use it. Furthermore, washback studies can also provide evidence for beneficial consequences of using such tools across multiple studies.

### 5.3. Research Questions

Following the description of the assumptions for each of the warrants at every stage of the interpretive argument chain, research questions were formulated in order to
guide the process of obtaining empirical backing. It is worth noting that, although all inferences and their assumptions were evaluated, the present study focused mainly on the evaluation, generalization, explanation, and extrapolation inferences, and thus this section summarizes the research questions that emerged from the main assumptions for these four inferences.

**Evaluation inference**

1. To what extent is the C-test text selection procedure during the test development stage appropriate for selecting items that can distribute learners of varying degrees of language abilities?
2. What are the statistical characteristics of the C-test items and measures?
3. To what extent do items on the C-test accurately assess test-takers and distinguish their wide range of ability levels?

**Generalization inference**

1. To what extent does the Korean C-test produce scores with acceptable reliability?
2. What configuration of items is the most appropriate for providing reliable estimates of performances from learners with varying degrees of language abilities?
3. Does the C-test include a sufficient number of items to provide stable estimates of performances from learners with varying degrees of language abilities?
4. Do observed scores on the C-test consistently reflect the Korean language abilities of Korean learners regardless of whether they are heritage or foreign language learners?
Explanation inference

1. What characteristics of the C-test items appear to influence the difficulty of the items?

2. Do these item characteristics systematically influence the difficulty of the items regardless of learner type?

3. How do the test-takers’ performances on the Korean C-test relate to their performance on other literacy-based measures?

4. To what extent can speaking and writing language skills predict performance on the C-test?

Extrapolation inference

1. How do the test-takers’ performance on the C-test relate to criterion measures of oracy-based language proficiency?

2. How do C-test scores relate to criterion measures generally used in the Korean language learning context?

3. How do the test-takers’ performances on the C-test relate to their self-assessment ratings?
Chapter 6

Methodology

A large-scale validation study was conducted to answer the research questions stated in the previous chapter. This chapter describes the methodology used to collect empirical backing for the assumptions in the aforementioned interpretive argument, including a description of participants, instruments, and the procedures of administering a battery of Korean language tests.

6.1. Participants

A total of 93 Korean language learners, 41 heritage and 52 foreign language learners from a wide range of proficiency levels, were recruited for the purposes of this study in the span of two semesters, Spring 2017 and Fall 2017. In keeping with the definition of HLL adopted for the purposes of this dissertation, Korean learners who had at least some exposure to Korean language at home during their childhood and at least some level of proficiency in Korean were initially classified as HLLs. In order to avoid recruiting HLLs who had spent almost all of their formative years in Korea, another important criterion used when recruiting participants was that they had to be enrolled in a Korean language course. Generally, HLLs enrolled in Korean language courses are

---

2 Initially, HLLs who had been exposed to several years of formal instruction in Korea were considered not to fit the profile of learners being investigated in the current study because it was observed during the pilot-test in Son et al. (2018) that Korean learners who had formal instruction in Korea for more than four years behaved more similar to NSs of Korean than to other advanced learners of Korean. The C-test had been unable to provide useful information about these particular learners, as they showed no variation in their performance on the C-test. A ceiling effect was observed for these particular learners. Nevertheless, this criterion was reassessed when considering the purported uses of this C-test. That is, these learners are commonly included in the studies on Korean language learners as they are enrolled in language courses, and thus the C-test should also be able to assess their proficiency and provide a measurement of their general language proficiency.
seeking to learn Korean in a formal context because they have had little or no experience with the language in such a context. It was also crucial to try to recruit a balanced number of Korean language learners across varying degrees of language ability. Previous research investigating the differences in Korean learners’ language skills have pointed out that a true comparison of performances between proficiency levels (e.g., intermediate vs. advanced) and learner types (HLLs vs. FLLs) is challenging as it is difficult to find advanced levels of Korean learners, particularly advanced FLLs. For this reason, similar number of participants were recruited from four different curricular levels for both the HLL- and the FLL group. Although curricular levels are not a reliable indicator of language proficiency, it was considered an appropriate strategy for the first stage of participant recruitment.

Participants were recruited from Korean language programs across 10 different universities in the United States. As mentioned earlier, the context was restricted to the United States to control for the possible influence of instructional context on the results. All learners were either taking Korean language courses at the moment of data collection or had taken Korean language courses during the semester prior to data collection. Furthermore, all learners had taken at least one year of Korean courses. The learner sample consisted of a total of 72 female and 21 male students. Participants’ ages ranged from 17 to 35 years old. In terms of their educational level, 89% of the participants (N = 83) were undergraduate students and 9% (N = 8) of them were graduate students enrolled in a Master’s or PhD program. The remaining 2% (N = 2) of the participants had graduated with a bachelor’s degree the semester before data collection started. Table 5
summarizes the demographics of the participants based on learner type, gender, educational level, and instructional level.

Table 5

**Participant Demographics**

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Age Range</th>
<th>Educational Level</th>
<th>Instructional Level&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Undergrad</td>
<td>Grad</td>
</tr>
<tr>
<td>HLLs</td>
<td>31</td>
<td>10</td>
<td>18-35</td>
<td>37</td>
</tr>
<tr>
<td>FLLs</td>
<td>41</td>
<td>11</td>
<td>17-29</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>21</td>
<td>17-35</td>
<td>83</td>
</tr>
</tbody>
</table>

Notes. <sup>a</sup>Instructional level was based on the curricular level of the Korean language course (i.e., first, second, third, or fourth year level Korean course as described by Korean instructors or Websites of each program)

The majority of participants (66%, N = 61) were born in the United States, 64% of them to families with no Korean heritage, and 36% of them to families where one or both parents were Korean. Nineteen participants (or 20% of the total sample) indicated being born in Korea. Only 5 of them (or 26% of the Korean-born) had received elementary school education for between 1 and 6 years and only 1 had been schooled in Korea up until middle school, whereas most of the Korea-born participants (68%, N = 13) had left the country prior to any formal schooling. The Korean-born heritage participants had a length of residence in the US that ranged from 3.5 to 25 years, and they had arrived in the US at an age that ranged between 1 month and 19 years. There was also a minority of participants (14%, N = 13) who were born neither in the US nor in Korea, but in other countries such as China, Turkey, and Singapore. They had been born in a family with no Korean heritage. Their length of residence in the US ranged from 6 months to 19 years and their age of arrival in the US was between 2 and 18 years old. Interestingly, from the US-born learners, 8 (or 13% of the US-born) indicated being FLLs of Korean who were
heritage learners of other languages, with a variety of L1s including Tamil, Mandarin, Vietnamese, Spanish, and Ejagham.

It is also worth noting that the place of birth or the first language learned did not always define the language participants were most comfortable using. As represented in Figure 8, while 41% and 26% of the participants indicated English and Korean as their first language learned, respectively, the great majority of them (83%, \( N = 77 \)) chose English as their only most comfortable language and only two (or 2%) chose Korean as their most comfortable language, both Korea-born and schooled in Korea until the age of 7 and 15 respectively. Similarly, there was a decrease in the number of learners who reported both English and Korean as first languages learned simultaneously (\( N = 11 \)) to having both languages as their most comfortable ones (\( N = 2 \)). This is in line with previous literature (e.g., Kim, Montrul, & Yoon, 2009; Valdés, 2001, 2005) that describes how HLLs gradually shift from learning and using the HL to using English as their most dominant (i.e., comfortable) language. This transition appeared to also be present in other participants (\( N = 8 \)) who were FLLs of Korean and heritage learners of languages other than Korean, as observed through the larger variety of languages reported as first languages learned (e.g., Ejagham, Karachay, Spanish, Tamil, Turkish, Vietnamese mostly in combination with English) when compared to the variety of those reported as their most comfortable languages. Furthermore, Chinese was the second most comfortable language reported with 10% (\( N = 9 \)) of participants choosing it. There were also three students who indicated English and another language (either Spanish or Turkish) as being both most comfortable.
Figure 8. Comparison between participants’ first and most comfortable language

6.2. Instruments

Five assessment instruments were used in the current study, namely, the (1) Korean C-test (Son et al., 2018); (2) Korean Elicited Imitation Test (Kim et al., 2016), (3) ACTFL Korean Oral Proficiency Interview-computer (OPIc), (4) ACTFL Korean Writing Proficiency Test (WPT), and (5) a background questionnaire which also included a self-assessment component. The test instruments (2) through (5) were administered to serve as criterion measures for the Korean C-test. They were primarily useful for gathering empirical backing for the explanation and extrapolation inferences in the interpretive argument.
6.2.1. **Korean C-test (Son, Kim, Cho, & Davis, 2018)**

As the development and characteristics of the Korean C-test (Son et al., 2018) have already been described in detail in previous chapters (see sections 3.3.2 and 5.1), this section mainly focuses on its scoring and administration procedures. As a reminder, the C-test was composed of several passages (or items) a paragraph in length, which contained 25 deleted words (or micro-items) each. The seven best fitting items from the 10-item Korean C-test in Son et al. (2018) were chosen to be administered for the purpose of this study. Each of the seven items contained 25 micro-items or blanks, which were scored dichotomously (i.e., 1 point for right answers and 0 points for wrong answers). No partial credit was given to partially completed deletions (see Table 6).

Some items also allowed an alternative answer determined by Korean NS answers gathered during the NS pilot as described in section 3.3.2. The maximum possible score for each super-item was 25, and that of the whole C-test was 175. Appendix A presents all seven texts included in the Korean C-test with their answer keys. In terms of test administration, participants were given 35 minutes to complete the test and were instructed to try to spend no more than 5 minutes per text so that they could reach the very last item. The test was administered in paper-based format to avoid disadvantaging learners unfamiliar with the Korean keyboard. It was assumed that the majority of learners were more familiar and comfortable with the paper-based format than the Korean keyboard, as in most of the Korean language programs assignments are submitted by hand and Korean keyboards are not easily accessible in the US.
Table 6
Scoring Method and Examples

<table>
<thead>
<tr>
<th>Target response</th>
<th>Sample response</th>
<th>Score</th>
<th>Error type</th>
</tr>
</thead>
<tbody>
<tr>
<td>돈도 그라 요.</td>
<td>돈도 그라 요.</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>돈도 그라 __.</td>
<td>돈도 그라 __.</td>
<td>0</td>
<td>Missing particle</td>
</tr>
<tr>
<td>돈도 그 __ 요.</td>
<td>돈도 그 __ 요.</td>
<td>0</td>
<td>Missing lexicon</td>
</tr>
<tr>
<td>돈도 그레 요.</td>
<td>돈도 그레 요.</td>
<td>0</td>
<td>Incorrect spelling</td>
</tr>
</tbody>
</table>

In order to ensure a consistent and accurate scoring process, the answers were manually key-entered into a Microsoft Excel spreadsheet which was connected to another spreadsheet that would automatically score each answer as correct or incorrect (i.e., 1 or 0) based on the answer key provided. The manual input of the responses for each participant was triple-checked for accuracy.

6.2.2. Korean elicited imitation test (Kim, Tracy-Ventura, & Jung, 2016)

Elicited Imitation Tests (EITs), also referred to as sentence repetition tests, are another type of shortcut measure that also are presumed to offer a good estimate of learners’ global language proficiency (e.g., Jessop et al., 2007; Vinther, 2002; Yan et al., 2015). In these tests, test-takers listen to a sentence and then repeat it as closely as possible to the original stimulus for a full series of up to about 30 sentences with either varying lengths or a certain constant length. Like C-tests, EITs require test-takers to understand and decode the overall meaning of the sentences, whose length exceeds their working memory capacity, in order to be able to reconstruct and repeat them. That is, the successful completion of the task depends on test-takers’ understanding of the meaning of the sentences and their ability to use implicit linguistic knowledge to reconstruct them (e.g., Erlam, 2006; Wu & Ortega, 2013; Zhou, 2012).
Due to their practicality in design, administration, and scoring, EITs have often been used in the L2 research as a shortcut measure of general language proficiency. Even though this type of test has been criticized for its lack of authenticity and resemblance to real-life communicative tasks, a considerable amount of research has indicated that EITs have high internal consistency and high reliability in distributing learners across different proficiency levels (e.g., Tracey-Ventura et al., 2014; Ortega, 2000; Bowden, 2016). Yan et al.’s (2015) meta-analysis indicated that the EITs are effective in distinguishing speakers of lower and higher proficiency levels with an overall weighted average effect size of 1.34. In other words, this meta-analytic study indicated that the EITs were able to consistently differentiate the performance of learners of low- and high-proficiency with a very large effect size.

The Korean Elicited Imitation Test administered in this study was developed by Kim et al. (2016) as one of the iterations of EITs originally designed by Ortega, Iwashita, Rabie, and Norris (1999). The test was composed of 30 Korean sentences with increasing lengths (between 7 and 17 syllables, typically) and difficulty levels. Kim et al. (2016) reported that the test items showed satisfactory reliability estimates, with a Cronbach’s alpha of .96, and that all items, except for the easiest one (item 1) showed item discrimination levels within the acceptable ranges as determined by Ebel (1979). Moreover, test-takers’ performances on the EIT had a positive and moderate relationship with other measures of proficiency, such as a speaking test \( (r = .77) \), a listening test \( (r = .62) \), and the fluency measures of the speaking test \( (r = .62) \). No significant correlations were found between the EIT measures and Phonological Short-Term Memory (PSTM).
measures, which meant that PSTM did not play a significant role in test-takers’
performance on the EIT.

The scoring of the Korean EIT for the current study followed the same rubric
reported in Kim et al. (2016). Two raters, a Korean native speaker and the researcher,
rated all of the speech samples. Even though Kim and colleagues offered a detail
description of the scoring rubric, there were certain features of responses for each original
stimulus that needed further discussion in order for the ratings to reach to higher
agreement rates. Therefore, the following steps were taken as rater training sessions.
First, a small portion of participants (N = 8; 4 HLLs and 4 FLLs) were randomly selected
from each curricular level and their EIT performances were rated by both raters while
discussing the rubric. At this first stage, raters were getting themselves familiarized not
only with the rubric but also with some of the features in each sentence that appeared to
be most important when scoring. This step helped develop a more detailed version of the
rubric, which included examples and more specific features for some items (see
Appendix B). Second, a stratified sample of participants by learner type and curricular
level (N = 14; 6 HLLs and 8 FLLs) was chosen and each rater rated the speech samples
individually. At this stage, the exact agreement rate between raters was of 76%. Third,
raters discussed the responses that did not reach perfect agreement and they reached a
consensus on a single score. Fourth, a matrix of responses by items and scores was
developed (see Appendix C). With this matrix in hand, raters scored another 15% of
participants (N = 14), also selected through stratified sampling. This time, the agreement
rate increased to 83%. Fifth, raters discussed once again the responses that were not in
perfect agreement and reached consensus. More examples of responses were added to the
matrix in order to show at least one example per score possible for each of the 30 items. Finally, the raters rated all the remaining participants individually ($N = 57$) and the exact agreement rate remained at 83%. Raters discussed again to reach consensus on final scores for the responses that were not in perfect agreement.

It is worth noting that the matrix of sample responses brought additional clarity to the scoring process. It helped raters discuss certain differences in the perception of some of the responses that were particularly more difficult to rate due to mispronunciations of words or differences in the nuance between the test-takers’ responses and the original stimulus. It is suggested that this matrix be used in future studies administering the Korean EIT in order to have more transparency of scores across studies. In addition, the matrix could be improved by adding more sample responses in each score, that is, by adding at least one example for each possible score for every sentence. In addition, the main source of rater disagreement in the present study was mispronunciation, particularly when the learners mispronounced words due to the unclear original stimuli (there were words in the original stimulus that were mispronounced or that followed a regional Korean accent unfamiliar to typical Korean FLLs in the US). It is recommended that in the future, the original stimuli are corrected to reflect a more standard Korean pronunciation of words.

6.2.3. Korean ACTFL Oral Proficiency Interview-computer and Writing Proficiency Test

The Korean ACTFL OPIc and WPT were also administered as additional measures of Korean speaking and writing abilities, respectively. The ACTFL OPI is a large-scale standardized proficiency test that assesses speaking proficiency following the
ACTFL language proficiency guidelines (see Chapter 2, Section 2.1). Its primary intended use is to assess language learners’ ability to communicate “effectively and appropriately in real-life situations” (ACTFL, 2014, p. 5). The ACTFL OPI is used for both academic and working contexts, including commercial enterprises, international organizations, and government agencies. While the OPI is administered by a certified ACTFL interviewer, the Oral Proficiency Interview-computer (OPIc), as the name indicates, is administered via computer with an avatar serving as the guide. Both tests are scored by certified ACTFL raters according to the ACTFL Proficiency Guidelines, based on three main assessment criteria, namely, function, accuracy, and type of discourse produced (i.e., words, sentences, extended discourse). The questions in the OPIc range from talking about familiar topics to supporting an opinion and hypothesizing on abstract subjects. The test is about 20 to 30 minutes long depending on the ability of each test-taker.

The ACTFL Korean WPT is a standardized test of functional writing ability. Its main intended use is to assess learners’ “ability to write effectively and appropriately for real-life writing purposes” (ACTFL, 2012b, p. 4). The writing samples are scored by certified ACTFL raters based on the ACTFL Proficiency Guidelines 2012 – Writing. These guidelines mainly focus on four criteria, namely, (1) functions produced, (2) social contexts and specific content areas within which test-takers produced language, (3) accuracy, and (4) the length and organization of the sample. The structure of the WPT consists of an introduction and warm-up section, where instructions and a warm-up novice-level task are presented, and a writing test section, which presents the prompts. There is a total of four prompts in the test, each suggesting the main ideas that should be
included in the response and its length (e.g., 2 – 3 paragraphs). Similar to the ACTFL OPIc, the time allotted for this test varies according to test-takers’ Korean language abilities, but it ranges from approximately 40 to 90 minutes. Table 7 describes the types of tasks that are included in the WPT and the type of ability expected for each level.

Table 7

*Tasks and Abilities of the WPT by Proficiency Level. (Adapted from ACTFL, 2012b)*

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>Type of task</th>
<th>Expected ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>Elicits limited information on a simple form using isolated words or phrases.</td>
<td>Can produce lists, notes, and limited formulaic information on simple forms and documents. Writing is typically limited to words, phrases, and memorized material.</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Elicits a simple description on a familiar topic.</td>
<td>Can meet a range of simple and practical writing needs, e.g., simple messages and letters, requests for information, notes, etc. Can communicate simple facts and ideas in a loosely connected series of sentences on topics of personal interest and social needs, primarily in the present. Because of vocabulary limitations and errors in basic structures, writing is comprehensible to those accustomed to the writing of non-natives.</td>
</tr>
<tr>
<td>Advanced</td>
<td>Asks for paragraph-length narration of a factual nature using appropriate time frames.</td>
<td>Can write routine, informal, and some formal correspondence, narratives, descriptions, and summaries of a factual nature in all major time frames in connected discourse of a paragraph in length. Writing is comprehensible to all native speakers due to breadth of generic vocabulary and good control of the most frequently used structures.</td>
</tr>
<tr>
<td>Superior</td>
<td>Asks to state and support an opinion and to hypothesize in a cohesive and structured text.</td>
<td>Can produce informal and formal writing on practical, social, and professional topics treated both abstractly and concretely. Can present well-developed ideas, opinions, arguments, and hypotheses through extended discourse. Can control structures, both general and specialized/professional vocabulary, spelling, punctuation, cohesive devices, and all other aspects of written form and organization with no pattern of error to distract the reader.</td>
</tr>
</tbody>
</table>
Before taking the ACTFL OPIc and the ACTFL WPT, participants were asked to respond to a brief self-assessment questionnaire composed of questions about their interests and hobbies as well as six can-do statements, that determined which OPIc test form (i.e., Form 1, Form 2, Form 3, Form 4, Form 5) or WPT writing prompts, were appropriate for each test-taker. For the ACTFL WPT, participants were asked to take this questionnaire two weeks in advanced because the Language Testing Institute (LTI) required about 2 weeks to compile a test booklet containing the appropriate prompts for each learner. Once the booklets were ready, they were sent electronically and the researcher printed them for administration. Like the Korean C-test, it was important that the WPT be administered in paper-based format to avoid disadvantaging learners unfamiliar with the Korean keyboard. The test responses were scanned and mailed to a secure server for trained certified ACTFL raters to score. The ACTFL OPIc was internet-based and thus participants could take the self-assessment questionnaire immediately before taking the test. The responses to the OPIc were sent directly through a secure server to the LTI.

### 6.2.4. Background and self-assessment questionnaire

A questionnaire was created for the purposes of collecting background information about the Korean learners (see Appendix D). The main objectives of the questionnaire were to identify which learners could be considered HLLs and which FLLs, to collect information about participants’ acquisition and use of the Korean language, to learn more about their family language history, and to examine their perception on their Korean language ability. Thus, the questionnaire was composed of fours parts, namely:

---

3 https://www.languagetesting.com/
(1) personal background, which consisted of 15 questions regarding learners’ personal background information (e.g., where were you born?); (2) linguistic history and Korean language use, which consisted of 13 questions regarding learners’ experiences learning languages and the frequency with which they used Korean (e.g., when did you first begin to read Korean? How often do you watch TV in Korean?); (3) family language background and Korean language use, which consisted of 6 questions concerning the language that learners usually use with their families and relatives (e.g., In what language(s) do you speak to your parents?); and (4) self-assessment of Korean proficiency, which consisted of five questions about their ability to use Korean and English language (e.g., How would you rate your proficiency in Korean: reading?) as well as their experiences with Korean language proficiency tests. This part also included 18 can-do statements (e.g., I can introduce myself) which intended to estimate participants’ level of difficulty or ease when using Korean in different situations. The whole questionnaire was administered online through the Qualtrics system and the majority of learners took an average of approximately 5 to 10 minutes to complete it.

Most of the items in the questionnaire were adapted from instruments used in previous studies, such as Marian et al.’s (2007) Language Experience and Proficiency Questionnaire (LEP-Q) and E. J. Kim’s (2006) identity questionnaire, as well as those provided by the National Heritage Language Center from Montrul (2012) and Torres (2012). The can-do statements within the self-assessment part of the questionnaire were reflective of the types of tasks usually encountered by speakers of Korean in everyday life as identified by the TOPIK rating scale (NIIED, 2014). The Likert scale for the can-do statements was adapted from Kondo-Brown (2005). While participants’ responses to
their language backgrounds, language learning experiences, and language use were analyzed to develop a better definition of who heritage language learners are, the responses to the self-assessment part of the questionnaire were used as another criterion measure for the C-test. These responses were particularly useful as part of the empirical backing for the extrapolation inference.

6.3. Data Collection Procedure

The data collection was carried out in two different phases. During Phase 1, participants were asked to complete two different questionnaires, the ACTFL WPT self-assessment survey and the background and self-assessment questionnaire developed by the researcher. As these questionnaires were made available online, participants were required to complete them at home. During Phase 2, participants were asked to come to a language laboratory or a quiet classroom where all four language tests were administered. Due to scheduling and logistics during visits at different institutions, sometimes two to three participants were asked to come in at the same time but were seated separately at different ends of the classroom or computer lab. First, participants were asked to complete the C-test for which 35 minutes were allotted. They were given a 10-minute warning before the time was up. Next, they completed the Korean EIT, which was administered through a YouTube link that would play the audio with the instructions, English sentence practice, and the actual test. The responses were recorded on the computer as well as on a manual recorder. If several participants were taking the tests at the same time, the researcher made sure that they would start the audio at the exact same time in order to avoid hearing each other’s responses or listening to the sentences twice.
The test was approximately 10 minutes long. Following the EIT, participants took the ACTFL OPIc. For this test, they were first asked to take the OPIc self-assessment survey and to test whether the recording of audio was working properly. The time to complete this test varied across learners because the test prompts were based on the results of the self-assessment survey. Participants spent about 20 to 35 minutes on this test. After completing the ACTFL OPIc, participants were given a 10-minute-break. Lastly, they were asked to take the ACTFL WPT. Similar to the ACTFL OPIc, the time to complete the test varied across learners depending on their self-assessment survey responses. The prompts in each test booklet reflected participants’ interests and hobbies as well as their Korean language abilities. Test-takers with higher Korean language abilities were asked to write longer responses and therefore were given about 60 to 80 minutes to complete the test, whereas learners with lower Korean language abilities were guided to write shorter responses, taking them about 40 minutes to complete the whole test. All participants were compensated with a $50 Amazon gift card. Figure 9 illustrates the process of data collection, including information about the tests administered in each phase and the time allotted for them.

Figure 9. Data collection procedure
Chapter 7

Results

Similar to the format of previous studies applying Kane’s (2006, 2011, 2013) Argument-Based approach to validation (e.g., Chapelle, Cotos, & Lee, 2015; Chapelle et al., 2008; Voss, 2012), the results presented in this chapter will follow the interpretive argument chain described in Chapter 5, starting with the theoretical grounds inference and continuing through all inferences until the extrapolation inference. Given that the Korean C-test has yet to be used operationally, the utilization inference will be only briefly described at the end of this chapter as a set of suggestions for future studies. Like the bridge analogy presented in Chapter 4, each inference will be described as being composed of grounds that lead to an intermediate conclusion, which in turn act as the grounds for the next inference until reaching a final conclusion. In order to go from one inference (or bridge) to the next, a “valid ticket” in the form of theoretically- or empirically-based evidence will be presented. All of the inferences in the Korean C-test interpretive argument will be explored separately following a similar structure, which begins by restating the warrant, assumptions, and research questions and then answering the research questions.

7.1. Outcomes Relevant to the Theoretical Grounds Inference

A particular challenge when evaluating the use of shortcut types of language assessments, such as C-tests, cloze tests, or elicited imitation tests is their abstract relationship to a target domain. This is especially true for shortcut measures that are being used as indicators of general language proficiency in research studies dealing with
a wide range of topics. In contrast to other types of language proficiency assessments such as educational-types of assessment (e.g., the TOEFL iBT®), which have a concrete target domain (e.g., academic domain, specifically North American universities), these shortcut measures are purposefully not domain-referenced or skill-specific. For this reason, they are more versatile for research purposes, as they can provide a more holistic measurement of language proficiency that is detached from a specific domain or skill. At the same time, this does not mean that these types of tests can be used for any given purpose. On the contrary, the use of these shortcut measurements of global language proficiency should be limited specifically to providing a quick estimate of global language proficiency (Norris, 2018).

Given its non-domain-referenced quality, the first inference for the Korean C-test interpretive argument chain could not refer to a “domain description.” This is different from domain-referenced tests, where it is common that validation studies using Kane’s (2006, 2011, 2013) validity model evaluate at this first stage, what types of language abilities the test should target and how the test tasks should reflect what test-takers might encounter in the target domain. Generally, empirical backing is collected for the identification of language skills used and types of tasks encountered by learners in a specific domain. In the case of the Korean C-test, there was no specific domain from which language skills and tasks can be identified. Therefore, rather than collecting empirical backing at the baseline of the interpretive argument, it was necessary to theoretically identify C-test constructs and closely examine how they have been characterized and investigated in the literature. These theoretical representations (and associated empirical support from the accumulated research) of the C-test constructs,
then, served as the grounds to the intermediate conclusion that the observations obtained on this C-test can be associated to scores reflecting Korean global language proficiency.

Table 8 presents the warrant, assumption, and the research question for the theoretical grounds inference.

Table 8

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumption</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations of the performance on the C-test reflect language learners’ global language proficiency as characterized by a large body of literature from C-test experts</td>
<td>The C-test format has been subjected to sufficient research to provide a trustworthy foundation for believing that it is a good measure of global language proficiency.</td>
<td>What type of evidence has research on C-test formats provided as a trustworthy foundation for believing that it is a good measure of global language proficiency?</td>
</tr>
</tbody>
</table>

7.1.1. Theoretical characterizations of C-test constructs

One of the main goals and at the same time one major challenge of C-test validation studies has been to determine what it measures. C-tests, like other tests of reduced redundancy (e.g., dictation, noise test, cloze tests), were originally designed to provide a measure of general language proficiency in a first, second, or foreign language (Grotjahn, Klein-Braley, & Raatz, 2002; Klein-Braley, 1997; Raatz & Klein-Braley, 2002). They were based on the assumption that high performing learners will be better than low performing learners at making use of the natural redundancy of language to reconstruct incomplete or distorted texts or messages (Spolsky, 1971). Given that specific rules of text selection and text deletion for constructing a C-test did not target any one specific type of text, genre, word type, or linguistic form, C-tests were generally viewed as integrative tests that targeted multiple linguistic phenomena and language knowledge at the same time. In addition, researchers argued that C-tests measured both receptive and
productive skills because test-takers are required to read and understand the C-test passages (receptive skill) as well as reproduce the missing part of the words (productive skill). Research also revealed that C-tests can be associated with discrete language skills like grammar, vocabulary, listening, and speaking (Chapelle, 1994, Harsch and Hartig; 2015; Janebi Enayat & Babaii, 2018; Kamimoto, 1992; Klein-Braley & Raatz, 1984; Singleton & Little, 1991), as suggested by overall moderate to strong and positive correlations, typically ranging from .64 to .88 (see Eckes & Grotjahn, 2006), between C-tests and other skill-specific tests or subtests. One important caveat of interpreting C-test results that Norris (2018) pointed out is that the interpretation of C-tests should be restricted typically to literacy-based skills. Language learners might develop language skills in one modality (e.g., speaking proficiency) but not in another (e.g., writing proficiency), and thus C-test results should be interpreted accordingly.

Due to their great versatility in measuring learners’ general language proficiency and predicting other language abilities, C-tests have sparked considerable scholarly attention in investigating what exactly it is that they measure. On the one hand, a number of studies has explored the usefulness of C-tests to measure one or two specific constructs, for example learners’ vocabulary skills (e.g., Chapelle, 1994; Karimi, 2011; Singleton & Little, 1991) or reading skills (e.g., Cohen et al., 1984). These studies mostly analyzed C-tests in terms of their relationship with other tests of language proficiency that measure specific language skills. Harsch and Hartig (2015), for example, investigated the predictive power of a German C-test in assessing receptive skills, such as reading and listening comprehension, by comparing it to another reportedly widely used test, the Meara’s X-Lex test (i.e., Yes/No vocabulary test). Four instruments were used in
the study, namely, a listening and a reading comprehension test, a C-test, and the X-Lex test. Correlation analyses showed that C-test correlated higher with listening and reading comprehension measures at $r = .76$ and $r = .72$, respectively, than did the X-Lex with the same listening and reading measures, which yielded correlation coefficients of only $r = .49$ and $r = .39$. In addition, traditional regression analyses and Structural Equation Model (SEM) analyses revealed that the C-test had more predictive power than the X-Lex, as it explained more of the variance in the receptive skills. The authors argued that unlike the X-Lex test, the contextualized and integrative nature of the C-test requires test-takers to use internal processes that are more closely related to spoken and written language processing.

On the other hand, several studies have investigated the validity and reliability of a C-test for assessing a general construct of language proficiency (Chihara et al., 1996; Dörneyi & Katona, 1992; Grotjahn, 1987; Klein-Braley, 1985; Lee-Ellis, 2009; Lei, 2008; Raatz & Klein-Braley, 2002; Roos, 1994). The issue at stake here has been whether C-tests have high reliability and internal consistency for assessing learners’ general language ability and their effectiveness in distributing learners into a range of different proficiency levels. This line of inquiry has been approached in several ways.

One approach has been to conduct factorial analyses and analyze C-test results together with results on other tests measuring a variety of different skills (Grotjahn, 1987; Klein-Braley, 1994; Klein-Braley & Raatz, 1984; Raatz, 1984). Generally, these studies have found that C-test scores together with scores on other language tests or subtests tend to load high on one factor, later defined as general language proficiency. Eckes and Grotjahn (2006), for example, developed a German C-test, which they administered to
843 participants together with the TestDaf (Test of German as a Foreign Language). The latter test was used as a criterion measure. The rationale behind using the TestDaf was that it was a standardized test that measured language proficiency in four different skills, namely, reading, listening, writing, and speaking. Unlike previous studies, Eckes and Grotjahn applied more stringent statistical measures with a rather large sample size. The results of a confirmatory factor analysis revealed that the C-test and TestDaf scores could be explained with a one-factor or two-factor model. There were only slight differences between the two models, and the researchers preferred the more parsimonious one-factor model. Accordingly, they concluded based on the results of the satisfactory fit of the scores to the one-factor model, that the C-test was measuring one single construct, which they referred to as general language proficiency.

Another approach to exploring the general proficiency construct has been through examining test-takers’ processing strategies when completing C-tests. Verbal protocols after or during the completion of a C-test have allowed researchers to assess how test-takers approached the completion of a C-test and what kind of skills they used to do so (e.g., Babaii & Ansary, 2001; Babaii & Jalali Moghaddam, 2006; Feldmann & Stemmer, 1987; Stemmer, 1991). For example, Babaii and Ansary (2001) analyzed the verbal protocols from 32 Iranian students of English and found that test-takers mentioned the use of four main types of cues, which represented both macro- and micro-level features of language. These cues were classified as (a) automatic processing, described as the automatic completion of the blanks due to the high frequency and familiarity of a word; (b) lexical adjacency, defined as filling in the blank by relying on the words immediately before and after the deleted word; (c) sentential cues, explained as the reliance on
grammatical features such as tense or cohesive devices to complete the words; and (d) top-down cues, described as using background knowledge and topic familiarity to fill in the blanks. Although learners used these cues to varying degrees, the researchers concluded that the test-takers referred to all four types. Accordingly, they indicated that the C-test was measuring both micro- and macro-level features of language (p. 216).

Additional research has also investigated the role of language proficiency as a moderating factor that affects processing strategies during the completion of C-tests (Babaii & Fatahi-Majd, 2014; Janebi Enayat & Babaii, 2018). As an illustration, Sigott (2004, 2006) characterized constructs underlying C-tests as fluid, arguing that while C-tests trigger a more holistic language proficiency, high- and low-performing C-test-takers process C-test items differently. In his 2006 study, Sigott administered to 60 EFL learners four C-test items in two different conditions, namely, a decontextualized condition (i.e., items were presented only in the sentence that they appeared on) and a fully contextualized condition (i.e., items appeared the same way as in a normal C-test). Results indicated that while high-proficiency test-takers were able to solve items in both conditions, low-proficiency test-takers were more likely to correctly solve the items in the contextualized condition. Hence, Sigott explained that high-proficiency test-takers are able to solve more C-test items because they can process decontextualized items (i.e., items at a sentence level), whereas low-proficiency learners need context beyond the sentence level in order to complete the mutilated words. Similarly, Babaii and Fatahi-Majd (2014) conducted think-aloud protocols on both high- and low-proficiency test-takers in an English C-test and also found differences between both learners in the way they approached the task of completing C-test items. They focused on the instances of
failed restorations of words in the C-test. Unlike Sigott (2006), they found that low-proficiency test-takers resorted mostly to the local cues surrounding the incomplete words, particularly when they could not understand the passage as a whole. On the other hand, high-proficiency test-takers approached the task of completing words by making use of linguistic, textual, and contextual resources.

Furthermore, in a recent study Janebi Enayat and Babaii (2018) investigated the extent to which depth and breadth of vocabulary knowledge helped test-takers from lower- and upper-intermediate proficiency levels solve C-test items. They conducted multiple regression analyses between C-test scores and a battery of tests including the Oxford Quick Placement Test, the Word Associates Test, and the Vocabulary Levels Test. They found that the breadth and depth of vocabulary knowledge played different roles depending on learners’ proficiency levels. For lower-intermediate proficiency level learners, neither breadth nor depth predicted C-test performance. In contrast, for upper-intermediate level learners, depth of vocabulary was able to significantly predict C-test performance, explaining 13% of the variance of the C-test performance.

All in all, these studies agree with the aforementioned basic assumption of reduced-redundancy tests that higher-proficiency level learners will perform better at solving C-test items than lower-proficiency level learners. They also suggest that certain language skills or strategies (e.g., understanding textual and contextual cues) might not be acquired until learners reach a certain level of proficiency. Moreover, these studies also underscore the importance of taking into account differences in language proficiency when making interpretations about C-test performance.
7.1.2. Summary of theoretical grounds inference

Despite the ongoing debate on what C-tests measure, the general consensus is that C-tests can provide an estimate of general language proficiency as they target multiple linguistic phenomena simultaneously. Furthermore, researchers have pointed out that these tests assess both receptive and productive skills and that learners make use of both micro- and macro-level language cues to complete C-test items. There are, however, two important caveats to the interpretation of C-test results. First, C-tests require test-takers to have at least basic literacy skills to be able to read C-test passages and to write the missing parts of the words. Although studies have found considerably high correlations with criterion measures such as speaking (e.g., Arras, Eckes, & Grotjahn, 2002; Eckes & Grotjahn, 2006), the interpretation of C-test scores should be limited to literacy-based skills due to the written modality of the test. Some learners might have exceptionally good oracy skills but lack literacy skills altogether. C-tests would not be able to provide any information about general language proficiency for those kinds of learners. Second, findings from studies on C-test item processing point to the crucial role of language proficiency as a moderating factor (e.g., Janebi Enayat & Babaii, 2018; Sigott, 2004, 2006). C-tests appear to trigger different skills and strategies depending on the test-takers’ language proficiency level. Even though this type of research has mostly focused on English C-tests, it is important to keep this caveat in mind when making interpretations and claims of what C-tests are measuring when targeting different language learners.

The warrant associated with the theoretical grounds inference requires the Korean C-test (Son et al., 2018) to reflect language learners’ global language proficiency as
theoretically characterized by C-test research. As described throughout this section, extensive research has been conducted to identify the underlying constructs of C-tests. The Korean C-test (Son et al., 2018) was designed with these theoretically identified constructs in mind. It followed the C-test construction principles laid out by C-test creators (Raatz & Klein-Braley, 2002) under the same assumptions that the C-test would provide a global measure of language proficiency based on test-takers’ literacy-based skills. In other words, the observation of test-takers’ performances on the Korean C-test can reflect their general language proficiency.

7.2. Outcomes Relevant to the Evaluation Inference

The second inference in the Korean C-test interpretive argument was the evaluation inference. It was based on the warrant that test-takers’ performance on the C-test can be accurately evaluated and scores can be provided to measure this performance. In addition to the evaluation of performance, the text selection process was also implicated in this inference because texts needed to be appropriately selected in order for them to target different levels of proficiency. Thus, backing for this inference was gathered from the development stage of the C-test as well as the statistical characteristics of the items on the test. Table 9 summarizes the assumptions underlying this warrant and the research questions that emerge from them.
Table 9

Evaluation Inference Warrant, Assumptions, and Research Questions

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumptions</th>
<th>Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations of the performance on the Korean C-test can be evaluated to provide scores that reflect general ability of the Korean language.</td>
<td>1. The C-test text selection procedure is appropriate for selecting texts that can distribute learners of varying degrees of language abilities. 2. The statistical characteristics of items in the C-test are appropriate for norm-referenced decisions. 3. Items on the C-test provide an accurate measurement for distinguishing test-takers with a wide range of language abilities.</td>
<td>1. To what extent is the C-test text selection procedure during the test development stage appropriate for selecting items that can distribute learners of varying degrees of language abilities? 2. What are the statistical characteristics of the C-test items and measures? 3. To what extent are items on the C-test accurately assessing test-takers and distinguishing their wide range ability levels?</td>
</tr>
</tbody>
</table>

7.2.1. Text selection

To answer the first research question associated with this inference, it is necessary to refer back to Chapter 3 and review some of the decisions taken by the designers of the test in their exploratory study with 38 learners (Son et al., 2018) in terms of selecting the texts to be included on the C-test for the larger-scale validation study conducted on 93 Korean learners. As briefly described in Chapter 3, the text selection process in the development stage described in Son et al. (2018), started with a pool of 24 texts collected from a wide range of contexts from newspaper articles, internet blogs, advertisement, and textbooks. This pool of texts was later reduced to 15 texts according to the authenticity of the text source, the structure of the text, and the facility to apply the C-test deletion rule to the text. Nine highly experienced Korean language instructors were asked to rate and rank the 15 texts in term of how difficult they would be for KFL learners in their classroom. They were instructed to classify all texts into five groups: Easiest Text,
Somewhat Easy, Medium, Somewhat Difficult, and Most Difficult. The 15 texts were then pilot-tested on 37 NSs of Korean. The results of both the ranking of the texts based on difficulty level as well as the accuracy rate of NSs and their perception of each task, helped reduced the pool further to 10 texts. It was considered important to have two texts per difficulty level. Son et al. then administered the 10-text Korean C-test to 38 KFL learners. Results indicated that the ranking based on expert reports and the NS pilot test was accurate as observed through the descriptive statistics and exploratory Rasch measures.

Out of the 10 texts, seven were chosen for the purposes of the current validation study. These seven texts were selected based on the results of the exploratory Rasch analysis on the 10-text C-test in Son et al. (2018). Fit model statistics as well as Item Characteristic Curves (ICCs) were taken into account in order to select the seven texts. In Son et al. (2018), five texts were selected as the most fitting. However, given that the sample size was rather small ($N = 38$), it was decided that two more texts would be added for the current large-scale validation study ($N = 93$). Two additional texts would only add about 10 more minutes to the total test administration time. Therefore, it was considered beneficial to have a larger pool of texts from which to choose for the final version of the Korean C-test.

According to the results of the exploratory Rasch analysis based on the sample in Son et al. (2018), the best fitting seven texts of the original 10 that were able to distribute learners across at least six levels of proficiency showed a person separation span of 6.88 logits and a separation reliability of .98. The item separation index was 11.16 logits with a separation reliability of .99. Three person outliers were excluded from the final sample
size for this analysis. As presented in Table 10, all seven texts fitted the model with infit and outfit measures well within the acceptable ranges (Linacre, 2012). The point-biserial correlation coefficients were generally high ranging from .76 to .94, indicating that all items were contributing in equivalent ways to the overall C-test scores.

Table 10

*Item Fit Statistics for Seven Items (N = 35)*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Infit</th>
<th>Outfit</th>
<th>Point-Biserial Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SE</td>
<td>MnSq</td>
<td>Z-Std</td>
</tr>
<tr>
<td>Text 1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-2.27</td>
<td>0.13</td>
<td>0.97</td>
</tr>
<tr>
<td>Text 4</td>
<td>-0.92</td>
<td>0.09</td>
<td>1.24</td>
</tr>
<tr>
<td>Text 5</td>
<td>-0.01</td>
<td>0.09</td>
<td>0.97</td>
</tr>
<tr>
<td>Text 6</td>
<td>0.28</td>
<td>0.09</td>
<td>0.64</td>
</tr>
<tr>
<td>Text 7</td>
<td>0.58</td>
<td>0.11</td>
<td>0.67</td>
</tr>
<tr>
<td>Text 9</td>
<td>1.29</td>
<td>0.10</td>
<td>0.69</td>
</tr>
<tr>
<td>Text 10</td>
<td>1.03</td>
<td>0.09</td>
<td>1.08</td>
</tr>
</tbody>
</table>

*Notes.*<sup>a</sup> After removal of three outliers from the original N = 38 in Son et al. (2018).<sup>b</sup>Text labels are based on the 10-text C-test (Items 2, 3, and 8 were eliminated)

Item Characteristic Curves (ICCs) were also examined in the pilot study as another criterion for item selection into the present study (see Appendix E). Most of the items had an overall good fit to the expected curves with no observations outside of the 95% confidence interval. Only Text 9 showed some deviations from the expected curve. However, it was decided that keeping both advanced-level items (Texts 9 and 10) was necessary for the current large-scale validation study as the KFL learner pilot-test in Son et al. (2018) showed that the test was not appropriately discriminating among the highest-performing learners. In other words, in the pilot study, the C-test might have been too easy for learners in the upper ability range (see item map in Appendix F). To avoid large ceiling effects, which are commonly observed in C-test studies (e.g., Grotjahn, 1987,
Klein-Braley, 1985), it was important to keep all texts at the highest level of difficulty for the larger-scale validation study.

In sum, the rigorous process of text selection followed by Son et al. (2018) yielded seven items that appropriately and accurately assessed test-takers from a wide variety of proficiency levels. As described throughout this section, several criteria were taken into account to make this selection, including the authenticity of text source, the structure and difficulty of the texts, as well as the extent to which the items fit the Rasch model. The next section presents the results of the large-scale validation of the Korean C-test, which included these seven items.

7.2.2. Statistical characteristics of the C-test items

The second research question associated with the evaluation inference refers to the statistical characteristics of the C-test and its items in the present sample of \( N = 93 \) Korean learners. Descriptive statistics indicated that the Korean C-test was effective in eliciting a broad range of test-taker abilities, as could be observed through the wide range of scores which extended from 26 to 169 points (see Table 11). Moreover, the mean score was higher than the midpoint of the maximum possible score, indicating that a large proportion of learners exhibited considerable proficiency in Korean. This was expected as the lowest proficiency level participants recruited for the study were learners who had taken at least two semesters of Korean language courses (i.e., one year of Korean courses) or had been placed into an equivalent course level. The standard deviation was relatively large suggesting considerable variability of performance among the learners. Furthermore, C-test performance was also examined for each type of learner group (i.e., HLL and FLLs) separately. Descriptive statistics revealed that the HLL group (\( N = 41 \)
performed better than the FLL group \((N = 52)\), with a mean difference of 17.97 points (less than one standard deviation unit). In addition, in terms of within-group variation, the HLLs showed more variation in their performance with a larger standard deviation as well as a wider score range, when compared to the FLLs.

Table 11

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>HLLs</th>
<th>FLLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>93</td>
<td>41</td>
<td>52</td>
</tr>
<tr>
<td>(k)</td>
<td>175</td>
<td>175</td>
<td>175</td>
</tr>
<tr>
<td>Mean</td>
<td>99.68</td>
<td>109.71</td>
<td>91.77</td>
</tr>
<tr>
<td>SD</td>
<td>35.51</td>
<td>36.22</td>
<td>33.18</td>
</tr>
<tr>
<td>Min</td>
<td>26</td>
<td>26</td>
<td>34</td>
</tr>
<tr>
<td>Max</td>
<td>169(^a)</td>
<td>169</td>
<td>157</td>
</tr>
<tr>
<td>SEM</td>
<td>±7.32</td>
<td>±11.43</td>
<td>±9.24</td>
</tr>
</tbody>
</table>

Notes: The maximum possible score was 175

In addition to descriptive statistics, an independent t-test was conducted in order to examine whether the differences in C-test means scores between the HLL- and FLL group were statistically significantly different. Results showed that there was a statistically significant difference between the two learner groups with \(t(91) = 2.49, p = .02\). The magnitude of this difference, however, was small to medium with \(d = 0.52\) (Oswald & Plonsky, 2010), which corroborates the results of the descriptive statistics. It is worth noting that the results of the inferential statistics should be interpreted with caution as the distribution of C-test scores of the HLL- and FLL group showed slight negative and positive skewness, respectively. All in all, these findings indicated that there was indeed a difference in C-test performance between both learner groups, with HLLs performing better overall than the FLLs.

The difference in C-test performance between the two groups could be attributed to the C-test assessing HLLs and FLLs in different ways. Accordingly, it was important
to take a closer look at the C-test performance for each learner group separately and examine whether the C-test was successful in achieving its main intended use for each learner group, which was to distribute learners along a wide range of proficiency levels. One way to do so was to inspect whether C-test scores were normally distributed. Non-normal distributions for either learner group would indicate that the C-test was not functioning appropriately or that the learner sample was not representative of a wide range of abilities. Accordingly, Kolmogorov-Smirnov’s tests of normality were conducted for C-test performance of both learner groups combined as well as for each learner group separately. The results showed no deviation from normality for any of the groups, combined or separately (see Table 12 and Figure 10). At the same time, these findings showed that the overall statistical characteristics for the C-test appeared to be appropriate for norm-referenced decisions.

*Figure 10. Korean C-test score distribution (N = 93)*
Table 12

*Kolmogorov-Smirnov Test of Normality*

<table>
<thead>
<tr>
<th></th>
<th>Statistics</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.06</td>
<td>93</td>
<td>0.20</td>
</tr>
<tr>
<td>HLLs</td>
<td>0.07</td>
<td>41</td>
<td>0.20</td>
</tr>
<tr>
<td>FLLs</td>
<td>0.10</td>
<td>52</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Additional analyses were conducted to assess the extent to which the items on the C-test were functioning appropriately in assessing learners from different language abilities. As a reminder, the term “item” here refers to the polytomous or super-items or texts, which contain 25 micro-items each. Table 13 presents the descriptive statistics per item ordered by level of difficulty from easiest to most difficult according to Korean instructors (see previous section). The mean scores indicated that the test-takers’ performance on each item gradually decreased as items increased in predicted difficulty. The only exception was Text 7, which was found to be an easier item than Text 5 and Text 6. Moreover, the standard deviations gradually increased as the level of difficulty for each item increased, suggesting that there was a ceiling effect on easier texts and that more variation in performances among test-takers as items became more challenging. The range of scores also pointed to the same phenomenon. The scores for items 1 through 3 spanned nearly all possible score range, while scores for items 4 through 7, which were considered most difficult, extended across the entire available score range (0 -25). In Figure 11, the mean scores and 95% confidence intervals for items 1 through 4 confirmed that the texts were noticeably different in terms of their difficulty levels. On the other hand, the last four texts (Text 4, 5, 6, 7) yielded similar means, suggesting that these items could be of very similar difficulty levels.
Table 13

*Descriptive Statistics for C-test Items*

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1*</td>
<td>21.57</td>
<td>3.21</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>Text 2</td>
<td>19.34</td>
<td>5.04</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>Text 3</td>
<td>16.41</td>
<td>5.28</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>Text 4</td>
<td>11.70</td>
<td>6.35</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Text 5</td>
<td>10.19</td>
<td>6.36</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Text 6</td>
<td>9.83</td>
<td>6.60</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Text 7</td>
<td>10.63</td>
<td>6.25</td>
<td>0</td>
<td>23</td>
</tr>
</tbody>
</table>

*Notes.* *Texts were relabeled (Texts 2 through 7 are equivalent to Texts 4, 5, 6, 7, 9, 10 in Table 10)*

All in all, the descriptive statistics provided useful information of the general trends of the C-test and the way the underlying items on the C-test were functioning. The C-test appeared to target test-takers with a broad range of language abilities as observed by the mean scores of test-taker performances as well as the score range for each item. Furthermore, most of the items varied in difficulty levels with mean scores gradually decreasing as the predicted item difficulty levels increased. Overall, the statistical
characteristics of the C-test appeared to be appropriate for norm-referenced decisions with means following a normal distribution. Nevertheless, for more stable estimates of test-takers and item measures, IRT analyses were conducted and are described in the next section.

7.2.3. Item response theory measures for C-test items

In order to answer the third research question for the evaluation inference, which asked the extent to which items on the C-test accurately assessed and distinguished among test-takers with a wide range of ability levels, Rasch model IRT analyses were conducted using WINSTEPS (Linacre, 2012). Given that the texts within the C-test differed in difficulty level and the range of scores for each item was different, a Partial Credit Model (PCM) was considered most appropriate for the Rasch analysis. In line with previous C-test research (Lee-Ellis, 2009; Norris, 2006) and the Korean C-test development study (Son et al., 2018), each text was considered a polytomous item or super-item scored on a 0-25 scale. Each blank or micro-item within these super-items could not be considered as an item on its own due to the assumption of local dependency. That is, within a passage, it was assumed that the completion of one incomplete word depended on the completion of another. As Rasch modeling assumes local independence of items (Draney, 1996), it was necessary to consider each text to be polytomous.

The summary statistics for persons and items, shown in Table 14, indicated that the C-test scores overall fit the model predicted by the Rasch model with infit and outfit measures within the acceptable ranges (Linacre, 2012). The person separation value of

---

4Linacre (2012, p. 25) describes the interpretation of parameter-level as follows: “>2.0 Distorts or degrades the measurement system; 1.5-2.0 Unproductive for construction of measurement, but not degrading; 0.5-1.5 Productive for measurement; <0.5 Less productive for measurement, but not degrading. May produce misleading good reliabilities and separations.”
5.50 suggested that about five different levels of test-taker ability could be distinguished by the seven items. The item map in Figure 12 helped illustrate the test-taker language abilities, which appeared to span over 5 logits. Moreover, the high item separation index showed that the test-taker sample included learners with varying language abilities. The very high separation reliability indices for both persons and items indicated that the model was reliable in separating test-takers and items based on difficulty levels.

Table 14

*Summary Statistics for Person and Item Fit*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Infit</th>
<th>Outfit</th>
<th>Separation</th>
<th>Separation reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>MnSq</td>
<td>Z-Std</td>
</tr>
<tr>
<td>Person</td>
<td>0.17</td>
<td>0.19</td>
<td>0.92</td>
<td>-0.20</td>
</tr>
<tr>
<td>Item</td>
<td>0.00</td>
<td>0.05</td>
<td>0.96</td>
<td>-0.50</td>
</tr>
</tbody>
</table>
Upon closer inspection, several outliers were identified in the data for both items and persons. Following Linacre (2012), the most aberrant items and persons were excluded from the data first until no noticeable difference was found in the overall measures with or without the misfitting item or person. Just to illustrate, Text 7 was considered one of the most problematic items. Namely, not only in terms of its fit statistics but also when examining the content of the passage, it was found that one word was repeated four times throughout the text and one of these times was in the first

Figure 12. Seven-text C-test item map (N = 93)
sentence that was left untouched. In other words, some test-takers with good test-taking strategies might have guessed the answer for three of the blanks without needing to process the whole passage for its meaning and structure. This in turn could have made it easier to guess other answers within the passage. For this reason, as observed through the descriptive statistics, this item was found to be easier than predicted by the experts.

Hence, Text 7 was the first to be eliminated. Following the same line of reasoning, a total of 3 items (i.e., Text 1, 5, 7) were eliminated from further analysis. In addition, a total of five people (i.e., two FLLs and three HLLs) were eliminated as outliers. These were five examinees who had either taken a 4th-year level of Korean (N = 3; 2 FLLs and 1 HLL) or the first year of Korean courses (N = 2 HLLs). In other words, some of the learners at each extreme of the language ability continuum appeared to have created some noise in the IRT model. The resulting fit statistics for each final item retained and without the five outliers are presented in Table 15.

Table 15

*Item Fit Statistics for Four Items (N = 88)*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Infit</th>
<th>Outfit</th>
<th>Point-Biserial Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>MnSq</td>
</tr>
<tr>
<td>Text 2</td>
<td>-1.19</td>
<td>0.06</td>
<td>1.07</td>
</tr>
<tr>
<td>Text 3</td>
<td>-0.38</td>
<td>0.06</td>
<td>0.76</td>
</tr>
<tr>
<td>Text 4</td>
<td>0.43</td>
<td>0.06</td>
<td>0.83</td>
</tr>
<tr>
<td>Text 6</td>
<td>1.13</td>
<td>0.06</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Note. Person Separation: 5.53, Reliability: .97; Item Separation: 13.64, Reliability: .99

As displayed in Table 15, all four chosen items showed good fit to the predicted Rasch Model with infit and outfit measures within the acceptable ranges (Linacre, 2012). These four items had a person separation index of 5.53 logits with a reliability coefficient of .97. Furthermore, the item separation spanned 13.64 logits with a reliability coefficient of .99. These results suggested that the C-test containing four items varied in difficulty
levels and they could discriminate learners up to about five levels of proficiency. In addition, there appeared to be a wide range of ability levels represented in the test-taker sample as observed through the item separation index as well as the item map (Figure 13).

Figure 13. Four-Text C-test item map (N = 88)
It is worth mentioning that ceiling effects were found in the data. As mentioned before, this is commonly observed in C-test studies that include very advanced language learners for whom the C-test might be too easy (see Grotjahn, 1987). Grotjahn, Klein-Braley, and Raatz (2002) explained that these ceiling effects are in line with the theory underlying C-test because “as the learner becomes more proficient, so s/he approaches the proficiency of the adult educated native who is expected to attain a score which varies from 100 per cent only by chance” (p. 97). The six very high performers in the Wright map in Figure 13 were identified as advanced-level learners based on the course they were enrolled in when participating in the study. Five of these test-takers were HLLs and one was a FLL. Their responses to the background questionnaire revealed that three out of the five HLLs had received almost all of their elementary school education in Korea (i.e., 6 years), whereas two had received kindergarten and 1st year of elementary school education in Korea. The FLL reported having studied Korean in Korea during 1.5 years in university and in Korean language institutes. It is important to note that years of education in Korea for this FLL might not have been the most important factor that explained her high performance on the C-test, as other FLLs in the advanced level also reported similar long periods of university study in Korea in the background questionnaire. All in all, then, the results of these high performers agree with what Grotjahn and colleagues have pointed out, which is that as advanced-level learners become more proficient in the language, their performances resemble more to that of educated NSs. The language constructs underlying C-test performance will be further discussed in the explanation inference section.
7.2.4. **Summary of the evaluation inference**

The empirical backing presented to support the assumptions underlying the claim that *observations of the performance on the Korean C-test can be evaluated to provide scores that reflect general ability of the Korean language for learners with a wide range of ability levels* was based on the text selection procedure and the descriptive statistics as well as the IRT measures of the items on the C-test. First, the rigorous procedure of selecting texts that would be included in the C-test for the present larger-scale validation study yielded seven texts with good fit to the Rasch model. These seven items functioned well in distributing learners from a broad range of ability levels. Second, the descriptive statistics of the larger-scale administration of the C-test suggested that the texts varied in difficulty level and that they were able to distribute learners along a normal distribution. Third, the Rasch Model analyses corroborated the results of descriptive statistics indicating that the items included in the C-test varied in difficulty and were appropriate for test-takers with different ability levels.

It was mentioned before that a possible rebuttal to the evaluation inference claim was that the test-takers in this sample did not represent a wide range of language abilities. Nevertheless, as demonstrated by the descriptive statistics and the IRT measures, as well as inferred from background questionnaire information about their learning experiences, there appears to be no backing for this rebuttal. The descriptive statistics indicated an overall normal distribution with learners ranging from low-performers to very high performers, who performed closed to NS-levels (as based on the NS pilot in Son et al., 2018). This also held true when examining the HLL and FLL learner population separately. As indicated by the tests of normality, the two learner populations were
distributed along a single normal curve. In addition, the IRT measures, specifically the person and item separation indices, revealed that the Korean learner population in this study represented varying proficiency levels.

7.3. Outcomes Relevant to the Generalization Inference

The third inference in the Korean C-test interpretive argument chain was the generalization inference, which was based on the claim that the observed scores on the C-test consistently reflect Korean language learners’ general language proficiency. This inference was mainly concerned about the reliability measures of the overall C-test and its items. Most important, it also included the assumption that the C-test can consistently measure both types of learners, heritage and foreign language learners, without any noticeable biases. Moreover, this inference focused on the replicability of the results in future administrations of the C-test. Table 16 summarizes the assumptions underlying the generalization claim and states the research questions that originated from them. While empirical backing is presented in this section for Assumptions 1 through 4, the evidence for assumption 5 was already described in detail in Chapter 3 section 3.2.2 (also see Son et al., 2018).
Warrant | Assumptions | Research Questions |
---|---|---|
Observed scores are estimates of expected scores that consistently reflect Korean learners’ (both HLLs and FLLs) language proficiency across items on the C-test. | 1. The items on the C-test can provide reliable and consistent estimates of test-takers’ performance. | 1. To what extent does the Korean C-test produce scores with acceptable reliability? |
| 2. The configuration of items on the C-test is the most appropriate for providing reliable estimates of test-takers’ performances. | 2. What configuration of items is the most appropriate for providing reliable estimates of performances from learners with varying degrees of language abilities? |
| 3. The number of tasks on the C-test is sufficient to provide stable estimates of test-takers’ performances. | 3. Does the C-test include a sufficient number of items to provide stable estimates of performances from learners with varying degrees of language abilities? |
| 4. The observed scores on the C-test consistently reflect the language abilities of both Korean heritage and foreign language learners. | 4. Do observed scores on the C-test consistently reflect the Korean language abilities of Korean learners regardless of whether they are heritage or foreign language learners? |
| 5. The C-test text selection and deletion method are sufficiently detailed and consistent to produce equivalent test forms. | |

### 7.3.1. Reliability measures of C-test items

The first, second, and third research questions associated with the generalization inference were closely related to each other and thus are discussed together in this section. To address the first research question, reliability measures were calculated for the overall C-test when taking into account all seven items and then when just taking into account the four best fitting items identified in the previous section. Table 17 displays Cronbach’s Alpha coefficients for all test-takers as well as the coefficients when HLLs and FLLs were examined separately. Although the reliability coefficients decreased when deleting three items, the overall very high reliability coefficients suggested that the C-test
was consistently and appropriately measuring Korean language learners. This held true for both learner groups when examining them separately. In other words, the Korean C-test consistently distributed both heritage and foreign language learners.

Table 17

*Reliability Measures for C-test when Including Seven and Four Items*

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>HLLs</th>
<th>FLLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ for Seven-item C-test</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>$\alpha$ for Four-item C-test</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
</tr>
</tbody>
</table>

In regard to the second research question on the appropriate configuration of items in the C-test, the results of the reliability measures and IRT measures were both taken into account. First, Cronbach’s Alpha reliability coefficients were examined when certain items were deleted. As observed in Table 18, a slight improvement in reliability measures was observed when deleting Text 2, whereas deleting Text 3, 4, and 6 showed a decrease in reliability. Despite the usefulness of reliability coefficients when deciding which items to select, this measurement on its own did not enough provide enough information for item retention, as it does not take into consideration the difficulty levels of the items. Since the main intended use of the C-test is to provide an overall estimate of general language proficiency for learners with a wide range of proficiency levels, the items included in the C-test should not only be reliable but also target learners with varying levels of proficiency. Therefore, IRT measures were also examined when making a decision about the final configuration of items on the Korean C-test. For this reason, items such as Text 5, which was also reliable, was deleted because it was found to be very similar to Text 6 in terms of level of difficulty. In addition, the Rasch Model suggested that Text 6 had a better fit in the model than Text 5. This final configuration of
only four items was still nearly as effective in distributing learners across at least 5 different proficiency levels as the configuration with all seven items, with a person separation index of 5.53 (see Table 15).

Table 18

<table>
<thead>
<tr>
<th>Reliability Measures for each Item on the C-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s Alpha if item deleted</td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td>Text 2</td>
</tr>
<tr>
<td>Text 3</td>
</tr>
<tr>
<td>Text 4</td>
</tr>
<tr>
<td>Text 6</td>
</tr>
</tbody>
</table>

In order to answer the third research question regarding the appropriate number of items on the C-test, it was important to go back to reviewing the main intended uses of the test. Having a sufficient number of items would mean that the items that have been selected are appropriately targeting all learners with low-, intermediate-, and advanced-levels of proficiency. As could be observed through the overall performance of test-takers on the C-test (see Figure 13), for several test-takers the C-test items might have not been challenging enough. It could then be claimed that an insufficient number of items was included on the C-test to be able to discriminate among these very high performing test-takers. Nevertheless, the main intended use of the C-test as stated in section 5.1 is to provide an estimation of Korean language learners’ general language proficiency that is useful for conducting research on SLA and Heritage Language Acquisition (HLA). In other words, the C-test should be able to distribute learners into different proficiency levels ranging from low to advanced-levels. Thus, rather than discriminating learners within the high-proficiency levels, C-test scores should be used as broad indicators of language proficiency. In addition, as aforementioned, these high-performing learners
were expected to score high as they were identified as being enrolled in the most advanced levels of Korean courses and some had received formal education in Korea for several years. The underlying structure of the C-test allows learners with higher proficiency levels to score as high as NSs with accuracy rates higher than 90%, as observed for the native speaker group in the pilot in Son et al. (2018). As a result, for the primary intended use of the Korean C-test, four items appear to be enough to distribute learners into different proficiency levels as observed in SLA and HLA research. It is important to also underscore the scope of the intended use and be specific about what C-test results should be (or should not be) used for. More specifically, the Korean C-test should be used only for the purposes of assessing learners’ general language proficiency and distributing learners across a wide range of proficiency levels. At the same time, the C-test should not be used for discriminating learners’ language abilities at the advanced levels.

7.3.2. Differential item functioning for heritage language learners and foreign language learners

The fourth research question associated with the generalization inference explored the extent to which the results of the Korean C-test can be generalized to different types of language learners, such as HLLs and FLLs. This question was most relevant to this dissertation study as it was important for the Korean C-test to be able to assess both HLLs and FLLs without items being biased towards one or the other group. To explore whether the C-test items were functioning the same way for both learner groups, first, the descriptive statistics of C-test performance across learner groups were examined for each of text, and second, a Differential Item Functioning (DIF) analysis was conducted. DIF
analysis helped evaluate the item difficulty level for each group on a given item when holding other item difficulties and person measures constant (Linacre, 2012).

Descriptive statistics showed that HLLs’ performance on all four C-test items was slightly better when compared to the FLLs’ performance (Table 19). While for the HLLs there appeared to be a more gradual decrease in scores across items, for the FLLs there was a big decrease of scores between Text 3 and Text 4 (i.e., 5.48 score point difference) and a very small decrease between Text 4 and Text 6 (i.e., 1.15 score points of difference). In other words, for the HLLs, items gradually increased in difficulty, whereas for the FLLs, this increase in difficulty was more abrupt from Text 3 to Text 4. All in all, like the previous Rasch model results (see Table 15 and Figure 13) the items increased in difficulty for both learner groups following the same pattern, with Text 2 being the easiest item and Text 6 being the most difficult one.

Table 19

<table>
<thead>
<tr>
<th></th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
<th>Text 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HLL</td>
<td>FLL</td>
<td>HLL</td>
<td>FLL</td>
</tr>
<tr>
<td>Mean</td>
<td>21.24</td>
<td>17.85</td>
<td>17.59</td>
<td>15.48</td>
</tr>
<tr>
<td>SD</td>
<td>4.51</td>
<td>4.96</td>
<td>5.38</td>
<td>5.06</td>
</tr>
<tr>
<td>Min</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Max</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

As observed through the descriptive statistics, compared to the FLLs, the HLLs performed better on all items. It was necessary then, to explore whether these items were biased towards a specific learner group, that is, whether certain items were unusually more difficult for one learner group than for the other. Therefore, DIF analysis was conducted for all four items on the Korean C-test between learner groups. The DIF
analysis indicated that there was no noticeable DIF contrast between HLLs and FLLs for any of the items. DIF contrasts can be interpreted as the effect size in logits (Linacre, 2012) and a difference greater than 0.5 logits indicates that there is a statistical and substantive variance across groups or classes (Bond & Fox, 2007). The DIF contrast for the Korean C-test items ranged from 0.16 to 0.38 (see Figure 14), thus no noticeable DIF was detected. In other words, all items were functioning similarly for both learner groups with no noticeable bias.

![Figure 14. Differential item functioning for HLLs vs. FLLs](image)

7.3.3. Possible rebuttal: Inaccurate definition of heritage language learners

A possible rebuttal to the claim that the Korean C-test scores can be generalized to different types of learners such as HLLs and FLLs could be the inaccurate classification of learners into the heritage and foreign language learner groups. Although this classification was based on the definition of HLL presented in the heritage learner literature (Kagan, 2005; Valdés, 2005), it was crucial to examine whether this definition
could be empirically supported. That is, it was important to explore whether other possibilities for grouping learners would be more accurate given that this definition depends on a complex series of factors that interact with one another, including family background, Korean language use, and Korean language learning experience. For this reason, a hierarchical cluster analysis (HCA) was conducted on the data collected through the background questionnaire in order to examine whether target variables would identify new classifications of learners. Then, the items on the C-test were evaluated again for DIF based on the new groupings.

Cluster analysis is a useful exploratory statistical tool that allows researchers to group cases (or learners) based on variables such as linguistic and family background, attitudes, proficiency, etc. In other words, groups can be identified in a bottom-up fashion according to these variables. This analysis is especially useful when, as is the case in many of the heritage language learner studies, “there is evidence to suggest that different subgroups of learners may utilize different pathways to language learning, including different strategies, aptitudes, motivational profiles, […] among other questions relevant to L2 research” (Staples & Biber, 2015, p. 244).

Accordingly, an agglomerative HCA with squared Euclidean distances was conducted using Ward’s method as the distance measure. Unlike other types of cluster analysis, such as K-means cluster analysis, HCA does not require an a priori knowledge of the best number of clusters. In the present study, for example, the number of clusters was not presumed from the outset. Even though two groups, HLLs and FLLs, were initially identified, this classification might not have been as robust as one made based on data specifically collected for this study. Furthermore, the Ward’s method, which is the
most commonly used in the field of L2 research, measures the distances between clusters by looking into the dissimilarities between clusters, and thus was considered more appropriate than other methods such as single linkage (Aldenderfer & Blashfield, 1984; Staples & Biber, 2015). For the current study, it was important to explore which learners could be associated with each other but also which groups of learners were dissimilar to others.

### 7.3.3.1. Cluster analysis methodology

Before conducting the cluster analysis, it was important to examine whether the data had multicollinear variables, or variables with correlation coefficients equal or higher to .90 with each other (Jeon, 2015; Mooi & Sarstedt, 2011; Tabachnick & Fidell, 2013). Cluster analysis is sensitive to multicollinearity given that if, for instance, two variables are highly correlated, it would be difficult to determine which variable is contributing in the identification of the clusters. Moreover, due to the great similarity between these variables, some of their characteristics might be overrepresented in the cluster analysis (Mooi & Sarstedt, 2011). Seventeen variables were examined first for multicollinearity, and it was found that seven of the variables demonstrated multicollinearity. Therefore, some of these variables were either eliminated or combined into a single variable. In the case of variables such as frequency of Korean use when (a) ‘your mother talks to you’ and (b) ‘your father talks to you’ together with frequency of Korean use when (c) ‘you talk to your mother’ and (d) ‘you talk to your father’ were highly correlated to each other. Thus, it was decided to combine these variables into a single variable of ‘mean of Korean use to and from parents’ (i.e., Parents Mean). After combining or eliminating variables, a total of 10 variables were selected (see Table 20).
Table 20

*Variables Considered for Cluster Analysis*

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Measurement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Oral Proficiency</td>
<td>ACTFL OPIc scores</td>
<td>ACTFL OPIc scores were used as indicators of oral proficiency</td>
</tr>
<tr>
<td>2. Writing Proficiency</td>
<td>ACTFL WPT scores</td>
<td>ACTFL WPT scores were used as indicators of writing proficiency</td>
</tr>
<tr>
<td>3. Study in Korea</td>
<td>Weeks of stay in Korea</td>
<td>How long participants studied Korean in Korea.</td>
</tr>
<tr>
<td>4. Parents</td>
<td>Mean of four items</td>
<td>How often participants speak and are spoken to in Korean by their mother and father.</td>
</tr>
<tr>
<td>5. Siblings</td>
<td>Mean of two items</td>
<td>How often participants speak and are spoken to by their siblings.</td>
</tr>
<tr>
<td>6. Grandparents</td>
<td>Mean of two items</td>
<td>How often participants speak and are spoken to by their grandmother and grandfather.</td>
</tr>
<tr>
<td>7. Relatives</td>
<td>One item</td>
<td>How often participants speak and are spoken to by their relatives.</td>
</tr>
<tr>
<td>8. Computer Use</td>
<td>Mean of seven items</td>
<td>How often participants use Korean to do a series of activities on the computer (social media) and phone</td>
</tr>
<tr>
<td>9. Watch/Listen</td>
<td>Mean of five items</td>
<td>How often participants use Korean to watch and listen to TV shows, dramas, movies, news, and music.</td>
</tr>
<tr>
<td>10. Read</td>
<td>Mean of six items</td>
<td>How often participants use Korean to read newspapers, internet news, novels, magazines, emails, and internet sites.</td>
</tr>
</tbody>
</table>

It is worth noting that the selection of these variables was also theoretically motivated. More specifically, previous studies found that HLLs tend to be more proficient in speaking and listening when compared to reading and writing (Friedman & Kagan, 2008; Kondo-Brown, 2003, 2004; Montrul, 2008). It was important then, to include items that would represent this type of learner characteristic. In addition, as Kondo-Brown (2005) noted, HLLs vary greatly in terms of their HL use with the family,
which in turn has great influence on how proficient HLLs are in their HL. Hence, the items in the questionnaire related to Korean language use with the family were extremely relevant for the purposes of this analysis.

For running the HCA, z-scores were used to standardize the measurements of all 10 variables. In terms of missing data, several participants were contacted again in order to complete parts of the questionnaire that had either not been clearly answered or were overlooked. As a result, only one learner was excluded from the analysis due to missing data (e.g., responses to questions like “how often do your parents speak to you in Korean?” or “how often do you speak to your parents in Korean?”). In addition, it is important to point out that the outliers identified by the Rasch analysis ($N=5$) were not excluded when conducting HCA, as they did not introduce noise to this particular analysis. In the end, data for the HCA consisted of 92 learners of Korean from various Korean language learning backgrounds, family backgrounds, and Korean language use.

7.3.3.2. Cluster analysis results

Results from the HCA identified four different clusters of learners in the data (see Appendix G for the dendrogram). The cluster solution derived from the difference between agglomeration coefficients at each stage. As illustrated in Figure 15 after four clusters the difference in coefficients started to flatten out, which meant that not much new information was gained by adding more clusters. All cluster solutions, namely the two-, three-, and four-cluster solutions were analyzed in terms of their relationship with the previous classification of HLLs vs. FLLs as well as the descriptive statistics of each of the ten variables. It was found that the four-cluster solution was the most efficient and informative.
Table 21 shows the crosstabulations between the results of the HCA, considering the two-, three-, and four-cluster solutions, and the prior classification of learners (HLLs vs FLLs). As can be seen, this procedure revealed that some HLLs who had been grouped into the HLL group were actually better represented in the FLL group (Table 21). More specifically, in the two-cluster solution, six of the initially considered HLLs were clustered into the group composed mostly of FLLs. In fact, the same six learners were always clustered with a mainly-FLL-clustered group in all of the cluster solutions.

Another important finding was that while FLLs were mostly clustered with each other as observed in the two- and three-cluster solution ($N = 52$), HLLs were either grouped with FLLs or separated into more than two groups. Interestingly, the six learners belonging to Cluster 4 in the four-cluster solution or Cluster 3 in the three-cluster solution were always clustered separately within the HLL group and were separated earlier in the hierarchy as displayed in the dendogram (Appendix G).

Figure 15. Distance between agglomeration coefficients at each cluster stage
Table 21

*Crosstabulation of Learner Types and Cluster Groups*

<table>
<thead>
<tr>
<th>Learner Type</th>
<th>HLL</th>
<th>FLL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-Cluster Solution 1</td>
<td>34</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>Two-Cluster Solution 2</td>
<td>6</td>
<td>52</td>
<td>58</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>52</td>
<td>92</td>
</tr>
<tr>
<td>Three-Cluster Solution 1</td>
<td>28</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Three-Cluster Solution 2</td>
<td>6</td>
<td>52</td>
<td>58</td>
</tr>
<tr>
<td>Three-Cluster Solution 3</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>52</td>
<td>92</td>
</tr>
<tr>
<td>Four-Cluster Solution 1</td>
<td>28</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Four-Cluster Solution 2</td>
<td>0</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Four-Cluster Solution 3</td>
<td>6</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>Four-Cluster Solution 4</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>52</td>
<td>92</td>
</tr>
</tbody>
</table>

In order to explore the underlying characteristics of the learners within each cluster, descriptive statistics for the ten variables were closely examined across all clusters for the four-cluster solution (see Appendix H) and these are represented in the 95% CI error bar plots in Figure 16 through Figure 19. The first error bar plot (Figure 16) presents a comparison of the performance on ACTFL tests (OPIc and WPT) among the four clusters. The next graph (Figure 17) shows the differences across the four clusters in terms of their length of stay (in weeks) in Korea learning Korean. Then, Figure 18 presents the four-cluster differences in the frequency of Korean language use between learners and different family members (i.e., parents, siblings, relatives, and grandparents). Finally, Figure 19 displays the amount of Korean language use by learners in the four clusters when performing different tasks, such as when using the computer, watching TV shows, reading novels, etc.
Figure 16. ACTFL test scores comparison across four clusters

Figure 17. Length of stay studying in Korea across four clusters
Figure 18. Learners’ Korean language use with family members across four clusters

Figure 19. Learners’ frequency of Korean language use across four clusters
The 10 variables explored in this section served as indicators for the most distinctive properties for each of the clusters. Cluster 1 was characterized by 28 learners who seem to fit one of the typical heritage language learner profiles present in the SLA literature. They used Korean language mostly with some family members but less in other contexts. This group of learners had an overall Intermediate Mid to Intermediate High level of speaking and writing performance. They appeared to have slightly better speaking skills than writing skills. Furthermore, these learners had very little experience studying in Korea with an average of only 2 weeks of stay in Korea. In terms of Korean language use with the family, this group of learners indicated frequent use of Korean to talk to parents and relatives but less to grandparents and almost never to siblings. In regard to the use of Korean for different tasks, these learners reported some use when watching or listening to Korean TV shows, dramas, movies, news, and music. Nevertheless, compared to the other clusters, they reported using Korean in this context with less frequency. Although there appeared to be some use of Korean for emailing and social media, it was very minimal. Particularly, they reported almost never reading in Korean. Accordingly, most of the learners in Cluster 1 could be classified as HLLs who use Korean language at home but who do it less in other contexts.

Cluster 2 was composed of 29 learners who seem to fit one of the typical foreign language learner profiles present in the SLA literature. They reported using Korean with frequency in most of the contexts but having no Korean family connections. Like Cluster 1 learners, they appeared to perform at an Intermediate Mid to Intermediate High level of Korean speaking and writing. However, these learners appeared to perform slightly better at writing than speaking. They indicated having some experience studying abroad in
Korea with an average of about 22 weeks of stay in Korea. As mentioned before, these learners indicated almost no use of Korean with the family. However, they appeared to watch and listen to Korean entertainment frequently and to use Korean for some computer-related tasks. Nevertheless, as with Cluster 1 learners, they reported lower frequency of Korean language use for reading than for other activities. In brief, this cluster appeared to describe FLLs with high enough levels of motivation to have studied abroad and to want to try and use the Korean language in different contexts, especially for watching and listening.

Cluster 3 was composed of 29 learners with very little use of Korean language across all contexts. This cluster was comprised of almost only FLLs and six learners who had been recruited as HLLs. Their scores on the ACTFL OPIc and WPT indicated that most of the learners in this group performed at the ACTFL Novice Mid to Novice High levels in speaking and writing. They also reported having minimal experience formally studying Korean in Korea with an average of 2 weeks. Nevertheless, there was large variability among learners in the group in terms of their ACTFL test scores (which could go as high as Advanced-High and as low as Novice-High ACTFL proficiency level) and the length of stay studying in Korea (ranging from zero to 26 weeks). Overall, these learners indicated almost no frequency in using Korean with family members. There were, however, some learners who indicated very minimal use of Korean particularly with their parents. These learners were identified as the six learners who had initially been classified as HLLs. Upon closer examination of their responses to the background questionnaire, these learners reported using Korean to speak to their parents with some frequency but they were either just spoken to in Korean, they spoke Korean to only one
parent, or spoke Korean very minimally with both parents. Furthermore, these learners used almost no Korean with other family members, including, siblings and grandparents. In terms of their Korean language use in other contexts, learners in this cluster showed high frequency of use when it came to watching and listening but less so for social media and computer use. As learners from the other clusters, very little frequency was reported for reading in Korean. In sum, this cluster was composed mainly of FLLs who mostly used Korean to watch and listen to Korean entertainment. Even though some HLLs fell in this group, they used Korean language very minimally with their family members and thus behaved more similarly to FLLs than to HLLs.

Cluster 4 consisted of 6 learners who all had Korean in their family heritage. They were highly-literate Korean language learners who often used Korean in all contexts. These learners were different from all other learners of Korean as they performed at the highest levels in the ACTFL OPIc and WPT. Furthermore, these were learners who spent a considerable amount of time receiving formal education in Korea with a range of approximately 2 to 9 years. They all reported being born in Korea and their arrival age in the US ranged between 7 to 15 years old. In addition to their high proficiency in Korean, they also appeared to use the HL frequently with their family members, including their siblings, as well as in other contexts such as for social media, computer use, and watching and listening to TV shows, dramas, etc. Interestingly, although still less frequent than other activities, these group of learners reported reading in Korean with some frequency and more than learners from the other clusters. Overall, these learners appeared to be HLLs who had very strong ties to Korean language through family connection but also through being immersed in the Korean culture.
7.3.3.3. Descriptive statistics, reliability measures, and Differential Item Functioning (DIF) analysis on the new clusters

Bringing the HCA results back to the interpretive argument, descriptive statistics, reliability estimates, and DIF were analyzed again with the new groupings identified in the HCA in order to revisit some of the claims in the generalization inference. In addition to exploring what types of learners were represented in the data, the results of the HCA were also important in assessing how different types of learners performed on the C-test and whether C-test items showed appropriate reliability estimates for each of type of learner. Furthermore, as there was a new classification of learner types, DIF analysis was conducted again to evaluate whether any noticeable biases could be found that affected a particular group of learners.

For the descriptive statistics and reliability measures, both the two-cluster and four-cluster solution were explored separately. Although the four-cluster solution was found to be the most informative in terms of describing different learner types, the two-cluster solution was also explored in order to compare results between the prior and new grouping of HLLs vs. FLLs. As a reminder, the two-cluster solution was composed of (1) Cluster 1 ($N = 34$), which was a combination of Cluster 1 and Cluster 4 in the four-cluster solution and (2) Cluster 2 ($N = 58$), which was a combination of Cluster 2 and Cluster 3 also in the four-cluster solution (see previous section). In other words, the main difference between the two groups was that while Cluster 1 was characterized by learners who used Korean with most of their family members with frequency (i.e., all being HLLs), Cluster 2 was composed of learners who had no or very minimal interactions with family members in Korean (i.e., mostly FLLs but also six HLLs). Unless otherwise
indicated, all future analyses concerning a comparison between heritage and foreign language learners will be conducted on the two-cluster solution. Table 22 presents the results of the descriptive statistics and reliability measures conducted on C-test scores for these two groups.

Table 22

<table>
<thead>
<tr>
<th></th>
<th>Two-Cluster Solution</th>
<th></th>
<th>Four-Cluster Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1 (^a) ((N = 34))</td>
<td>C2 ((N = 58))</td>
<td>C1 ((N = 28))</td>
</tr>
<tr>
<td>Mean</td>
<td>66.59</td>
<td>51.10</td>
<td>60.61</td>
</tr>
<tr>
<td>SD</td>
<td>19.76</td>
<td>20.22</td>
<td>16.26</td>
</tr>
<tr>
<td>Min</td>
<td>23.00</td>
<td>12.00</td>
<td>23.00</td>
</tr>
<tr>
<td>Max</td>
<td>97.00</td>
<td>91.00</td>
<td>85.00</td>
</tr>
<tr>
<td>SEM</td>
<td>±7.19</td>
<td>±5.32</td>
<td>±6.30</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.92</td>
<td>0.95</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Notes. \(^a\)C1 = Cluster 1, C2 = Cluster 2, C3 = Cluster 3, C4 = Cluster 4; \(^b\)The maximum possible score for four items was 100.

As with the prior HLL vs. FLL classification of learners, the two-cluster solution indicated that the C-test was able to elicit a wide range of scores from different populations of learners, with a score point range of 23 to 97 for Cluster 1 and 12 to 91 for Cluster 2. The variation of performance among learners within each cluster was very similar with a standard deviation of 19.76 and 20.22. Learners in Cluster 1, mainly composed of HLLs, performed better than those in Cluster 2, mostly composed of FLLs. Furthermore, even though the HCA did not include C-test performance as one of the variables for clustering learners, descriptive statistics for C-test scores described learners in similar ways as the HCA. In other words, in the four-cluster solution learners from Cluster 4 (i.e., highly literate HLLs) performed exceptionally better on the C-test than learners in the other clusters. It also worth noting that the standard deviations of C-test scores
scores for each of the subgroups in the four-cluster solution were smaller than those in
the two-cluster solution, suggesting that there was less variability in performance among
learners in each cluster.

In addition to examining the descriptive statistics, an independent samples t-test
was also conducted to determine whether there were any statistically significant
differences in C-test performance between the two clusters in the two-cluster solution.
Results showed a statistically significant difference between Cluster 1 (N = 34) and
Cluster 2 (N = 58) with \( t(90) = 3.58, p = .00; d = .92 \). Nevertheless, as with the
results of the comparison between HLLs and FLLs in the evaluation inference (see
section 7.2.2), these results need to be interpreted with caution. First, the sample sizes per
cluster were rather small. Second, upon closer inspection of the distribution of the C-test
performance for both clusters, it was found that Cluster 1 was negatively skewed and
Cluster 2 was positively skewed. One issue with the current data was the very small
number of HLLs who were Novice-level learners, on the one hand, and FLLs who were
highly literate, on the other. In other words, it was difficult to make a direct comparison
of C-test performance between Cluster 1 and Cluster 2 (in the two-cluster solution) as the
learners who composed these clusters differed in language proficiency levels. Thus, to
make a more accurate comparison between clusters, an independent samples t-test was
conducted again only between learners identified as Cluster 1 (made up of 29 HLLs) and
Cluster 2 (made up of 28 FLLs) in the four-cluster solution, as they both showed average
intermediate levels of ACTFL proficiency (i.e., Intermediate Low to Intermediate High).
The results of the new independent samples t-test indicated that there was no statistically
significant difference in the C-test performance between Cluster 1 and Cluster 2 with 
\[ t(55) = 1.20, p = .24; d = .32. \]

Furthermore, very high reliability estimates were found for most of the clusters in both two- and four-cluster solution, which indicated that the C-test items were reliably distinguishing test-takers’ ability levels even within different types of learner populations. As expected given the very small sample size of six, the reliability coefficient for Cluster 4 was lower than that of the other subgroups (\( \alpha = .53 \)). However, rather than reflecting a low reliability of C-test items, these results together with the descriptive statistics indicated that the learners within this subgroup were very few and moreover varied very minimally in language ability levels. The range of C-test scores for this group was 87 to 97 with a standard deviation of 3.78. In other words, as these learners were so few and varied minimally in their proficiency levels, the C-test was unable to reliably distinguish their abilities as there was almost no room for differentiation. In sum, C-test items were able to reliably assess different types of learners with a wide range of ability levels.

Another important claim of the generalization inference that needed revisiting was the possibility of C-test item bias on a given learner type. Therefore, DIF analysis was run on the two-cluster solution. The four-cluster solution was not considered for this analysis because the sample sizes of the subgroups was smaller than 30. Linacre (n.p.)\textsuperscript{5} stated that “to compare reasonably robust item difficulties for each sub-group in a somewhat informal way, then 30 per group-classification” is recommended. Usually, DIF analyses are conducted on sample sizes of 200 or 300 per group for adequate power and

\textsuperscript{5} http://www.winsteps.com/winman/difconcepts.htm
to reduce Type I error rates. Thus, it was important to keep subgroups as large as possible given the overall sample size of 93. Moreover, given that the purpose of using DIF analysis was to examine whether the C-test items were biased towards a specific learner group, the two-cluster solution was considered as informative as the four-cluster solution.

The results of the DIF analysis found no noticeable DIF contrast when comparing Cluster 1 and Cluster 2. Like the previous DIF analysis in section 7.3.2, the four items on the C-test behaved in similar ways across the new groups (see Figure 20). The DIF contrast (or the effect size in logits) ranged between 0.08 and 0.42. As mentioned before, a DIF contrast big enough to be noticeable is usually more than 0.50 logits. Accordingly, the observed scores of all four items on the C-test appeared to consistently reflect the language abilities of both Korean heritage and foreign language learners with no noticeable bias. At the same time, these results also suggested that the difference in C-test performance between these two learner groups, as indicated by the independent t-test, was not due to the C-test being biased toward one learner group or the other, but rather was because the sample of learners in the clusters differed in their language proficiency.

*Figure 20. Differential item functioning for Cluster 1 vs. Cluster 2*
7.3.4. Summary of the outcomes relevant to the generalization inference

Four different claims were investigated for the generalization inference: (1) the C-test can provide reliable and consistent estimates of test-takers’ performances; (2) the configuration of four items on the C-test is the most appropriate for providing reliable estimates of test-takers’ performances; (3) the number of tasks on the C-test is sufficient to provide stable estimates of test-takers’ performances; and (4) the scores on the C-test can consistently reflect the language abilities of both Korean heritage and foreign language learners without biases.

For the first claim, reliability estimates were calculated for the C-test. Overall, Cronbach’s alpha coefficients were .94 and above, indicating that the test as a whole was highly reliable in distinguishing Korean language learners’ abilities. Reliability estimates were also very high when examining different population of learners separately, with $\alpha = .93$ for HLLs and $\alpha = .95$ for FLLs. Slightly higher reliability measures were found for the C-test when it included seven items than when it included four. Nevertheless, the difference in the estimates was very minimal, suggesting that the four-item C-test was similarly providing scores with high reliability (i.e., the lowest reliability coefficient was .91) for both learner populations.

In regard to the second claim, reliability estimates and IRT measures were taken into consideration to evaluate whether including Text 2, Text 3, Text 4, and Text 6 on the C-test was most appropriate for assessing Korean language learners with a wide range of ability levels. Deciding on the best configuration of items on the C-test depended on two factors, which were that the items be reliable and that they would target different proficiency levels. Therefore, Cronbach’s alpha was calculated for different scenarios
when items were deleted. The reliability estimates with the final four items appeared to be the most appropriate. Even though deleting certain items improved the overall reliability estimate, they were not eliminated without taking into account the IRT measures. For example, although deleting Text 2 would have increased the overall Cronbach’s alpha coefficient, Rasch model analyses suggested that this item was the most appropriate in targeting the low-proficiency level learners. As a result, it was decided to keep this item on the C-test. After examining the reliability estimates, fit statistics, and ICCs for each item, the final four texts (i.e., Text 2, Text 3, Text 4, and Text 6) were found to be the most reliable and efficient in distributing learners from a wide range of ability levels into at least five different proficiency levels.

In terms of the third claim, the overall performance of learners on the C-test was examined in order to evaluate how items behaved and whether the four items that were selected to be in the final C-test version were sufficient to appropriately target all levels of Korean learners represented in the data. As the descriptive statistics and IRT measures showed, there was a wide range of learner language abilities represented in the data and the four items appeared to be sufficient to reliably assess these learners. However, these four items were not sufficient when it came to discriminating learners at the highest proficiency level, that is, almost all of the six HLLs in Cluster 4. For these learners, even the most difficult item on the C-test was not as challenging as it was for other learners. It is, however, critical to reiterate that discriminating learners within the upper ability level was not part of the intended use of the Korean C-test. Rather, this test intended to assess the Korean general language proficiency of learners from a wide range of ability levels,
and thus the four items on the C-test were sufficient and worked appropriately and reliably in doing so.

The fourth claim was based on the assumption that if the C-test scores could be generalized to provide an estimate of global proficiency for all types of learners, then the items on the C-test should behave similarly for all types of learner populations (i.e., HLLs and FLLs). Therefore, a DIF analysis was conducted on the C-test in order to explore whether any biases could be detected on the items included on the C-test for HLLs and FLLs. Findings indicated that no noticeable DIF contrast was detected in any of the items for HLLs and FLLs. All items on the C-test behaved in similar ways for both learner populations.

Finally, following Kane’s (2006, 2011, 2013) validity model, a rebuttal to the argument underlying the generalization inference was also explored in order to assess whether it could weaken the strength of this inference. This rebuttal focused on the classification of learners into the HLL and FLL groups and the extent to which this classification, which was based on definitions found in the literature, was accurate. A hierarchical cluster analysis (HCA) was conducted using participants’ information about their Korean language learning experience, Korean language use in different contexts including with family members, and their speaking and writing proficiencies. The HCA identified two-, three-, and four-cluster solutions for describing different types of learners in the data. The two-cluster solution classified learners in a very similar way to the original classification of HLLs vs. FLLs. Nevertheless, there were some differences in the HLL group. That is, six learners who had originally been selected as HLLs were associated more closely to the FLL group within the HCA. Moreover, the four-cluster
solution was more informative in describing learner types. Briefly, it described four unique learner types: (1) learners who are typically described as HLLs in the SLA literature, who mainly used Korean language with family members but very minimally in other contexts, (2) FLLs who used Korean mainly to watch and listen to Korean, and to use for social media but not at all with family members, (3) mostly FLLs and six HLLs who used Korean language with very minimal frequency in all contexts, and (4) highly-literate HLLs who frequently used Korean in all contexts. Descriptive statistics, reliability estimates, and DIF were analyzed again for the new groupings found in the HCA. Overall, very high reliability estimates were found for the C-test items even within the new grouping of learners, suggesting that the items could reliably distinguish language abilities of distinct learner types. In addition, no noticeable DIF contrasts were detected for any of the clusters in any of the items. Hence, despite some discrepancies in the classification of learners between the initially defined HLL vs. FLL groups and the HCA classification, the C-test items could still reliably and consistently measure general language proficiency for different types of Korean learners. The rebuttal, therefore, did not weaken the strength of the generalization inference.

7.4. Outcomes Relevant to the Explanation Inference

The fourth inference in the Korean C-test interpretive argument was the explanation inference, which was based on the warrant that expected scores on the C-test are attributed to the constructs of general language proficiency. Therefore, this inference mainly focused on the theoretical constructs underlying the C-test and gathered evidence to support (or reject) the claims related to the accuracy of the test in targeting those
constructs. Table 23 presents a summary of the claims and the four research questions that emerged from them.

Table 23

*Explanation Inference Warrant, Assumptions, and Research Questions*

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumptions</th>
<th>Research Questions</th>
</tr>
</thead>
</table>
| Expected scores are attributed to a global construct of Korean language proficiency. | 1. The difficulty of C-test items is systematically influenced by item characteristics regardless of the learner type.  
2. Performance on the Korean C-test relate to performance on other literacy-based measures of language proficiency as expected theoretically.  
3. Performance on the C-test can be predicted by speaking and writing proficiency. | 1. What characteristics of the C-test items appear to influence the difficulty of the items?  
2. Do these item characteristics systematically influence the difficulty of the items regardless of the learner type?  
3. How does the test-takers’ performance on the Korean C-test relate to their performance on other literacy-based measures?  
4. To what extent can speaking and writing language skills predict performance on the C-test? |

7.4.1. **Korean C-test item characteristics and item difficulty**

C-test researchers have used various approaches to examine the difficulty of C-test passages. In English C-tests, for example, the difficulty of the passages has been attributed to the sentence length, the type-token ratio (Klein-Braley, 1984), and syntactic complexity (Dörnyei & Katona, 1992). In a more recent study by Beinborn, Zesch, and Gurevych (2014) multiple characteristics of texts, including average word and sentence length, number of certain parts of speech (POS), number of certain phrase patterns, word familiarity, inflections, and cognates, were associated with the difficulty of the items. These studies suggest possible item characteristics that could influence item difficulty. Nevertheless, given that these studies were focused on the English language, it was
necessary here to consider Korean language-specific characteristics, such as its agglutinative morphology, complex syntax, as well as different word definitions and classifications, when examining possible item characteristics that may affect the difficulty of the Korean C-test passages.

Accordingly, to answer the first and second research questions in the explanation inference, the four selected items were analyzed in terms of their grammatical structure and lexical difficulty. For grammatical structure, sentences in the passage were manually-coded for number of clauses. In terms of lexical difficulty, first, words in a passage were classified in terms of content words or function words as defined by Sohn (1994, 1999). Second, all content words were coded based on their vocabulary level as classified by the National Korean Language Institute (NKLI, 2005) in a list of basic Korean words. This list categorizes Korean words into group A, B, or C according to their difficulty for Korean learners and has been used in several studies (e.g., Kim, et al., 2016) to gauge the level of Korean vocabulary.

Results of the item characteristics statistics showed that items varied in terms of the number of clauses per sentence. From easiest to most difficult, while texts decreased in the number of sentences, the number of clauses increased. Thus, the number of clauses per sentence increased together with the difficulty of the item, as observed by the C-test mean scores. More specifically, as displayed in Table 24, while Text 2, the easiest text, had the most number of sentences and the least number of clauses per sentence, Text 6, the most difficult text, had the least number of sentences and the most number of clauses per sentence.
Table 24

*Number of Clauses per Sentence across All C-test Passages*

<table>
<thead>
<tr>
<th></th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
<th>Text 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test Mean Scores(^a)</td>
<td>19.34(^b)</td>
<td>16.41</td>
<td>11.70</td>
<td>9.83</td>
</tr>
<tr>
<td>C-test Scores SD</td>
<td>5.04</td>
<td>5.28</td>
<td>6.35</td>
<td>6.60</td>
</tr>
<tr>
<td>Number of sentences</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Mean number of clauses/sentence</td>
<td>1.30</td>
<td>1.40</td>
<td>3.38</td>
<td>4.00</td>
</tr>
</tbody>
</table>

*Notes.* \(^a\)C-test mean scores are based on 93 test-takers; \(^b\)Maximum possible score was 25 for each text.

The following two sample sentences from Text 2 and Text 6 provide a more concrete example. While the majority of the sentences in Text 2 were simple sentences containing only one or two clauses, the sentences in Text 6 showed more complexity including up to 6 clauses in one sentence. Similarly, sentences in Text 3 were mostly composed of one or two clauses. Although this text had more sentences that contained 2 clauses when compared to Text 2. Sentences in Text 4 resembled those in Text 6 with more than 2 clauses per sentence, up to 6 clauses in one sentence.

**Text 2 sample sentence:**

Original sentence: [앞으로는 지원이에게 답장을 꼭 쓸게]. 1 clause

Yale romanization: [Aphulo-nun ciwen-i-eykey tapcang-ul kkok ssul-key.]

English Translation: In the future, I will always reply to Jiwon.

**Text 6 sample sentence:**

Original sentence: [최고급 시설로 회사에 오래 머물게 해서] [가정에서 보내야 하는] 시간을 빼앗는 것이 아니나[는 의견이다]. 4 clauses

Yale romanization: [choykokup sisel-lo hoysa-ey olay memwul-key hayse] [kaceng-eyse ponay-ya hanun] [sikan-ul ppayas-nun kesi aninya]-nun uykyen-ita.]
English Translation: This is the type of opinion that questions whether they are taking away people’s time, which should be spent at home, by making them stay longer hours in offices with state-of-the-art facilities.

Lexical features in the texts were also analyzed in terms of word classes and vocabulary level. First, words were defined and classified based on Sohn’s (1994, 1999) word classification into content and function words. As per Sohn, Korean content words include nouns, pronouns, verbs, adjectives, adverbs, determiners, and copulas, whereas function words are comprised of particles (i.e., case particles, delimiters, and conjunctives). It is worth mentioning that the design of the C-test and thus the application of the deletion rule also followed Sohn’s definition of words, as particles were considered words on their own. Table 25 presents the item statistics in terms of the number of content and function words. Given that more types of words belong to the content word class than to the function word class, it was expected that there would be more content words than function words across all four C-test passages. Word counts also revealed that except for Text 3, all texts had more content words than function words deleted. The ratio of deleted content to function words was larger as the texts increased in difficulty. In other words, there were more content words deleted when texts were more difficult. The only exception was Text 3, which had more function words deleted than Text 2.
Another item characteristic explored was the level of difficulty of content words. As mentioned earlier, the NKLI basic word list for Korean learners was used to classify words into different difficulty levels A, B, and C, with A being the easiest words and C the most difficult ones. There were some content words that were not found in the list and those included copula (affirmative copula 이다, and negative copula 아니다), as well as proper nouns (e.g., Google, Facebook), and low-frequency sino-Korean words (e.g., 족쇄 or handcuffs). Those words were excluded from the count. Figure 21 represents the counts of words that belonged to the different levels of vocabulary across texts. As can be observed in the graph, as the item difficulty increased so did the number of words belonging to Level B and C. On the other hand, the number of words of Level A decreased as items were more difficult. Text 2, for example, contained only one word belonging to level C, whereas Text 6 had 14 words.
Another important piece of evidence that would corroborate the relationship between the item characteristics and item difficulty was the analysis of the individual gap difficulty. Accordingly, Rasch model IRT analyses were used to examine C-test scores for each micro-item and map all items into a single underlying logit scale (Figure 22). Even though the individual analysis of gaps is not recommended for C-tests as all micro-items are presumed to be interdependent, this analysis was considered to be useful for providing an overview of which gaps were the easiest and most difficult to complete. This new analysis yielded 10 misfitting persons, which were removed from the micro-item IRT analysis. As observed in Figure 22, there was a progression of difficulty according to the overall ranking of the polytomous items. That is, the micro-items (i.e., gaps) in the upper end of item map, which were the most difficult items, belonged mostly to Text 6, whereas the micro-items in the lower end of the item map, which were the easiest items, belonged mostly to Text 2. This text difficulty progression also helped

Figure 21. Word count by vocabulary level
illustrate that indeed C-test items are interdependent and that the solution to the majority of micro-items depends on an overall comprehension of the passage as a whole.

In addition, the 10 most difficult and the 10 easiest micro-items were further examined to explore what type of syntactic or morphological characteristics they were targeting. As observed in Table 26, the most difficult items included more complex verb inflections and connective endings (i.e., morphemes used to connect clauses or

![Figure 22. Item map for all micro-items in the C-test (N = 83)](image_url)
sentences), as well as content words belonging to category B or C in the NIKL vocabulary list. In the case of items like 6.24, the word 오래 (yale romanization: olay; meaning: “long time”) was considered difficult even though it belongs to category A in the NIKL list, because it was followed by the word 머물다 (yale romanization: memwulta; meaning: “stay”), which is categorized as belonging to level C. Likewise, item 4.05, the most difficult item, consisted of the word 어려서 (yale romanization: elyese; meaning: “being young”), which belongs to level A. This word was part of the construct 어려서부터 (yale romanization: elyese pwuthe; meaning: “since being young”), which was confused by many learners as 어릴 때부터 (yale romanization: elil ttay pwuthe), another construct that has the same meaning. The latter construct is incorrect in this context due to spacing rules. In other words, learners needed to process not only the meaning and morphology of the item but also the spacing rules in order to provide the correct answer.

On the other hand, the 10 easiest micro-items included content words from level A that had already been provided as undeleted stimuli, as well as common sentence final endings (e.g., -어요 yale romanization: -eyo; prefinal ending and polite sentence ending) and subject and object case markers. There were many repeated micro-items for Text 2, which made items easier for all test-takers.
### Table 26

**Ten Most Difficult and Easiest Micro-Items**

<table>
<thead>
<tr>
<th>Item number</th>
<th>Item</th>
<th>Meaning</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.05</td>
<td>어려서 a</td>
<td>being young</td>
<td>Level A b adjective + inflection from the grammatical construct 어려서부터</td>
</tr>
<tr>
<td>6.21</td>
<td>거세다</td>
<td>strong (opinion)</td>
<td>Level C adjective</td>
</tr>
<tr>
<td>4.11</td>
<td>용돈의</td>
<td>money + genitive case marker</td>
<td>Level A noun + genitive case marker</td>
</tr>
<tr>
<td>6.02</td>
<td>자체</td>
<td>the matter itself</td>
<td>Level B noun</td>
</tr>
<tr>
<td>6.15</td>
<td>효율</td>
<td>efficiency</td>
<td>Level C noun</td>
</tr>
<tr>
<td>4.21</td>
<td>사용할지</td>
<td>(if) to use</td>
<td>Level B noun + verb derivational suffix + connective ending</td>
</tr>
<tr>
<td>6.16</td>
<td>높이는</td>
<td>increasing</td>
<td>Level B verb + verb inflection</td>
</tr>
<tr>
<td>6.24</td>
<td>오래</td>
<td>long time</td>
<td>Level A adverb</td>
</tr>
<tr>
<td>4.12</td>
<td>대부분을</td>
<td>most + object marker</td>
<td>Level B noun + object marker</td>
</tr>
<tr>
<td>6.10</td>
<td>설치했다</td>
<td>installed</td>
<td>Level C noun + verb derivational suffix</td>
</tr>
<tr>
<td>2.01</td>
<td>엄마</td>
<td>mother</td>
<td>Level A noun</td>
</tr>
<tr>
<td>2.16</td>
<td>엄마는</td>
<td>mother + subject marker</td>
<td>Level A noun + subject marker</td>
</tr>
<tr>
<td>2.24</td>
<td>엄마</td>
<td>mother</td>
<td>Level A noun + subject marker</td>
</tr>
<tr>
<td>3.13</td>
<td>있어요</td>
<td>there is</td>
<td>Level A verb + polite sentence ending</td>
</tr>
<tr>
<td>3.20</td>
<td>많이</td>
<td>a lot</td>
<td>Level A adverb</td>
</tr>
<tr>
<td>3.04</td>
<td>수도있어요</td>
<td>was the capital</td>
<td>Level A noun + copula + prefinal ending + polite sentence ending</td>
</tr>
<tr>
<td>2.03</td>
<td>편지를</td>
<td>letter + object marker</td>
<td>Level A noun + object marker</td>
</tr>
<tr>
<td>2.21</td>
<td>편지</td>
<td>letter</td>
<td>Level A noun</td>
</tr>
</tbody>
</table>

**Notes.** a The underlined syllables indicate the gaps in the C-test; vocabulary levels are based on the list by NKLI (2005). 

In brief, item characteristics appeared to systematically influence the difficulty of the items. Results indicated that as item difficulty increased so did the number of clauses.
per sentence, content words deleted, and words belonging to level C of the vocabulary list. In other words, the item difficulty was associated with sentence complexity, word class, and vocabulary level.

The second research question in the explanation inference referred to the extent to which item characteristics systematically influence the difficulty of the items regardless of the learner type. Even though the HLL group appeared to do slightly better in all texts when compared to the FLL group, the ranking in difficulty of the (polytomous) items was the same for both groups (see Table 19). Therefore, the item characteristics examined thus far, namely the number of clauses per sentence, number of content vs. function words within each text, and the number of words belonging to different vocabulary levels, appear to have influenced the performance of both learner groups (i.e., the two groups identified in the two-cluster solution in the HCA) in similar ways.

In order to further examine whether there was a difference between learner groups in the way they perceived specific micro-items, the difficulty of all 100 micro-items (i.e., 25 items x 4 texts) was analyzed for each learner group (i.e., in the two clusters found through HCA) separately (Figure 23). The comparison of the item maps across learner groups suggested that, as when analyzing both learner groups together, the texts progressed in difficulty in similar ways. That is, for both learner groups, the majority of items at the upper end of the item map were part of Text 6, whereas the majority of the items at the lower end of the item map were part of Text 2. Furthermore, when closely examining the most difficult and easiest items, there appeared to be general agreement between both groups. As the number of test-takers for each learner group was different, the direct comparison of items based on logit scales would not be recommended. Hence,
it can be concluded from the broader trends of the difficulty of gaps showed that the item characteristics influenced item difficulty in similar ways for both learner groups.

<table>
<thead>
<tr>
<th>MEASURE PERSON - MAP - ITEM</th>
<th>MEASURE PERSON - MAP - ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>X</td>
<td>4.05</td>
</tr>
<tr>
<td>T</td>
<td>6.21</td>
</tr>
<tr>
<td>XX</td>
<td>4.11</td>
</tr>
<tr>
<td>4</td>
<td>6.14</td>
</tr>
<tr>
<td>X</td>
<td>6.19</td>
</tr>
<tr>
<td>3</td>
<td>6.24</td>
</tr>
<tr>
<td>XX</td>
<td>6.13</td>
</tr>
<tr>
<td>X</td>
<td>5.18</td>
</tr>
<tr>
<td>2</td>
<td>6.17</td>
</tr>
<tr>
<td>XX</td>
<td>6.19</td>
</tr>
<tr>
<td>X</td>
<td>4.18</td>
</tr>
<tr>
<td>1</td>
<td>6.08</td>
</tr>
<tr>
<td>XX</td>
<td>6.11</td>
</tr>
<tr>
<td>X</td>
<td>4.03</td>
</tr>
<tr>
<td>0</td>
<td>6.12</td>
</tr>
<tr>
<td>XX</td>
<td>4.01</td>
</tr>
<tr>
<td>X</td>
<td>2.20</td>
</tr>
<tr>
<td>8</td>
<td>3.10</td>
</tr>
<tr>
<td>XX</td>
<td>4.07</td>
</tr>
<tr>
<td>X</td>
<td>2.00</td>
</tr>
<tr>
<td>7</td>
<td>3.15</td>
</tr>
<tr>
<td>XX</td>
<td>2.19</td>
</tr>
<tr>
<td>X</td>
<td>3.15</td>
</tr>
<tr>
<td>6</td>
<td>3.12</td>
</tr>
<tr>
<td>XX</td>
<td>3.21</td>
</tr>
<tr>
<td>X</td>
<td>2.18</td>
</tr>
<tr>
<td>5</td>
<td>3.15</td>
</tr>
<tr>
<td>XX</td>
<td>2.15</td>
</tr>
<tr>
<td>X</td>
<td>2.14</td>
</tr>
<tr>
<td>4</td>
<td>3.12</td>
</tr>
<tr>
<td>XX</td>
<td>2.08</td>
</tr>
<tr>
<td>X</td>
<td>2.10</td>
</tr>
<tr>
<td>3</td>
<td>3.10</td>
</tr>
<tr>
<td>XX</td>
<td>2.06</td>
</tr>
<tr>
<td>X</td>
<td>2.08</td>
</tr>
<tr>
<td>2</td>
<td>3.10</td>
</tr>
<tr>
<td>XX</td>
<td>2.06</td>
</tr>
<tr>
<td>X</td>
<td>2.10</td>
</tr>
<tr>
<td>1</td>
<td>3.08</td>
</tr>
<tr>
<td>XX</td>
<td>2.07</td>
</tr>
<tr>
<td>X</td>
<td>2.04</td>
</tr>
<tr>
<td>0</td>
<td>3.03</td>
</tr>
<tr>
<td>XX</td>
<td>2.00</td>
</tr>
<tr>
<td>X</td>
<td>2.05</td>
</tr>
<tr>
<td>-1</td>
<td>3.02</td>
</tr>
<tr>
<td>XX</td>
<td>2.00</td>
</tr>
<tr>
<td>X</td>
<td>2.01</td>
</tr>
<tr>
<td>-2</td>
<td>3.11</td>
</tr>
<tr>
<td>XX</td>
<td>2.00</td>
</tr>
<tr>
<td>X</td>
<td>2.05</td>
</tr>
<tr>
<td>-3</td>
<td>3.17</td>
</tr>
<tr>
<td>XX</td>
<td>2.00</td>
</tr>
<tr>
<td>X</td>
<td>2.09</td>
</tr>
<tr>
<td>-4</td>
<td>3.20</td>
</tr>
<tr>
<td>XX</td>
<td>2.00</td>
</tr>
<tr>
<td>X</td>
<td>2.01</td>
</tr>
</tbody>
</table>

HLLs Micro-Item Difficulty (N = 31)

FLLs Micro-Item Difficulty (N = 56)

Figure 23. Comparison of micro-item difficulty between HLLs and FLLs

In addition, another method for comparing the perceived level of item difficulty between both learner groups was by computing Spearman rho correlation coefficients. The IRT item measures for all 100 micro-items were examined in terms of the difficulty ranking for each learner group and a rank correlation coefficient was calculated between these measures. The results of the correlation analysis showed a strong positive correlation with $\rho = .88$, further backing the previous analysis of the item difficulty by
learner group. In other words, it was found that the difficulty ranking of the micro-items was similar for both learner groups.

7.4.2. **Relationship between Korean C-test and other proficiency measures**

If the expected scores on the C-test could be attributed to test-takers’ general language proficiency, then a strong relationship between Korean C-test performance and other measures of Korean language proficiency should be expected. Therefore, the third and fourth research questions in the explanation inference focused on, first, the extent of the relationship between Korean C-test performance and other literacy-based measures and, next, the predictability of C-test performance by other measures of proficiency.

To answer the third research question, correlation coefficients were calculated between C-test scores and ACTFL WPT scores. As the ACTFL test provides a measurement of writing proficiency, it was considered an appropriate measurement of literacy. Unfortunately, no measurement for reading was available for the purposes of this study. It is recommended that future studies also gather this information for more robust results on the relationship of C-test scores and literacy-based measures.

Before examining the correlation coefficients, it was important to explore the descriptive statistics of the ACTFL WPT. Table 27 presents test-takers’ performance on the test based on the numerical equivalent of ACTFL proficiency scores (see Vande Berg, Connor-Linton, & Paige, 2009; Appendix I). HLLs were found to perform better overall than the FLLs. Nevertheless, there was more variability within the former learner group, with larger standard deviations. In addition, it is worth noting that the mean differences between the HLL and FLL group appeared larger because of two HLLs who reached the Superior ACTFL proficiency level (i.e., 3.0). One of these learners indicated that she had
received formal education in Korea until middle school. The other learner indicated that
she had attended four years of elementary school in Korea, although that these schools
were American schools. However, she also reported having attended one year of
undergraduate education at a university in Korea.

Table 27

\textit{ACTFL WPT Descriptive Statistics}

<table>
<thead>
<tr>
<th>Learner Type</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>All (N = 93)</td>
<td>1.55</td>
<td>0.53</td>
<td>0.8</td>
<td>3.0</td>
<td>±0.11</td>
</tr>
<tr>
<td>HLL(^a) (N = 34)</td>
<td>1.73</td>
<td>0.57</td>
<td>0.8</td>
<td>3.0</td>
<td>±0.20</td>
</tr>
<tr>
<td>FLL (N = 58)</td>
<td>1.42</td>
<td>0.44</td>
<td>0.8</td>
<td>2.8</td>
<td>±0.12</td>
</tr>
</tbody>
</table>

Notes: The classification of HLLs and FLLs is based on the HCA.

An independent samples t-test was also conducted to determine whether the
differences in ACTFL WPT performance between the HLL- and FLL group were
statistically significant. As with the previous independent samples t-tests comparing these
two groups, results initially showed a statistically significant difference with \( t(90) = 2.94, p = .00; d = .61 \). However, these results must be interpreted with caution as
learners’ performance on the ACTFL WPT were positively skewed and not normally-
distributed for both clusters. For a more accurate comparison between learner groups
with similar proficiency levels, the ACTFL WPT performances for Cluster 1 (\( N = 29 \)
HLLs) and Cluster 2 (\( N = 28 \) FLLs) in the four-cluster solution were examined, and no
statistically significant differences were found with \( t(55) = 1.12, p = .27; d = .29 \).

Before conducting the Pearson correlation analyses, the statistical assumptions
were evaluated. C-test scores and ACTFL WPT scores were first analyzed for extreme
outliers. Jeon (2015) describes extreme cases as those with z-scores equal or larger to
\( ±3.29 \). Therefore, scores on both tests were converted into z-scores and all scores \( ±3.29 \)
were identified as outliers. Following this procedure, no outliers were found. Nevertheless, the boxplot for the ACTFL WPT showed two outliers who had scored 3 (i.e., equivalent to Superior in the ACTFL proficiency guidelines; see Appendix J). Thus, further outlier analyses were conducted with and without these two data points. Again, no noticeable differences were found, and therefore they were not excluded from the analyses. However, it was found that the ACTFL WPT violated the assumption of normality. According to Bachman (2002), “as a rule of thumb, values for skewness and kurtosis of between -2 and +2 indicate a reasonably normal distribution” (p. 74). According to this rule of thumb, scores on both tests were normally distributed. However, Bachman also indicated that another way to test for normality is to examine the ratio between skewness and kurtosis by their respective standard errors. If the value of the ratio is between -2 and +2 then the distribution is considered normal with 95 percent confidence. Based on this measure, the ACTFL WPT scores were not normally distributed. Thus, the interpretation of the correlation coefficients should be accompanied by caution. In addition, non-parametric correlations coefficients (Spearman’s Rho) were also calculated to accompany the results of the Pearson correlations.

Rasch model measurements for the Korean C-test performances were used in the Pearson correlation between the Korean C-test performance and writing proficiency measured through the ACTFL WPT, because, unlike raw scores, they reflect estimates of true intervals of item difficulty and person ability by creating linear measures (Granger, 2008). Results indicated that there was a strong positive correlation between these two variables with $r = .85$ and a Spearman rank order correlation of $\rho = .87$. Although there were some outliers falling outside of the 95% CI range, as observed in the scatterplot in
Figure 24, there was an overall strong positive correlation between both variables. The overall strong correlation coefficient indicated that learners who performed well on the ACTFL WPT tended to also do well on the Korean C-test. In other words, there appeared to be a strong relationship between writing proficiency, which was the literacy-based measurement, and C-test performance.

![Figure 24. Scatterplot of bivariate correlation between ACTFL WPT and C-test scores](image)

To answer the fourth research question in the explanation inference, which addressed the predictability of C-test performance by other proficiency measures, a standard multiple regression analysis was conducted with ACTFL OPIc and ACTFL WPT scores as predictor variables and C-test scores as the dependent or criterion variable. Given the claim that the Korean C-test can provide an estimate of general
language proficiency, it was expected that both speaking and writing proficiency would
be able to predict performance on the C-test.

Data were first screened to evaluate whether they met the assumptions for
conducting multiple regressions. The first assumption was related to the sample size.
Tabachnick and Fidell (2013) recommend following $N \geq 50 + 8m$ (where $m$ is the
number of independent or predictor variables) as one simple rule of thumb for testing
regression. The sample size in the current study was 93 and thus the first assumption was
met. Furthermore, a univariate data screening was conducted to identify whether there
were any univariate outliers that needed to be removed from further analysis. Scores on
all tests (i.e., C-test, ACTFL OPIc, and ACTFL WPT) were converted into z-scores and
no scores were identified as extreme or as equal or larger than $\pm 3.29$. In terms of
normality, linearity, and homoscedasticity, the residual scatterplot (i.e., scatterplots
plotting predicted dependent variable scores and residual scores) was explored (see
Appendix K). The overall shape of the residual scatterplot was rectangular, indicating
that the assumptions of normality, linearity, and homoscedasticity were met. The test of
normality for residuals (i.e., Shapiro-Wilk) also showed no significant results, which
provided further evidence for normality.

Data were additionally screened for multivariate outliers. Mahalanobis distance
was computed for each case and evaluated against the critical value of chi-square.
According to the chi-square table (see Tabachnick & Fidell, 2013, p. 952), the critical
chi-square value for the present study was $16.27$ ($\chi^2[3] = 16.266, p < .001$). None of
the cases were identified as having a Mahalanobis distance equal to or higher than the
chi-square critical value. Data were also examined for multicollinearity. Bivariate
correlation analyses between ACTFL OPIc and ACTFL WPT scores yielded a Pearson
correlation coefficient lower than .90 ($r = .71$), which indicated no multicollinearity.
Given that bivariate correlations could be misleading at times, Tolerance and variance
inflation factors (VIF) were also inspected. As a rule of thumb, a tolerance value lower
than .40 and a VIF higher than 2.50 was considered to indicate multicollinearity (Allison,
1999). No multicollinearity was detected between the predictor variables. In brief, the
data met all assumptions to conduct multiple regressions.

The results of the standard multiple regression indicated that both predictor
variables (i.e., ACTFL OPIc and ACTFL WPT) together were closely related to 80% of
the variance of C-test performance with $R^2 = .80$, $F(2,90) = 181.83, p = .000$ (Table
28). Furthermore, as shown in Table 29, it was found that writing proficiency
significantly predicted C-test performance with $\beta = .57, p = .00$ as did speaking
proficiency with $\beta = .39, p = .00$. Even though it appeared that writing proficiency
mattered slightly more than speaking proficiency when predicting C-test performance, it
is important to note that these results are relative to each other and should not be
interpreted separately. That is, both writing and speaking ability were jointly important in
predicting C-test performance. These results were similar to findings in previous C-test
studies (e.g., Eckes & Grotjhan, 2006; Grotjahn, 1987; Klein-Braley, 1994; Klein-Braley
& Raatz, 1984; Raatz, 1984), which have found that C-test performance can be explained
by multiple language skills, including literacy and oracy skills.
Table 28

*Multiple Regression Model Summary*

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.90</td>
<td>.80</td>
<td>.80</td>
<td>9.75</td>
<td>1.75</td>
</tr>
</tbody>
</table>

*Notes.* $^a$Predictor variables: ACTFL OPlc and ACTFL WPT

Table 29

*Regression Coefficients for the Criterion Variable: C-test Performance*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>95% CI for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>1.64</td>
<td>3.16</td>
<td></td>
</tr>
<tr>
<td>OPlc</td>
<td>13.47</td>
<td>2.33</td>
<td>.39</td>
</tr>
<tr>
<td>WPT</td>
<td>23.32</td>
<td>2.81</td>
<td>.57</td>
</tr>
</tbody>
</table>

In addition to exploring the predictability of writing and speaking proficiency on C-test performance for both learner groups (i.e., HLLs and FLLs) together, it was also important to assess whether the same predictor variables significantly predicted C-test performance for each learner group in the two-cluster solution. Therefore, a standard multiple regression was conducted for the HLL and FLL group separately. There were, however, concerns about the sample size of each learner group ($N = 34$ for HLLs and $N = 58$ for FLLs$^6$). As aforementioned, the recommended number of observations for conducting multiple regressions on two independent variables is 66, which is higher than the sample size for both learner groups, particularly for the HLL group$^7$. Therefore,

---

$^6$ As a reminder one learner was not classified in either of the clusters (see section 7.3.3.1.)

$^7$ Steven (1996) suggests a less conservative rule of thumb for determining the sample size with 15 observations per predictor variable
findings from these analyses must be interpreted with caution. Data were also screened for all other assumptions, that is, univariate outliers, normality, linearity, and homoscedasticity (see scatterplot of residuals in Appendix K), multivariate outliers, and multicollinearity, and no violations were found.

Table 30 displays the results of the standard multiple regressions for both HLLs and FLLs. Findings indicated that writing and speaking proficiency together were closely related to 78% and 77% of the variance in the C-test performance for HLLs and FLLs, respectively, with $R^2 = .78$, $F(2,31) = 53.47, p = .000$ for HLLs and $R^2 = .77$, $F(2,55) = 92.07, p = .000$ for FLLs. Like the previous results of the multiple regressions for the learner population as a whole, both predictor variables together significantly predicted HLLs’ and FLLs’ performance on the C-test to near identical degrees.

Table 30

<table>
<thead>
<tr>
<th>Learner group</th>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLL</td>
<td>1</td>
<td>.88&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.78</td>
<td>.76</td>
<td>9.67</td>
<td>1.83</td>
</tr>
<tr>
<td>FLL</td>
<td>1</td>
<td>.88</td>
<td>.77</td>
<td>.76</td>
<td>9.87</td>
<td>1.65</td>
</tr>
</tbody>
</table>

<sup>a</sup>Predictor variables: ACTFL OPIc and ACTFL WPT

In addition, as displayed in Table 31, for both learner groups writing and speaking proficiency were significantly predictive of Korean C-test performance. However, it appeared that writing proficiency mattered relatively more than speaking for the HLL group when compared to the FLL group. More specifically, for the HLL group writing proficiency significantly predicted C-test score variance with $\beta = .65, p = .00$ more
than speaking proficiency did with $\beta = .29, p = .03$. On the other hand, for the FLL group writing and speaking proficiency was overall equally important in the prediction of C-test score variance with respectively $\beta = .53, p = .00$ and $\beta = .42, p = .00$.

Accordingly, these results suggested that although writing and speaking proficiency were both significantly predictive of C-test performance, depending on the learner type, the degree of importance of these predictor variables differed slightly. While for the HLLs, it appeared that writing proficiency was more important than speaking proficiency in predicting C-test performance, for the FLLs both writing and speaking were almost equally important. However, it is worth noting again that the standardized beta coefficients for each predictor variable should be interpreted relative to each other. That is, the degree to which each predictor variable, in this case speaking and writing, can explain C-test performance should not be interpreted separate to each other. If the regression model would have included only one variable, for example speaking, then its degree of predictability of C-test performance could have been much higher than in the current model, which also includes writing as a predictor variable.

Table 31

<table>
<thead>
<tr>
<th>Learner Group</th>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>$t$</th>
<th>Sig.</th>
<th>95% CI for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>HLL</td>
<td>(Constant)</td>
<td>7.46</td>
<td>6.35</td>
<td>1.17</td>
<td>0.25</td>
<td>-5.50</td>
</tr>
<tr>
<td></td>
<td>OPlc</td>
<td>10.72</td>
<td>4.59</td>
<td>.29</td>
<td>2.34</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>WPT</td>
<td>22.48</td>
<td>4.20</td>
<td>.65</td>
<td>5.35</td>
<td>0.00</td>
</tr>
<tr>
<td>FLL</td>
<td>(Constant)</td>
<td>-1.99</td>
<td>4.43</td>
<td>-0.45</td>
<td>0.66</td>
<td>-10.86</td>
</tr>
<tr>
<td></td>
<td>OPlc</td>
<td>15.71</td>
<td>3.41</td>
<td>.42</td>
<td>4.61</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>WPT</td>
<td>24.29</td>
<td>4.22</td>
<td>.53</td>
<td>5.75</td>
<td>0.00</td>
</tr>
</tbody>
</table>
7.4.3. Possible rebuttal: Measuring heritage language learners with ACTFL proficiency tests

Even though findings from the standard multiple regressions provided backing to the claim that C-test performance can be predicted by speaking and writing proficiency, one rebuttal could be the inappropriate measurement of these language skills. The ACTFL OPIc and ACTFL WPT were used to assess the proficiency of both HLLs and FLLs. Given that the ACTFL tests have been designed with FLLs in mind, using these tests for assessing HLLs has been a controversial issue. On the one hand, several scholars have pointed out that, as the design of the ACTFL tests followed FLLs’ language developmental continuum, they should not be used to assess a different type of learner population like HLLs (Fairclough, 2006; Draper & Hicks, 2000; Valdés, 1989). On the other hand, other scholars have argued that the ACTFL tests are suitable for assessing functional language ability of HLLs in speaking and writing (Alonso, 1997; Kagan, 2005; Kagan & Friedman, 2003; Martin et al., 2013; Swender et al., 2014). ACTFL Proficiency Guidelines are based on different criteria of assessment that are not only relevant to content and accuracy, but also to function. Therefore, the ACTFL tests can be used to assess HLLs if the main goal is to assess how functional these learners are in communicating in their heritage language. Moreover, it is also important to recognize that the ACTFL proficiency guidelines are widely accepted throughout institutions across the United States, making the results easily interpretable for the foreign language instruction and researcher community.

The present study cannot support or reject the rebuttal in the explanation inference since the ACTFL test score data alone was not enough to evaluate test-taker performance.
and rater behavior. Backing for the rebuttal would require in-depth analyses of test-takers’ speech and writing samples together with raters’ performance on each task. It must be acknowledged, however, that in practice the present study sided with the position espoused by Alonso (1997), Kagan (2005), Kagan & Friedman (2003), Martin et al. (2013), and Swender et al. (2014), who all have argued that the use of ACTFL tests must be considered appropriate for assessing both heritage and foreign language learners, given that the purpose of the ACTFL tests is to gauge learners’ functional ability in Korean. Based on the descriptive statistics, it appeared that overall both ACTFL tests, the OPIc and WPT, could accurately distribute learners into a wide range of proficiency levels. An additional argument in support of the use of ACTFL tests in the current study was that it was crucial to use standardized tests that generated scores interpretable by the foreign/second and heritage language learning research community, and tests based on the ACTFL guidelines were considered to be the most appropriate for this purpose. Future studies with access to speaking and writing samples should be able to explore raters’ interpretation of the ACTFL scales vis-à-vis learners’ performance on the tests, and to determine whether ACTFL tests could assess different types of learners equally and without any bias.

7.4.4. Summary of outcomes relevant to the explanation inference

The explanation inference was based on the warrant that expected scores were attributed to a global construct of Korean language proficiency. Three different claims were explored for this inference: (1) C-test item characteristics systematically influence the difficulty of the C-test; (2) performance on the Korean C-test relates to other literacy-
based measures of language proficiency, such as the performance on the ACTFL WPT; and (3) speaking and writing proficiency can predict C-test performance.

In terms of the first claim, the internal structure of the Korean C-test was examined in regard to theoretical views of language ability. In other words, it was important to investigate the characteristics of the C-test items and whether they were systematically influencing the difficulty of the C-test. For this purpose, the items were first analyzed in terms of their sentence structure and morphological and lexical complexity. With respect to sentence structure, it was found that C-test passages consisting of more complex sentences, that is, those with more clauses per sentence, were more difficult than passages with sentences containing only one or two clauses. Furthermore, passages with more content words than function words were generally more difficult. In particular, test-takers found C-test passages with more content words deleted harder than passages with more function words deleted. Additionally, difficulty of the passages was associated with the level of difficulty of the vocabulary used in each passage. According to the classification of words based on the NIKL Korean learner vocabulary list, the difficulty of the C-test passages increased as more words belonging to Category B and C (i.e., most challenging words) were present in the passage. In sum, the internal characteristics of the C-test items affected test-takers’ performance on the C-test. That is, the level of difficulty of C-test items was directly influenced by their syntactic, morphological, and lexical complexity.

Along the same lines, the second claim explored the extent to which C-test item characteristics affected item difficulty in similar ways for each learner group. All 100 micro-items were mapped out in a single logit scale in order to determine which micro-
items were most difficult and easiest for each learner group. Results yielded similar ranks of micro-item difficulty. In other words, the most difficult and easiest micro-items were generally the same when comparing the performances of each learner group. The micro-items considered to be most difficult were part of the most difficult passages (Text 6 and Text 4), while the easiest micro-items belonged to the easier passages (Text 2 and Text 3). A strong positive Spearman rank order correlation coefficient ($\rho = .88$) between item measures for both learner group corroborated the results observed in the Wright map, indicating that the item characteristics affected HLLs and FLLs in similar ways. These findings, at the same time, suggested that the completion of C-test micro-items are interdependent. Although some of the same content and function words were deleted in different passages, their difficulty was dependent on the comprehension of the passage as a whole.

With respect to the third claim, correlation analyses were conducted to evaluate the extent to which C-test performance was related to literacy-based measures. Thus, Pearson and Spearman rank order correlations were estimated between C-test scores and ACTFL WPT scores and results indicated that there was a strong positive correlation between both variables at $r = .85$ and $\rho = .87$. Accordingly, there appeared to be a close relationship between Korean learners’ C-test performance and a literacy-based measure, which in the present study was assessed through a writing test.

For the fourth claim, standard multiple linear regressions were conducted to provide backing for the assumption that the internal structure of the C-test scores can be associated to the theoretical view that general language proficiency consists of highly interrelated components (Bachman & Palmer, 1996; Hamp-Lyons & Henning, 1991;
Hinkel, 2002; Hulstijn, 2015; Purpura, 2008; Sasaki, 2002). First, a standard multiple regression was computed for both HLLs and FLLs together, using C-test performance as the criterion variable and writing and speaking proficiency as the predictor variables. Findings suggested that both predictor variables together were related to 80% of the C-test score variance. It was also found that both writing and speaking skills individually predicted C-test performance. Similarly, when conducting the same analyses for each learner group separately, results showed that a large percentage of the C-test score variance could be explained by writing and speaking proficiency together (78% for the HLL group and 77% for the FLL group). In addition, both predictor variables significantly predicted C-test performance but in different degrees depending on the learner group. For HLLs, writing proficiency mattered more than speaking proficiency for predicting C-test performance. In contrast, for FLLs, both speaking and writing proficiency were equivalently important for predicting C-test performance.

7.5. Outcomes Relevant to the Extrapolation Inference

Extrapolation was the fifth and the last inference in the Korean C-test interpretive argument chain examined in the present study. This inference generally connects assessment scores to “sample of performances from the target domain” (Kane et al., 1999, p. 10). Nevertheless, as explained in the theoretical ground inference, given that the Korean C-test was not intended to be domain-specific, the evidence to support this inference did not originate from sample performances from the target domain. Rather, the backing for this inference came from test-takers’ performances on criterion tests that helped interpret the results of the Korean C-test in the general context of Korean
language learning. Thus, this inference was based on the warrant that the construct of Korean global proficiency as assessed by the Korean C-test can be related to other criteria of language proficiency that reflect general language proficiency. Table 32 summarizes the warrant, assumptions, and research questions for the extrapolation inference.

Table 32
Extrapolation Inference Warrant, Assumptions, and Research Questions

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumptions</th>
<th>Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The construct of Korean global language proficiency as assessed by the Korean C-test is related to other criteria of language proficiency that are reflective of general language proficiency.</td>
<td>1. Test-takers’ performance on the C-test has a positive relationship with their performance on other measures of language proficiency (i.e., speaking, writing, and oracy skills). 2. C-test performance can be interpreted in terms of other criterion measures generally used in the Korean language learning context. 3. C-test performance has a positive relationship with learners’ self-ratings of their Korean language proficiency.</td>
<td>1. How do test-takers’ performance on the C-test relate to criterion measures of oracy-based language proficiency? 2. How do C-test scores relate to criterion measures generally used in the Korean language learning context? 3. How does the test-takers’ performance on the C-test relate to their self-assessment ratings?</td>
</tr>
</tbody>
</table>

7.5.1. Relationship between Korean C-test performance and oracy-based criterion measures

The first assumption underlying the warrant in the extrapolation inference referred to the relationship between test-takers’ C-test performance and their performance on other tests that also measured oracy-based Korean language proficiency. In this case, the claim was that a strong positive relationship between the C-test performance and other criterion tests would indicate that the C-test is closely related to other measurements of
Korean language proficiency, including standardized measurements of language proficiency, such as the ACTFL proficiency tests. In order to assess the extent to which the C-test and other criterion measures are associated with each other, Pearson correlation coefficients were computed as evidence to back this claim. Thus, apart from the correlation analyses between C-test and ACTFL WPT scores presented in section 7.4.2, additional correlation analyses were estimated between C-test and other criterion measures (i.e., ACTFL OPIc scores, combined mean of both ACTFL tests, Elicited Imitation Test (EIT) scores, and self-assessment estimates).

Before conducting correlation analyses, however, descriptive statistics for both ACTFL tests and their average as well as the Elicited Imitation Test (EIT) were examined first in order to evaluate whether the data met the assumptions to conduct Pearson correlations. Table 33 presents the descriptive statistics for all learners as well as for each learner group (i.e., HLL- and FLL group) separately. Findings indicated that overall all tests were able to assess a wide range of ability levels, as observed through the minimum and maximum scores. Nevertheless, the ACTFL OPIc score range did not extend to all possible scores as there were no test-takers who scored Superior (or 3.00). Similarly, the ACTFL WPT minimum and maximum scores did not extend across the entire available scoring range (i.e., 0.10 – 3.00), suggesting that test-takers who were true beginners in Korean language writing were not represented in the data (see Figure 25). These findings were expected, though, as Korean learners who were recruited for this study had to be enrolled for at least one year in a Korean language program. In addition, results also indicated that the HLL group performed better than the FLL group in all tests, particularly in the OPIc and EIT. These results, however, should be interpreted with
caution because the way learners were clustered into the HLL group and the FLL group included performances on the ACTFL OPIc and ACTFL WPT as clustering factors. In other words, the six Korean language learners who were regrouped into the FLL group through the hierarchical cluster analysis were learners from the lowest curricular level (i.e., those who were enrolled in first-year-level Korean course) who had initially been considered beginner-level HLLs.

Table 33
Descriptive Statistics for ACTFL OPIc, WPT, OPIc & WPT Mean, and EIT

<table>
<thead>
<tr>
<th>Test</th>
<th>Learner Type</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>±SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPIc</td>
<td>All</td>
<td>1.45</td>
<td>0.63</td>
<td>0.10</td>
<td>2.80</td>
<td>±0.13</td>
</tr>
<tr>
<td></td>
<td>HLL</td>
<td>1.89</td>
<td>0.53</td>
<td>0.80</td>
<td>2.80</td>
<td>±0.18</td>
</tr>
<tr>
<td></td>
<td>FLL</td>
<td>1.19</td>
<td>0.54</td>
<td>0.10</td>
<td>2.30</td>
<td>±0.16</td>
</tr>
<tr>
<td>WPTa</td>
<td>All</td>
<td>1.53</td>
<td>0.51</td>
<td>0.80</td>
<td>3.00</td>
<td>±0.11</td>
</tr>
<tr>
<td></td>
<td>HLL</td>
<td>1.73</td>
<td>0.57</td>
<td>0.80</td>
<td>3.00</td>
<td>±0.19</td>
</tr>
<tr>
<td></td>
<td>FLL</td>
<td>1.42</td>
<td>0.44</td>
<td>0.80</td>
<td>2.80</td>
<td>±0.13</td>
</tr>
<tr>
<td>OPIc &amp; WPT</td>
<td>All</td>
<td>1.50</td>
<td>0.53</td>
<td>0.45</td>
<td>2.90</td>
<td>±0.11</td>
</tr>
<tr>
<td>Mean</td>
<td>HLL</td>
<td>1.81</td>
<td>0.51</td>
<td>0.80</td>
<td>2.90</td>
<td>±0.18</td>
</tr>
<tr>
<td></td>
<td>FLL</td>
<td>1.31</td>
<td>0.45</td>
<td>0.45</td>
<td>2.45</td>
<td>±0.12</td>
</tr>
<tr>
<td>EIT</td>
<td>All</td>
<td>52.09</td>
<td>35.60</td>
<td>0.00</td>
<td>117.00</td>
<td>±7.47</td>
</tr>
<tr>
<td></td>
<td>HLL</td>
<td>78.94</td>
<td>27.31</td>
<td>21.00</td>
<td>117.00</td>
<td>±9.53</td>
</tr>
<tr>
<td></td>
<td>FLL</td>
<td>36.34</td>
<td>30.15</td>
<td>0</td>
<td>105.00</td>
<td>±7.92</td>
</tr>
</tbody>
</table>

Notes. aWPT scores are presented here again for convenience.
Four different statistical assumptions were tested for conducting Pearson correlations, including absence of outliers, normality, and linearity. As done in the previous section, to evaluate whether there were outliers, the scores on the Korean C-test, both ACTFL tests, and the Elicited Imitation test were converted into z-scores and all scores ±3.29 were identified as outliers. No outliers were identified in any of the tests based on this measure. Nevertheless, as aforementioned, the boxplot for the ACTFL WPT showed two outliers who had scored 3 (i.e., equivalent to Superior in the ACTFL proficiency guidelines; see Appendix J). The analyses with and without these two observations yielded similar outcomes, and therefore they were not excluded from the analyses. The normality assumption was tested in three different ways. First, the Shapiro-Wilk test of normality was conducted for all tests and the results indicated that the scores of both ACTFL tests and the EIT scores violated this assumption. However, due to its sensitivity to sample sizes (Tabachnick & Fidell, 2013, p. 80), the descriptive statistics, more specifically skewness and kurtosis measures, were also taken into account as another measure of normality. Values for skewness and kurtosis fell between -2 and +2.
indicating a reasonably normal distribution (Bachman, 2002). The ratio between skewness and kurtosis by their respective standard errors were also examined and the statistics indicated that the ACTFL WPT and EIT scores violated this assumption. Finally, with respect to linearity, scatterplots of the bivariate correlations between each variable were examined (Appendix L). All scatterplots displayed linear relationships between variables. In brief, all assumptions except the normality of the ACTFL WPT and EIT scores were met. Therefore, Spearman rank order correlations were considered a better approach to correlation.

Table 34 presents the Spearman rank order correlation coefficients for all tests. Findings indicated that there was a strong positive correlation between the Korean C-test (four-item-based test) and all criterion measures, with Spearman rho correlation coefficients ranging from \( \rho = .81 \) to .90. More specifically, these results suggested that there was a strong association between the Korean C-test and tests that measure writing, speaking, and oracy skills. The strongest association was found between the C-test and ACTFL WPT scores. In addition, similar to the results of the multiple regressions (see section 7.4.2), there was a stronger relationship between C-test and ACTFL WPT than between C-test and ACTFL OPIc, which indicated that the C-test performance was more closely related to literacy-based skills, such as writing proficiency, than to oracy-based skills, such as speaking proficiency. Interestingly, correlation coefficients also revealed that the C-test performance was more closely correlated to the oracy construct measured by the EIT, than the speaking construct assessed by the ACTFL OPIc. This was a predictable outcome, given that the EIT, like the C-test, is a shortcut measure of language proficiency that requires test-takers to understand and decode the overall meaning of the
sentences. In other words, the successful completion of items on the EIT depends on test-takers’ understanding of the meaning of the sentences and their ability to use implicit linguistic knowledge to reconstruct them (e.g., Erlam, 2006; Wu & Ortega, 2013; Zhou, 2012). This type of cognitive process parallels that observed in C-tests, where learners also need to decode the overall meaning of a passage and reconstruct the missing parts of the text. There was, on the other hand, a stronger correlation between the EIT and the ACTFL OPIc than between the C-test and the ACTFL OPIc, which showed that the EIT was more closely associated to oracy skills than to literacy skills (Drackert, 2015; Wu & Ortega, 2013).

Table 34
Spearman Rank Order Correlation Coefficients between C-test and Other Criterion Measures

<table>
<thead>
<tr>
<th>C-test</th>
<th>OPIc</th>
<th>WPT</th>
<th>ACTFL Combined Mean</th>
<th>EIT Combined Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPIc</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WPT</td>
<td>.89</td>
<td>.78</td>
<td></td>
<td>.86</td>
</tr>
<tr>
<td>ACTFL Combined Mean</td>
<td>.90</td>
<td>.95</td>
<td>.92</td>
<td></td>
</tr>
<tr>
<td>EIT</td>
<td>.84</td>
<td>.85</td>
<td>.76</td>
<td></td>
</tr>
</tbody>
</table>

As observed in Table 34, the combined mean of both ACTFL tests was also included as a new variable. This new variable was considered more appropriate in representing a global measurement of language proficiency as it consisted of test-takers’ both written and spoken abilities. Correlation coefficients between C-test scores and ACTFL combined means (i.e., the average scores of ACTFL OPIc and ACTFL WPT) were the highest, with $\rho = .90$ (Figure 26). Like the findings of the standard multiple regressions suggested in the explanation inference, the combined mean of test-takers’ proficiency in speaking and in writing appeared to be strongly correlated to Korean C-test
performance. These findings suggest that both writing and speaking ability together are closely related to C-test performance.

Correlational analyses were also conducted for each learner group separately in order to evaluate whether the association between the Korean C-test performance and other criterion measures were similar. Table 35 presents the relationship between all tests by learner type; the non-shaded and shaded parts of the table respectively display the correlation coefficients for FLLs and HLLs as identified in the two-cluster solution (cf. section 7.3.3.3). Results showed that C-test performance was similarly associated with other criterion measures when comparing both learner groups. The strongest correlation coefficient was found between C-test performance and the mean scores of ACTFL tests for both heritage and foreign language learners. Furthermore, the performance on the C-

Figure 26. Scatterplot of bivariate correlation between C-test and ACTFL OPIc and WPT mean scores
test was more closely related to writing proficiency than to speaking proficiency for both learner groups. In terms of the relationship between C-test and EIT, the correlation coefficients were similar in magnitude for both learner groups.

Table 35

*Spearman Rank Order Correlation Coefficients between C-test and Other Criterion Measures by Learner Type*

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>OPIc</th>
<th>WPT</th>
<th>ACTFL Combined Mean</th>
<th>EIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test</td>
<td>—</td>
<td>.81</td>
<td>.87</td>
<td>.87</td>
<td>.84</td>
</tr>
<tr>
<td>OPIc</td>
<td>.78</td>
<td>—</td>
<td>.78</td>
<td>.94</td>
<td>.80</td>
</tr>
<tr>
<td>WPT</td>
<td>.87</td>
<td>.77</td>
<td>—</td>
<td>.93</td>
<td>.76</td>
</tr>
<tr>
<td>ACTFL Combined Mean</td>
<td>.88</td>
<td>.92</td>
<td>.95</td>
<td>—</td>
<td>.83</td>
</tr>
<tr>
<td>EIT</td>
<td>.86</td>
<td>.78</td>
<td>.76</td>
<td>.81</td>
<td>—</td>
</tr>
</tbody>
</table>

*Notes:* The shaded correlation coefficients belong to the HLL group; the non-shaded correlation coefficients belong to the FLL group.

Based on the results of the independent samples t-tests conducted to compare HLLs’ and FLLs’ performance on the C-test and on the ACTFL WPT (see sections 7.3.3.3 and 7.4.2), it was also important to analyze the relationship between C-test performance and other criterion measures for only the intermediate-level learners (Cluster 1 and Cluster 2 of the four-cluster solution). As mentioned in those sections, results of the independent samples t-tests suggested that a more accurate comparison between HLLs and FLLs could be made when comparing only the intermediate-level learners, because in this way, language proficiency could be controlled for. Therefore, Spearman rank order correlation coefficients were also computed between C-test scores and all other criterion measures for the two intermediate learner groups identified as Cluster 1 (N = 29.
HLLs) and Cluster 2 ($N = 28$ FLLs) in the four-cluster solution. Similar to the previous correlation analyses, the results indicated that the C-test was highly correlated to the ACTFL test means for both learner groups at $\rho = .81$ for the HLLs and $r = .77$ for the FLLs (see Table 36). In addition, writing proficiency appeared to be more closely related to the C-test performance than speaking proficiency was for the HLLs. On the other hand, writing and speaking proficiency were almost equally related to C-test performance for the FLLs. These results echoed results of the multiple regressions (see section 7.4.2), suggesting that while writing and speaking proficiency are both strongly correlated to C-test performance for both learner groups, for the HLLs, writing was more strongly correlated than the speaking.

Table 36

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>OPIc</th>
<th>WPT</th>
<th>ACTFL Combined Mean</th>
<th>EIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test</td>
<td>—</td>
<td>.62</td>
<td>.78</td>
<td>.77</td>
<td>.80</td>
</tr>
<tr>
<td>OPIc</td>
<td>.69</td>
<td>—</td>
<td>.54</td>
<td>.86</td>
<td>.65</td>
</tr>
<tr>
<td>WPT</td>
<td>.81</td>
<td>.67</td>
<td>—</td>
<td>.87</td>
<td>.71</td>
</tr>
<tr>
<td>ACTFL Combined Mean</td>
<td>.81</td>
<td>.89</td>
<td>.92</td>
<td>—</td>
<td>.76</td>
</tr>
<tr>
<td>EIT</td>
<td>.77</td>
<td>.67</td>
<td>.61</td>
<td>.68</td>
<td>—</td>
</tr>
</tbody>
</table>

*Notes.* The shaded correlation coefficients belong to the intermediate-level HLL group (Cluster 1 in the four-cluster solution); the non-shaded correlation coefficients belong to the intermediate-level FLL group (Cluster 2 in the four-cluster solution)
7.5.2. C-test score interpretation and ACTFL proficiency guidelines

In order for Korean C-test scores to be easily interpretable by the Korean heritage and foreign language learning research and education community, it was crucial to describe test-takers’ performance on the C-test in terms of their performance on more commonly-used standardized proficiency tests, such as the ACTFL OPIc and ACTFL WPT. At the same time, this allowed the construct of general language proficiency as measured by the C-test to be extrapolated to other criteria of language proficiency commonly used in the academic and research setting. Accordingly, to answer the second research question in the extrapolation inference, the Korean C-test composite scores were examined in terms of test-takers’ average scores on the ACTFL OPIc and ACTFL WPT, as a combined mean. As mentioned before, the average scores for both tests were considered more appropriate in representing test-takers’ Korean language abilities as they consisted of both spoken and written skills together.

After examining the descriptive statistics for the ACTFL OPIc and ACTFL WPT mean scores, they were recoded to fit the ACTFL proficiency guidelines. More specifically, following the numerical scale of the ACTFL proficiency levels (Vande Berg, Connor-Linton, & Paige, 2009), mean scores were transformed into the nominal scale from Novice Low to Advanced High. There were, however, several scores that fell outside of the suggested 10 numerical scale equivalencies (i.e., 0.1 to 3.00; see Table 37). Thus, a new more detailed scale was developed using the mid-point of two scale-points. For example, for scores that fell between 0.1 (Novice-Low, NL) and 0.3 (Novice-Mid, NM), a mid-point (i.e., 0.2) between these two points of the scale was added and scores equal or lower than 0.2 were considered NL and scores higher than 0.2 but lower or equal
to 0.55 (the next mid-point) were considered NM. Table 37 presents a comparison of the numerical scale suggested by Vande Berg, Connor-Linton, and Paige (2009) and the new scale.

Table 37  
Comparison of ACTFL Proficiency Numerical Scales

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice-Low (NL)</td>
<td>0.10</td>
<td>0.10 – 0.20</td>
</tr>
<tr>
<td>Novice-Mid (NM)</td>
<td>0.30</td>
<td>0.21 – 0.55</td>
</tr>
<tr>
<td>Novice-High (NH)</td>
<td>0.80</td>
<td>0.56 – 0.95</td>
</tr>
<tr>
<td>Intermediate-Low (IL)</td>
<td>1.10</td>
<td>0.96 – 1.20</td>
</tr>
<tr>
<td>Intermediate-Mid (IM)</td>
<td>1.30</td>
<td>1.21 – 1.55</td>
</tr>
<tr>
<td>Intermediate-High (IH)</td>
<td>1.80</td>
<td>1.56 – 1.95</td>
</tr>
<tr>
<td>Advanced-Low (AL)</td>
<td>2.10</td>
<td>1.96 – 2.20</td>
</tr>
<tr>
<td>Advanced-Mid (AM)</td>
<td>2.30</td>
<td>2.21 – 2.55</td>
</tr>
<tr>
<td>Advanced-High (AH)</td>
<td>2.80</td>
<td>2.56 – 2.90</td>
</tr>
<tr>
<td>Superior (S)</td>
<td>3.00</td>
<td>2.91 – 3.00</td>
</tr>
</tbody>
</table>

Next, the descriptive statistics for Korean C-test scores were computed across eight ACTFL proficiency levels from NM to AH. As observed in Table 38 and the error bar graph in Figure 27, performance on the Korean C-test was different across all eight ACTFL proficiency levels. The average difference in mean C-test scores between levels was 10.02. The largest mean score difference was between the NH and IL levels with 14.04 points of difference, followed by the difference in scores between IL and IM with 13.94 points of difference. Overall, C-test mean scores and minimum and maximum scores gradually increased as did the ACTFL proficiency levels. Moreover, the standard deviations were generally low for each level, indicating that there was small variation in C-test scores within each ACTFL level.
### Table 38

**Descriptive Statistics of the Korean C-test Scores across Eight ACTFL Proficiency Levels**

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM</td>
<td>1</td>
<td>22.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>NH</td>
<td>19</td>
<td>29.79</td>
<td>8.95</td>
<td>25.48</td>
<td>34.10</td>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>IL</td>
<td>12</td>
<td>43.83</td>
<td>11.13</td>
<td>36.76</td>
<td>50.90</td>
<td>25</td>
<td>59</td>
</tr>
<tr>
<td>IM</td>
<td>26</td>
<td>57.77</td>
<td>9.45</td>
<td>53.95</td>
<td>61.59</td>
<td>37</td>
<td>78</td>
</tr>
<tr>
<td>IH</td>
<td>16</td>
<td>67.69</td>
<td>10.38</td>
<td>62.16</td>
<td>73.22</td>
<td>47</td>
<td>80</td>
</tr>
<tr>
<td>AL</td>
<td>10</td>
<td>80.50</td>
<td>9.01</td>
<td>74.06</td>
<td>86.94</td>
<td>70</td>
<td>97</td>
</tr>
<tr>
<td>AM</td>
<td>7</td>
<td>91.57</td>
<td>6.58</td>
<td>85.49</td>
<td>97.66</td>
<td>84</td>
<td>99</td>
</tr>
<tr>
<td>AH</td>
<td>2</td>
<td>91.00</td>
<td>5.66</td>
<td>40.18</td>
<td>141.82</td>
<td>87</td>
<td>95</td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
<td>57.28</td>
<td>21.64</td>
<td>52.82</td>
<td>61.74</td>
<td>12</td>
<td>99</td>
</tr>
</tbody>
</table>

There were, however, some discrepancies in the data. First, the average C-test scores for learners who had performed at the AH ACTFL level were lower than those who had performed at the AM ACTFL level. In fact, the lowest mean score difference was between these two levels but in the opposite direction with a score difference of 50.
-0.57. Moreover, the mean scores at the AM level were only nine score points away from reaching the maximum score, which indicated that there was a ceiling effect. Given that the mean scores for test-takers at the AH level were based only on two observations, this finding should be interpreted with considerable caution. As observed in the error bar plot, the AH level showed a large error range. Second, the minimum C-test score for the NH level was 12 which was lower than the minimum scores at the NM level, 22. In other words, there was at least one test-taker at the NH level who had scored lower on the C-test than the test-taker at the NM level. Third, it is also worth noting that the data did not show the full range of all possible ACTFL proficiency levels. More specifically, there were no observations at the Novice-Low (NL) level or at the Superior (S) level.

Given that there was overlap between the lower and upper bounds of the 95% CI as well as between minimum and maximum scores, the ACTFL proficiency levels were regrouped in terms of their three main proficiency levels (i.e., Novice, Intermediate, and Advanced). Once regrouped, descriptive statistics of Korean C-test performance were reexamined and error bar plots were created again (see Table 39 and Figure 28). Similar to the previous results across eight ACTFL proficiency levels, C-test mean scores as well as minimum and maximum scores gradually increased with the increment in main ACTFL proficiency levels. This time, the increase in mean score between proficiency levels averaged 27.96 score points. Furthermore, the small standard deviations also indicated that the C-test performance in each of the proficiency level-groups were similar. The largest variation in C-test performance was found in the Intermediate group.
Table 39

Descriptive Statistics of the Korean C-test Scores across ACTFL Main Proficiency Levels

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>95% CI Lower Bound</th>
<th>95% CI Upper Bound</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>21</td>
<td>29.76</td>
<td>8.81</td>
<td>25.75</td>
<td>33.77</td>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>Intermediate</td>
<td>53</td>
<td>58.00</td>
<td>12.92</td>
<td>54.44</td>
<td>61.56</td>
<td>25</td>
<td>80</td>
</tr>
<tr>
<td>Advanced</td>
<td>19</td>
<td>85.68</td>
<td>9.34</td>
<td>81.15</td>
<td>90.21</td>
<td>70</td>
<td>99</td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
<td>57.28</td>
<td>21.64</td>
<td>52.82</td>
<td>61.74</td>
<td>12</td>
<td>99</td>
</tr>
</tbody>
</table>

In order to explore the extent to which C-test performance was different across these proficiency groups, a one-way Analysis of Variance was conducted. Evaluation of the assumptions underlying the one-way ANOVA, including homogeneity of variance (Levene’s statistics), normality, and absence of extreme outliers, did not reveal any substantial anomalies. Results indicated an overall statistically significant difference of C-test performance with large effect sizes across all three ability groups with $F(2,92) =$
118.95, \( p = .000 \), partial \( \eta^2 = .73 \). Post-hoc comparisons using Bonferroni test indicated that there was a statistically significant difference between all three proficiency levels (Table 40). These findings agreed with the initial analysis of descriptive statistics which showed that while differences in C-test performance within each ability group were small, difference between groups was significantly large.

Table 40

*Pairwise Comparisons of C-test Performance between ACTFL Main Proficiency Levels*

<table>
<thead>
<tr>
<th>ACTFL Main Proficiency Levels</th>
<th>Mean difference</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval Lower Bound</th>
<th>95% Confidence Interval Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>Intermediate</td>
<td>-28.24</td>
<td>2.96</td>
<td>.000</td>
<td>-35.45</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>-55.92</td>
<td>3.63</td>
<td>.000</td>
<td>-64.78</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Novice</td>
<td>28.24</td>
<td>2.96</td>
<td>.000</td>
<td>21.03</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>-27.68</td>
<td>3.07</td>
<td>.000</td>
<td>-35.16</td>
</tr>
<tr>
<td>Advanced</td>
<td>Novice</td>
<td>55.92</td>
<td>3.63</td>
<td>.000</td>
<td>47.07</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>27.68</td>
<td>3.07</td>
<td>.000</td>
<td>20.21</td>
</tr>
</tbody>
</table>

Based on these results, C-test cut-off scores were determined through a cross-tabulation of C-test scores and the three main ACTFL proficiency levels. This analysis showed that test-takers could be accurately classified into the Novice, Intermediate, and Advanced levels according to their C-test performance. Even though there were 11 test-takers who could not be accurately grouped into the ACTFL levels, as can be observed in Table 41, the overall accuracy rate of these cut-off scores was appropriate with 88.17% accuracy rate.
Furthermore, it was important to examine whether the same C-test cut-off scores could classify HLLs and FLLs into the three main ACTFL proficiency level. Therefore, cross tabulations were computed for each learner group separately. As observed in Table 42, the C-test score ranges appear to function similarly for both learner groups. For HLLs, only seven learners could not be accurately classified into these main levels. Six of these learners had performed better on the ACTFL tests than on the C-test, whereas only one learner performed better on the C-test than on the ACTFL tests. From the learners who performed better on the ACTFL tests, five had scored at an Advanced level but had not been able to perform as high on the C-test. Nevertheless, 79.41% of HLLs were accurately grouped. For FLLs, only four learners did not fit the ability groups. These learners had all performed at an Intermediate level on the ACTFL tests but performed like learners on the lowest level on the C-test. Only one of these learners was one of the FLLs who had originally been grouped as a HLL. The accuracy rate of the cut-off scores for this group was higher with 93.10% of test-takers being classified accurately.
Table 42

Cross Tabulation of C-test Mean Scores and ACTFL Main Proficiency Levels

<table>
<thead>
<tr>
<th>Learner Type</th>
<th>C-test Score Range</th>
<th>ACTFL Main Proficiency Levels</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Novice</td>
<td>Intermediate</td>
</tr>
<tr>
<td>HLLs</td>
<td>12 – 43</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>44 – 80</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>81 – 100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>FLLs</td>
<td>12 – 43</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>44 – 80</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>81 – 100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>17</td>
<td>36</td>
</tr>
</tbody>
</table>

7.5.3. Relationship between Korean C-test performance and self-assessment ratings

The third research question explored in the extrapolation inference investigated the relationship between test-takers’ overall self-assessment on their reading, listening, speaking, and writing ability and their performance on the C-test. Self-assessment ratings were considered another important indicator of their Korean language ability because, as Enright et al. (2008) explained, “learners have a much more extensive database to evaluate than do external observers” (p. 177). In order to gauge their overall Korean language ability in different modalities, four 4-point Likert scale items were included in the background questionnaire (see Appendix D).

The descriptive statistics for these items are presented in Table 43. Results indicated that learners generally self-rated their receptive skills (i.e., reading and listening) higher than they did their productive skills (i.e., speaking and writing). Moreover, they rated their listening skills the highest and their writing skills the lowest. However, the variation of self-ratings for the listening skills was high, indicating that learners varied in the way they perceived their listening skills. Given that these data
included self-ratings from learners with a wide range of proficiency levels, the large standard deviations were expected. In contrast, the variation of ratings for the writing skills was the lowest and learners did not use the full-range of the scale from 0 to 3. These results suggested that learners may have underestimated their writing abilities.

Table 43

*Descriptive Statistics: Self-Assessment of Four Korean Language Skills*

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>95% CI</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>1.42</td>
<td>0.70</td>
<td>1.28</td>
<td>1.56</td>
<td>0</td>
</tr>
<tr>
<td>Listening</td>
<td>1.54</td>
<td>0.73</td>
<td>1.39</td>
<td>1.69</td>
<td>0</td>
</tr>
<tr>
<td>Speaking</td>
<td>1.33</td>
<td>0.68</td>
<td>1.19</td>
<td>1.47</td>
<td>0</td>
</tr>
<tr>
<td>Writing</td>
<td>1.15</td>
<td>0.61</td>
<td>1.03</td>
<td>1.28</td>
<td>0</td>
</tr>
<tr>
<td>Overall</td>
<td>1.36</td>
<td>0.53</td>
<td>1.25</td>
<td>1.47</td>
<td>0</td>
</tr>
</tbody>
</table>

The over- or under-estimation of language abilities has been commonly observed in research using self-assessment instruments on HLLs. Findings from previous studies have indicated that certain language skills (e.g., speaking) are inaccurately self-assessed by HLLs and that the type of heritage language and proficiency level could moderate these results (Donovan et al., 2012; Kang & Kim, 2012; Martin et al., 2013, Swender et al., 2014). Hence, it was important to examine the descriptive statistics of self-ratings for the HLL group separately. At the same time, a comparison between both learner groups and their self-perception of Korean language ability was then possible. Findings suggested that comparatively, HLLs self-assessed their oracy skills (i.e., Listening and Speaking) as their strongest skills, whereas FLLs rated their literacy skills (i.e., Reading and Writing) as their strongest skills (see Table 44). This pattern has been commonly reported in the HLL literature (Beaudrie & Ducar, 2005; Kang & Kim, 2012; Kondo, 1997; Montrul, 2008). Kagan and Dillon (2012), for example, found that HLLs typically
self-assessed their listening skills as their best language skill followed by speaking, reading, and finally writing. This is the exact pattern that was observed in the present study for HLLs.

Table 44

*Descriptive Statistics: Self-Assessment of Four Korean Language Skills for Each Learner Type*

<table>
<thead>
<tr>
<th>Learner Type</th>
<th>Reading</th>
<th>Listening</th>
<th>Speaking</th>
<th>Writing</th>
<th>Overall ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLL Mean</td>
<td>1.47</td>
<td>1.94</td>
<td>1.62</td>
<td>1.00</td>
<td>1.51</td>
</tr>
<tr>
<td>HLL SD</td>
<td>.71</td>
<td>.60</td>
<td>.74</td>
<td>.65</td>
<td>.55</td>
</tr>
<tr>
<td>FLL Mean</td>
<td>1.38</td>
<td>1.29</td>
<td>1.17</td>
<td>1.24</td>
<td>1.27</td>
</tr>
<tr>
<td>FLL SD</td>
<td>.70</td>
<td>.70</td>
<td>.60</td>
<td>.57</td>
<td>.51</td>
</tr>
</tbody>
</table>

Once descriptive statistics were explored for the four items, data were analyzed in terms of the extent to which learners’ self-ratings on the four modalities related to their performance on the C-test. Thus, Pearson correlation coefficients were computed between the four items as well as the overall mean score for all four modalities and the C-test performance. As done in previous sections evaluation of the assumptions underlying the correlational analyses, including normality, linearity, and absence of extreme outliers were conducted and no substantial anomalies were identified.

As observed in Table 45, the strongest association was found between C-test performance and the overall language ability with \( r = .62 \), which indicated that the general construct of language proficiency measured by the Korean C-test was closely related to the global construct of language proficiency indicated by the self-assessment ratings. In general, correlation coefficients of this magnitude could be characterized as moderate to strong correlations (Plonsky & Oswald, 2014). Furthermore, Ross’s (1998) meta-analysis of self-assessment in second language testing found that the average correlation of overall proficiency in 60 self-assessments was .63 with a median of .49 and
an upper hinge of .65. When examining each language skill separately, reading was the most closely correlated variable followed by listening and speaking. Writing skills were found to be the least associated with the C-test performance with a low correlation coefficient of $r = .34$. This result somewhat contradicts the findings of the correlation between C-test and ACTFL WPT performance. As explained before when exploring the second claim in the explanation inference (see section 7.4.2), writing proficiency as measured by the ACTFL WPT was found to be strongly correlated to Korean C-test performance with $r = .85$. The discrepancy between these findings could be due to the inaccurate perception of writing ability by learners, or perhaps to a mismatch between what is measured in the WPT and what learners consider to be their goals for writing in Korean. Like the descriptive statistics suggested, learners, particularly HLLs, appeared to have underestimated their proficiency in writing.

Table 45

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Listening</th>
<th>Speaking</th>
<th>Writing</th>
<th>Overall ability</th>
<th>C-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Listening</td>
<td>.56</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Speaking</td>
<td>.50</td>
<td>.55</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Writing</td>
<td>.65</td>
<td>.31</td>
<td>.35</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Overall ability</td>
<td>.86</td>
<td>.79</td>
<td>.77</td>
<td>.71</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>C-test</td>
<td>.53</td>
<td>.51</td>
<td>.51</td>
<td>.34</td>
<td>.62</td>
<td>—</td>
</tr>
</tbody>
</table>

Correlation coefficients between the four language skills and C-test performance were also calculated separately for each learner group in order to examine the extent to which these associations were similar or different for distinct learner groups. Table 46 presents the Pearson correlation coefficients by learner groups. Just as when exploring
both learner groups together, C-test performance and overall language ability was the strongest correlation found for both groups with $r = .70$ for the HLL group and $r = .55$ for the FLL group. On the whole, correlation coefficients were higher in magnitude between all variables for HLLs than for FLLs. This was also true for correlations between the ACTFL tests and the four skills. This might, arguably, indicate that HLLs are able to self-assess their language skills with slightly higher accuracy than FLLs. These results could also indicate that HLLs were better at assessing their language abilities holistically than FLLs. In other words, as the self-assessment questionnaire contained only four items targeting very broad language skills (i.e., reading, listening, speaking, and writing), FLLs might have focused on specific contexts for these skills that might have contributed to the inaccuracy of their self-ratings. Nevertheless, these results should be interpreted with caution as further research must be conducted to determine the accuracy rate of Korean learners in rating their own language abilities.

Table 46

*Pearson Correlation Coefficients between C-test Performance and Four Language Skills by Learner Group*

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Listening</th>
<th>Speaking</th>
<th>Writing</th>
<th>Overall ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLL</td>
<td>C-test</td>
<td>.64</td>
<td>.55</td>
<td>.56</td>
<td>.55</td>
</tr>
<tr>
<td>FLL</td>
<td></td>
<td>.49</td>
<td>.37</td>
<td>.41</td>
<td>.48</td>
</tr>
</tbody>
</table>

7.5.4. **Summary of the outcomes relevant to the extrapolation inference**

The extrapolation inference was based on the warrant that the general language proficiency construct as assessed by the Korean C-test accounts for the learners’ performance in other language proficiency tests that are also reflective of global language proficiency. Three assumptions underlying this warrant were (1) that the performance on
the C-test is related to other criteria of Korean language proficiency such as the performance on the ACTFL OPIc, ACTFL WPT, and EIT; (2) that the C-test scores can be described in terms of ACTFL proficiency guidelines; and (3) that the performance on the C-test can also be related to learners’ self-assessment ratings on their Korean language proficiency. Backing for these assumptions was obtained from examining descriptive statistics as well as from correlational analyses between C-test scores and other criterion-measures.

With respect to the first claim, findings indicated that Korean C-test performance was strongly and positively correlated to all criterion measures, namely to learners’ performance on the ACTFL OPIc, ACTFL WPT, and EIT. These results showed that the construct of general language proficiency as measured by the C-test was strongly associated to other measures of Korean language proficiency in writing, speaking, and oracy skills. The strongest correlation was found between C-test and the ACTFL WPT performance, which suggested that the constructs assessed by the C-test were closely related to the literacy measure more so than with the oracy measures. These findings were in line with the theoretical perspectives of what C-tests are supposed to measure. In other words, due to their written modality, C-tests require learners to have literacy-based skills in order to complete them and thus it is expected that the performance on the C-test be more related to literacy than to oracy skills. Nevertheless, it should not be overlooked that the correlation with speaking measures was also really strong. In terms of the ACTFL proficiency tests, the correlation between C-test and the mean score of ACTFL OPIc and ACTFL WPT was even stronger with $\rho = .90$. Given that the mean scores of the ACTFL tests were reflective of both speaking and writing proficiency, these results
appeared to suggest that the construct measured by the C-test was strongly correlated to a measure of proficiency that consisted of both oracy and literacy skills.

The associations between Korean C-test performance and the three criterion measures were also examined for each learner group, separately. The trend in the relationship between the C-test performance and the criterion measures were similar for both learner groups. As when looking at both learner populations together, the ACTFL test mean scores were the most strongly correlated criterion measure to C-test performance. Likewise, writing proficiency appeared to be more strongly associated to C-test than was speaking proficiency for both learner groups.

In terms of the second claim, Korean C-test scores were described according to ACTFL proficiency guidelines. Because these guidelines are widely used in the Korean foreign language research and education, the results of the C-test could, then, be easily interpreted by the KFL research community. Learners’ mean scores on the ACTFL OPIc and ACTFL WPT combined were recoded to show 10 different sublevels of ACTFL proficiency from Novice Low (NL) through Superior (S) and they were subsequently grouped into these levels. Then, the descriptive statistics of C-test performances were examined across these proficiency groups. Eight different sublevels of ACTFL proficiency were found in the data from NL to Advanced High (AH). The performance on the C-test differed across all eight-level groups with a gradually increasing trend. The average mean score difference in C-test scores between adjacent proficiency levels across this range was 10.02 score points.

In order to examine classification accuracy between proficiency groups, data were regrouped according to the ACTFL proficiency main levels, that is into Novice-,
Intermediate-, and Advanced-levels. C-test descriptive statistics on the new groupings were reexamined. The new mean score difference between adjacent groups was 27.96 score points. In addition, a one-way Analysis of Variance (ANOVA) was conducted to assess the extent to which the C-test performance differed across the main language proficiency groups. Results indicated that there was indeed a statistically significant difference between groups with a medium effect size. Post-hoc comparisons also showed that the difference in C-test performance was statistically different across all proficiency groups.

Once it was determined that the main proficiency groups were different, C-test cut-off scores were defined by cross-tabulating the C-test mean scores and ACTFL proficiency main levels. The C-test score ranges considered Novice, Intermediate, and Advanced, were respectively 12-43, 44-80, and 80+. The accuracy rate to which these ranges grouped learners’ C-test performance was 88.17%. Results also indicated that the same cut-off scores appropriately distributed both HLLs and FLLs with a high degree of accuracy as observed through separate cross-tabulations.

With respect to the third claim in the extrapolation inference, learners’ self-assessment ratings on Korean language proficiency were analyzed as an additional indicator of proficiency. As with the correlational analyses between C-test performance and other criterion-tests, the claim was that a positive strong correlation between C-test and self-ratings would provide further evidence that the construct of language proficiency measured by the C-test can be closely associated to the construct of general language proficiency. Learners were asked to respond to four 4-point Likert scale items assessing their overall proficiency across four different language skills, namely reading, listening,
speaking, and writing. Descriptive statistics from these items revealed that learners perceived their listening ability as their strongest language skill, followed by reading, speaking, and lastly writing. When exploring each learner group separately, it was found that self-ratings from HLLs followed the typical trend described in previous HLL studies (e.g., Beaudrie & Ducar, 2005; Kagan & Dillon, 2012; Kang & Kim, 2012; Kondo, 1997; Montrul, 2008). Namely, HLLs regarded their listening skills as the strongest followed by speaking. Their writing skills were largely underestimated. This trend differed with self-ratings from FLLs. While they agreed with HLLs in that they thought their listening skills were the strongest, they regarded their speaking skills as their lowest ability.

Pearson correlation coefficients between self-ratings and C-test performance were overall moderate to strong ranging from .34 to .62. The mean score of all four skills was the most strongly associated with C-test performance, suggesting that the construct assessed by the C-test correlated strongly with a language construct consisting of a combination of four language skills, that is, reading, listening, speaking, and writing. The same results were found when exploring correlation coefficients for each learner group separately.

7.6. Test Use Inference: Warrant, Assumptions, and Research Questions to Be Explored

The last inference in the Korean C-test interpretive argument chain was the utilization inference. However, as has already been indicated before, the Korean C-test has yet to be operationalized and thus the evaluation of this inference fell outside of the scope of the current dissertation study. Despite the unfeasibility of exploring this
inference, it is crucial to lay out the warrant and assumptions as well as the possible research questions that could be investigated once the test becomes fully operational.

Table 47 presents a summary of the warrant and assumptions underlying this inference and the research questions that emerged from these claims.

Table 47

test Use Inference Warrant, Assumptions, and Research Questions

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumptions</th>
<th>Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Korean C-test performance results can be used as a measure of Korean learners’ general language proficiency for applied linguistics research that seeks to examine both Korean HLLs and FLLs.</td>
<td>1. The C-test scores/results are easily interpretable by applied linguistics researchers. 2. The C-test is useful for measuring language proficiency for research where language proficiency is a predictor or moderating variable. 3. The C-test is useful for research focusing on both HLLs and FLLs or on one of the learner groups. 4. The C-test helps to better understand research findings across multiple studies on Korean language learners.</td>
<td>1. To what extent can the C-test scores be easily interpreted by applied linguistics researchers? 2. What are the perceptions of the Korean C-test users in terms of its applicability in their research domains? 3. To what extent have C-test results been useful in assessing proficiency for the purposes of conducting research on Korean learners? 4. To what extent can findings from research on Korean language learners be generalizable when using the Korean C-test to measure general language proficiency?</td>
</tr>
</tbody>
</table>

The utilization inference, as its name implies, is based on the warrant that the Korean C-test results are useful as indicators of Korean general language proficiency for the purposes of a variety of applied linguistics research that include Korean language learners. In order for the test to be useful, it should meet at least four requirements: (1) that the C-test scores are easily interpretable by researchers; (2) that the C-test is useful in measuring language proficiency for research were language proficiency is a predictor or moderating variable; (3) that the C-test is useful for research that focuses on both heritage...
and foreign language learners; and (4) that the C-test contributes to the generalization of results across Korean applied linguistics research.

In future, backing for the aforementioned claims can be collected from the perceptions of the Korean C-test users, which would include not only researchers but also C-test-takers. These data would inform the C-test developers and evaluators to have a better understanding of the way the Korean C-test is being used in different research domains, as well as the advantages and disadvantages of this instrument. Furthermore, if the Korean C-test is widely used among Korean language learning researchers, a more comprehensive meta-analytic study can be conducted to explore the effectiveness of this tool for generalizing results across studies that are investigating the same phenomenon.
Chapter 8

Discussion and Conclusion

Thus far, the interpretive argument for the Korean C-test, which included six inferences (i.e., theoretical ground, evaluation, generalization, explanation, extrapolation, and utilization) was presented in terms of assumptions, warrants, and the empirical backing to support these warrants. Once an interpretive argument is sketched out and the evidence to support each claim is presented, it is possible to examine the validity argument, which can weigh the evidence provided in the interpretive argument and assess the coherence of the argument. This chapter describes the current validity argument for the Korean C-test by briefly summarizing the results relevant to each inference (focused more specifically on the evaluation, generalization, explanation, and extrapolation inferences) and explaining them in terms of previous research on Korean language learners and C-tests in general. In addition, it evaluates the strength of the links between each inference and the logic underlying the argument. Next, some of the limitations of the present study are acknowledged and suggestions for future research are provided. The chapter ends by discussing the implications of the current study for the use of the Korean C-test in applied linguistics research.

8.1. The Validity Argument for the Korean C-test

As mentioned in Chapter 5, the primary intended use of the Korean C-test developed by Son et al. (2018) was to provide a quick measurement of Korean language learners’ general language proficiency for research purposes. Therefore, following Kane’s (2002, 2006, 2013) validation model, the intended conclusion (or main claim) of
the validity argument examined in the present study was that the Korean C-test scores are valid for making interpretations of Korean language learners’ general language ability, specifically literacy-based skills, for research purposes. These scores would be useful for helping researchers in the field of applied linguistics to distribute learners with a wide range of language abilities into different proficiency groups. At the same time, the use of the Korean C-test would contribute to the comparability and generalizability of results across studies. To support this conclusion, backing for the warrants in the chain of inferences leading up to this conclusion were gathered and presented in Chapter 6. For example, the interpretive argument started with the theoretical grounds inference, which described the theoretical characterizations of C-test constructs, derived from extensive research investigating C-test design and implementation. The intermediate conclusion for this inference was that the extant literature on C-tests suggests that observations of C-test performance can be associated with variations in general language proficiency. Subsequently, defining the C-test construct served as the grounds for the next inference, evaluation. The interpretive argument followed the same structure of grounds and intermediate conclusions for each of the inferences, and it was crucial to provide empirical support (or backing) for each inference in order to advance from one inference to the next.

Figure 29 synthesizes the steps in the Korean C-test validity argument and the backing collected to support each step. This staircase structure was adapted from Chapelle et al. (2008, p. 349) to represent the different claims and backing collected to support each claim. As can be observed in the figure, each inference guided different investigations that led to the collection of evidence to support the inferences. Like a
staircase, each of the inferences served as the base for the next step or inference leading up to the intended conclusion related to the use of the Korean C-test. Given that the Korean C-test is yet to be operational, the evidence to support the last step, namely utilization, has not yet been collected and analyzed. Ultimately, the Korean C-test validity argument currently consists of the holistic analysis of all of the inferences together with their warrants and claims and how backing helped support (or reject) these claims. Accordingly, the validation argument approach involved evaluating the extent to which the backing for each inference was strong enough to support each claim.

Figure 29. Steps representing the Korean C-test validity argument. This representation is adapted from Chapelle et al. (2008; p. 349). The utilization step is in grey as it has yet to be explored.
One important component of the argument-based approach (ABA) to validity that is not represented in the staircase is the rebuttals. In Chapter 4, it was explained that Kane’s (2006, 2013) validity model follows Toulmin’s (1958, 2003) general framework for argumentation. More specifically, each inference in the validity model is based on this argumentation framework, which includes a claim, data supporting the claim, warrants, backing, rebuttals, and rebuttal data (Kane, 2006, 2013; see also section 4.4.1 for a detailed description of this components). In this framework, rebuttals and the rebuttal data play a crucial role, as they are the counterclaims that could weaken the validity argument (Bachman, 2005). The Korean C-test validity argument included rebuttals for the evaluation, generalization, and explanation inference and they were used to weigh the evidence that was collected to support each claim and the evidence gathered to support the counterclaims.

8.1.1. Theoretical grounds

One major difference between the way the ABA was implemented for the Korean C-test and the way it is generally implemented for other tests was the first step, defining the domain. As mentioned in Chapter 7, the Korean C-test was designed to be non-domain specific. Although the larger target domain of the C-test was applied linguistics research, it differed from validation research on tests for specific purposes, where a specific domain, such as academic or professional language use, can be identified. The Korean C-test was purposely designed to be domain-free in order for this instrument to be useful in a wide range of research domains within applied linguistics. It is important, however, to underscore that a domain-free test is not equivalent to a “multi-purpose” test instrument. In other words, the Korean C-test is not intended to be used for any possible
uses of language proficiency tests. For example, the C-test is not intended to be used as a placement test in a language program, because it has not been designed with a specific language program curriculum in mind. The use and the interpretation of results of the C-test should be guided by its main intended use of providing a quick measurement of Korean language learners’ general language proficiency for research purposes.

Given that a domain analysis was not possible for the Korean C-test, the first inference in the validity argument, the theoretical grounds inference was, as its name suggests, theoretically-based rather than empirically-based. It was built upon the warrant that the observations of the performance on the C-test reflect test-takers’ Korean general language proficiency as has been described by a large body of C-test literature. The underlying assumption for this warrant was that the C-test format has been subject to sufficient research that has been able to identify and define C-test constructs and has provided strong evidence for believing this format is a good measure of global language proficiency.

The extant literature on C-tests can be summarized into two major findings: (1) C-tests have been found to measure multiple language skills, including productive and receptive skills, and (2) learners’ language proficiency is a major moderator of the constructs being measured by C-tests. For the first major finding, there have been two broad approaches to investigating C-test constructs. One of them has been to measure the relationship between C-tests and other language skills. This type of research has shown that C-tests can be closely associated with discrete language skills, such as vocabulary, grammar, listening, and reading skills (Chappelle, 1994; Cohen et al., 1984; Harsch & Hartig, 2015; Karimi, 2011). The other approach has been to examine the extent to which
C-tests can measure general language proficiency. While defining general language proficiency remains a contentious topic, these previous studies have provided ample evidence which indicate that the constructs being measured by C-tests can be associated to multiple language skills. Numerous studies have found through confirmatory factor analyses that scores on various language tests (e.g., reading, listening, speaking, and writing tests) and C-test scores load high on one single factor, which has been described as general language proficiency (Eckes & Grotjahn, 2006; Grotjahn, 1987; Klein-Braley, 1994; Klein-Braley & Raatz, 1984; Raatz, 1984). Additionally, other studies, particularly those conducting verbal protocols, have also suggested that C-tests are able to provide a holistic measurement of language proficiency. These studies have found that C-test-takers use both micro- and macro-level cues in order to complete the missing part of the words in the C-test items (Babaii & Ansary, 2001; Babaii & Jalali Moghaddam, 2006; Feldman & Stemmer, 1987; Stemmer, 1991). Thus, test-takers need to use a more global knowledge of the language to process the C-test items as a whole and specifically address each of the blanks.

One word of caution worth mentioning here is that although previous research has found a close relationship between C-test constructs and discrete language skills–even those including oracy-based skills (i.e., speaking, listening)–the inherent design of the C-tests requires test-takers to have basic literacy-based skills to be able to take the test. Therefore, the interpretation of C-test constructs should be limited to learners’ literacy skills. That is, C-tests require learners to have basic literacy skills to complete the test, given that it is in the written modality, and thus learners who are only proficient in the spoken modality cannot be accurately assessed by this type of test.
The second major finding in C-test research suggests that the C-test constructs are moderated by test-takers’ level of language proficiency. Sigott (2006) described the constructs assessed by C-tests as fluid, explaining that learners with different language abilities approach C-test items in different ways. That is, the strategies used by low- and high-proficiency learners to complete C-test items vary. Studies on this specific issue have yielded mixed results, however. On the one hand, it has been suggested that low-proficiency learners are more successful in completing C-test items when they are provided with context whereas high-proficiency learners are successful in completing C-tests even when items are decontextualized (Sigott, 2004, 2006). On the other hand, studies implementing verbal protocols have indicated that low-proficiency learners use local cues to complete C-test items, particularly when they are unable to understand the overall meaning of a C-test passage, whereas high-proficiency learners use not only local cues but also linguistic, textual resources (Babaii & Fatahi-Majd, 2014). Even though more research is needed to examine the role of proficiency in the completion of C-tests, the sum of these studies has underscored the importance of taking into account proficiency levels when interpreting C-test results.

As described through the brief summary of the C-test literature, there is extensive research (see also Grotjahn, 2018 for a complete C-test bibliography) that has characterized C-test constructs as being holistic in terms of how they can assess general language proficiency, and as being fluid, given that they are moderated by learners’ language abilities. This research had helped inform the initial process of developing and validating the use of the Korean C-test by Son et al. (2018), thus serving as the theoretical grounds in the validity argument presented in this dissertation study.
8.1.2. Evaluation

The main objective in evaluation was to examine the accuracy of the scores on the C-test when reflecting learners’ Korean general language proficiency. Once the theoretical background for C-test construct was described (i.e., in the previous inference), the next step was to investigate whether observations of performance on the Korean C-test could be evaluated to provide scores that would reflect such constructs. Thus, the validity warrant for the evaluation inference rested on the support of three assumptions: (1) the C-test item selection procedure is appropriate in choosing items that could assess learners of varying degrees of language abilities; (2) the statistical characteristics of the C-test items are appropriate for norm-referenced decisions; and (3) items on the C-test can provide an accurate measurement for distinguishing the performance of test-takers with a wide range of proficiency levels. Backing for these assumptions came from examining the C-test text selection process through descriptive statistics and IRT measures, analyzing test-takers performance on the selected items through descriptive and inferential statistics, and investigating the item and person fit statistics of the IRT model for the Korean C-test. Figure 30 illustrates the validity of the warrant for the evaluation inference. Following the ABA framework to validation, a possible rebuttal to this warrant was also included. However, as will be explained later in this section, backing to support this rebuttal was not found, further strengthening the argument for this inference.
As represented in Figure 30, the psychometric characteristics of the Korean C-test scores in Son et al. (2018) and in the current dissertation study provided backing for all three assumptions in the evaluation inference. First, the process of selecting texts to be converted into C-test items as well as the psychometric qualities of the seven texts selected from the pool of 10 texts piloted in Son et al. (2018) supported the assumption that the final seven texts selected for the current larger-scale validation study were appropriate. The evidence supporting this claim came from fit statistics and the analysis.
of item difficulty, which indicated that the items performed well at targeting learners with a wide range of language abilities. Second, results of the descriptive statistics of learners’ performance on the seven C-test items indicated that the scores were normally distributed, suggesting that the C-test was able to successfully do what it was intended to do. Similarly, when examining the C-test performance of HLLs and FLLs separately, C-test scores were also normally distributed. In other words, the Korean C-test was able to distribute both HLLs and FLLs into a wide range of language abilities. Finally, the IRT measures for the C-test items helped identify the best fitting four items (i.e., texts) and the possible learner-outliers that could introduce noise to the item analysis. These four items showed good fit to the Rasch model and were able to target different levels of difficulty as well as to distribute learners along more than five logits worth of levels with high reliability.

These findings furnished evidence to suggest that the C-test items were able to contribute to accurate scores that reflected the general Korean language ability of learners, both HLLs and FLLs. One rebuttal to this claim, however, was that the psychometric characteristics of the Korean C-test items were not accurate due to the sample represented in the data in the sense that the learners recruited for this study might not have reflected a wide range of ability levels. Nevertheless, as displayed in Figure 30, no evidence was found that could support this rebuttal. The descriptive statistics and IRT measures of C-test scores for all learners combined and for HLLs and FLLs separately revealed a wide range of language abilities represented in the data. The wide range of scores, the large standard deviations, and the Rasch model person separation indices all provided evidence against the rebuttal.
One point of concern, however, was relevant to the results of the independent samples t-test that compared the HLLs’ and FLLs’ performances. As described in section 7.2.2, inferential statistics were computed to accompany the descriptive statistics and explore the extent to which HLLs’ and FLLs’ performance on the C-test were similar. Results showed that HLLs performed statistically significantly better than FLLs on the C-test. A significant difference in performance between both learner groups could indicate that the Korean C-test was assessing both learner groups in different ways, on the one hand, and on the other that the learner sample was not representative of a wide range of abilities. The difference in performance between both learner groups might suggest that the C-test was easier for HLLs than it was for FLLs, and that the C-test was not achieving its primary intended use, which was to distribute learners into different proficiency levels. Nevertheless, the results of the effect size calculations indicated that the statistical difference in performance was small in magnitude. Furthermore, when examining each learner group separately, the C-test scores were normally distributed as indicated by Kolmogorov-Smirnov’s test of normality. Therefore, although HLLs performed better than FLLs on the C-test, the normally distributed performances of each group suggested that the C-test was successful in doing what it was intended to do for both learner groups.

These findings, moreover, encouraged a closer look into the different language abilities of HLLs and FLLs. It was found that although there was a wide range of abilities represented within each learner group (i.e., the HLL group and the FLL group), when comparing both learner groups there were more beginner-level learners in the FLL group than in the HLL group and more advanced-level learners in the HLL group than in the FLL group, as measured by the Korean C-test. It is important to point out that although
the recruitment of participants was carefully conducted in order to have a balanced number of HLLs and FLLs of the same language ability levels (see section 6.1), using curriculum levels as an initial criterion of recruitment might not have been accurate. As numerous researchers (e.g., Norris & Ortega, 2012; Thomas, 1994, 2006; Tremblay, 2011) have pointed out, curriculum level is not the most reliable criterion when trying to group learners into different proficiency groups. Future studies should take this drawback into consideration when recruiting participants and use other indicators of proficiency that are more accurate in differentiating learners’ language abilities.

Another important finding that needs further discussion was the exceptionally good performance on the C-test of six learners. In section 7.2.3, the IRT measures for learners’ performance on the four items on the C-test identified six learners who had achieved almost perfect accuracy levels on the C-test and whose performances had, to a certain degree, set them apart from the rest of the learners (see Wright Map in Figure 13). On the one hand, from the language assessment perspective, particularly from a C-test expert’s standpoint, this was an expected outcome. As explained in Chapter 7, the C-test literature has commonly reported ceiling effects in C-test performance when the sample includes advanced-level language learners. Grotjahn, Klein-Braley, and Raatz (2002) asserted that as learners become proficient in a language they approach a level of proficiency similar to that of educated native speakers, who would be expected to achieve almost perfect scores on the C-test. Thus, if C-tests distribute learners along a continuum from beginner to advanced levels of proficiency, as the C-test in the current dissertation study is intended to do, ceiling effects are to be expected.
On the other hand, from an SLA and HLA perspective, these six very high performers bring up interesting questions for discussion regarding the effect of formal education on sequential bilinguals as well as the type of learners that the C-test can accurately assess. As described in Chapter 7, a closer inspection of these exceptional learners’ responses to the background questionnaire revealed that, out of the six learners, five were HLLs and one was a FLL. The HLLs reported having unique Korean language learning backgrounds. Three of the five HLLs indicated they had received formal education in Korea for almost all of their elementary school education. One of these learners had even completed two out of her three years of middle school education in Korea. These learners can be described as sequential bilinguals or late child L2 learners (De Houwer, 2009, 2011; Meisel, 2011; Montrul, 2016) given that they grew up in a Korean monolingual environment and received their formal education in Korea until they encountered English at a later age (all three after the age of 10) when they moved to the United States.

Furthermore, two of the five HLLs in the exceptional group of six very advanced learners indicated having received minimal formal education in Korea, with one of them reporting going only to kindergarten in Korea and the other finishing only the first year of elementary school. Thus, these two learners showed a profile similar to other HLLs in the sample who had been born in Korea and had moved to an English-speaking country at an early age. Nevertheless, unlike other Korea-born HLLs, these learners had shown advanced proficiency levels in Korean as assessed by the C-test. The learner who indicated having received only one year of kindergarten education in Korea was the learner with the highest score on the C-test. A brief conversation with her after the
administration of the tests revealed that although she had grown up in Great Britain, her parents had tried to expose her as much as they could to the Korean language and culture. Thus, she had received instruction from Korean private tutors while in Great Britain and she had visited Korea multiple times during her upbringing. This had likely increased the amount of Korean input she received and contexts where she could use the language.

The only FLL within these uniquely advanced performers had a learner profile similar to many other advanced-level FLLs who had experienced studying in Korea. Yet, like the five HLLs in the very advanced group, she had achieved an exceptionally high score on the C-test. Her responses to the background questionnaire indicated that she had studied Korean independently for two years during high school and had enrolled in Korean courses in her university for a year and a half. She also reported to have studied abroad in Korea, enrolling in a Korean language institute for one year and three months and in a university for about two and a half months. In addition to the extended stay in Korea, another factor, and one that set her apart from most of the other high but not exceptionally performing FLLs was her frequency of Korean language use in different contexts. For example, she reported often using Korean for texting, emailing, taking notes, talking informally and formally to people, reading novels, browsing the web, and listening to music, among other contexts. Similar to the case of the HLL who only went to kindergarten in Korea, this FLL worked to maximize input in Korean as she could and used the language in numerous contexts.

In brief, the different learner profiles helped explain the outcomes of these six learners’ performances on the C-test. Age of arrival appeared to be a major factor contributing to the exceptionally good performance on the C-test. Most participants who
arrived in the US (or another English-speaking country) after the age of seven showed advanced-levels of Korean language abilities. This is in line with several previous studies that have reported that this type of sequential bilingual tends to have a higher language proficiency than learners who are simultaneous early bilinguals, as they have been exposed to the HL for a longer period of time and have enjoyed the benefit of formal education in the medium of Korean, such that their retention of the language is less vulnerable to attrition (Allen, Crago, & Presco, 2006; Montrul, 2002, 2008, 2016; Zaretsky & Bar-Shalom, 2010). There were, however, exceptions, such as the case of the learner who immigrated to Great Britain at the age of five and who maintained an advanced-level of Korean proficiency throughout her life, and the one FLL who had availed herself of extensive and sustained opportunities for the use (including reading and writing) of Korean. Conversely, there were five cases of HLLs (50%, $N = 5$) who grew up in Korea until or past the age of five but had not reached the same exceptionally high-levels of proficiency.

As it can be surmised, an important factor that also helped explain these learners’ scores on the C-test was the years of formal education received in Korea. Numerous studies have described formal education in the HL country as one major variable (e.g., Caldas & Caron-Caldas, 2000; Gatti & O’Neil, 2017; Kondo-Brown, 2010; Montrul, 2016). This factor inevitably is also closely related to the age factor, as learners who immigrate to the United States at a later age receive some years of formal education. In addition, these learner profiles also revealed that apart from age and formal education, the amount and quality of input and language use played a distinct role in the mastery of the Korean language (Jia & Paradis, 2014; Tomasello, 2003). Given that the current study...
only included 10 cases of Korea-born HLLs (or 29% of HLLs) who had arrived in the United States at a later age, these results should be interpreted with caution. Future studies should include a large number of HLLs with more systematically varying ages of onset and amounts of formal schooling through the medium of the heritage language in order to shed light on the effects of age, education, input, and language use on the performance of HLLs on C-tests.

Moreover, one important question relevant to the validation of the use of the Korean C-test was whether the test was accurate in assessing these types of advanced learners. As could be observed in the Wright Map in Figure 13, there were no items on the C-test that could target the six very high-performing learners. Nevertheless, it seems important to reiterate the primary intended uses of the Korean C-test. Throughout this dissertation, it has been emphasized that the Korean C-test was develop for the purpose of providing a quick measurement of Korean general language ability for learners with varying degrees of language proficiency. At the same time, this measurement would allow the distribution of learners into different proficiency levels. In that sense, performance on the current Korean C-test does not seem to be an appropriate indicator of proficiency for differentiating the language abilities within very advanced-level learners of Korean that perform similar to monolingual native speakers. More on this issue will be discussed in the next section.

8.1.3. Generalization

Generalization was mainly concerned about the reliability of the C-test items in measuring Korean language learners’ (both HLLs and FLLs) general language proficiency. The strength of the validity argument for the generalization inference
depended on finding support for four assumptions: (1) the C-test items can provide reliable and consistent estimates of test-takers’ performance; (2) the current configuration of items on the C-test is the most appropriate; (3) the number of items on the C-test is sufficient to provide reliable estimates of test-takers’ performance; and (4) the C-test items can consistently reflect both HLLs’ and FLLs’ general language proficiency. Figure 31 illustrates the generalization inference, its warrant, assumptions, backing for the assumptions, as well as the possible rebuttal and backing for the rebuttal.

**Figure 31.** Representation of generalization with four assumptions, backing, and rebuttal
As observed in Figure 31, backing for the four assumptions in this inference came from item reliability analyses, Rasch model item fit statistics, and DIF analyses comparing the HLLs’ and FLLs’ performances on each C-test item (or text). For the first assumption, the overall high Cronbach’s Alpha coefficients provided evidence in support of the high reliability and consistency of the C-test items in assessing Korean language learners, both HLLs and FLLs. Although there was a slight decrease in reliability measures when reducing the number of C-test items from seven to four, the analysis still showed very high reliability estimates when examining the performances of both HLLs and FLLs combined and separately. In regard to the second assumption, which was focused on the appropriateness of the configuration of the items on the C-test, Cronbach’s Alpha coefficients were examined if items were deleted. It was found that Texts 2, 3, 4, and 6 provided the most reliable estimates. Furthermore, the Rasch Model analysis also indicated that these four items were the most appropriate configuration for the C-test given that they showed good item fit statistics and a wide range of item difficulty levels.

In addition, IRT measures offered partial support for the third assumption in the generalization inference, which was concerned with the sufficient number of items included on the C-test. The support is described as partial because, as was reported in Chapter 7, while IRT measures showed that the C-test was able to accurately and reliably assess Korean language learners (both HLLs and FLLs) and distribute them into a wide range of ability levels, there were no items on the C-test that could specifically target very high performing learners. In other words, one more item with a higher difficulty level could have been included to target the exceptionally good C-test performers. It is, however, crucial to point out that the primary intended use of the C-test was not to
discriminate language abilities among extremely high-proficiency level learners. Rather, it was to distribute learners from a wide range of ability levels, which it successfully accomplished with only four items.

Finally, a DIF analysis of the C-test items provided backing to support the fourth assumption, which was relevant to the consistency of the C-test in measuring both HLLs and FLLs. More specifically, DIF contrast of the items indicated there was no noticeable DIF detected between the performances of the HLLs and FLLs in any of the items. In other words, all items included on the C-test functioned similarly in assessing Korean language learners’ proficiency regardless of whether they were HLLs or FLLs. As discussed in the previous section, the results of the independent samples t-test, which indicated that HLLs had performed statistically significantly better than the FLLs, raised concerns about the possibility of the C-test assessing the learner groups in different ways. Nevertheless, results of the DIF analysis indicated that the C-test items were not biased towards a specific learner group and were able to assess both HLLs and FLLs in the same way. Accordingly, the Korean C-test scores were able to reliably and consistently reflect both HLLs’ and FLLs’ general language proficiency.

8.1.3.1. The rebuttal to the generalization inference

One important contribution of this dissertation study is the examination of the rebuttal of the main claim in generalization. As displayed in Figure 31, one possible rebuttal to the warrant in generalization was that the definition and classification of HLLs was not accurate. It was explained in Chapter 7 that the initial definition of HLLs and thus the criteria to categorize a learner as an HLL was based on definitions presented in previous literature (see Kagan, 2005; Valdés, 2005). However, for a more empirically-
based and bottom-up classification of learners, a hierarchical cluster analysis (HCA) was conducted using 10 variables describing Korean learners’ speaking and writing proficiency, language learning background, and frequency of language use in different contexts. The results of the HCA showed that a two-cluster solution and four-cluster solution were the most efficient in describing the Korean learners in the sample. The two-cluster solution was composed of a group of 34 learners that had been originally classified as HLLs and another group of 58 learners that had been mostly classified as FLLs. When comparing the two-cluster solution with the initial classification of HLLs and FLLs, there were only six learners who had been classified as HLLs but that the HCA had identified to be more similar to learners in the FLL group. In other words, there was partial backing for the rebuttal in the generalization inference. That is, the initial definition and classification of HLLs and FLLs was not completely supported empirically.

However, the results of the HCA were not sufficient to weaken the strength of the Korean C-test validation argument, given that it was also necessary to examine whether the C-test was able to reliably assess the general language proficiency of the two new learner groups in the two-cluster solution. In other words, this rebuttal could be described as being composed of two questions: (a) one challenged the accuracy of the original classification of learners into the HLL or FLL group; and if the first question was answered affirmatively (b) a follow-up question, which was whether the C-test could nonetheless reliably assess these new group of learners without noticeable biases. As a result, reliability and IRT analysis for the new groupings was conducted again and findings still supported the generalization inference. In other words, items on the C-test
were found to have a high-level of internal consistency and showed good item fit to the Rasch model, even for the newly configured learner groups. In addition, no noticeable DIF contrast was found for any of the items when comparing the C-test performance of HLLs and FLLs as redefined by the HCA.

8.1.3.2. Understanding the different learner profiles and their performance on the C-test

As mentioned in the subsection above, the HCA also revealed a four-cluster solution, which classified learners into four distinct learner groups. This cluster solution provided detailed information about the types of learners represented in the data and raised important questions about the usefulness and reliability of the C-test in assessing these distinct learner groups. The first cluster (Cluster 1) was composed of learners who were initially identified as HLLs. These learners can be described as prototypical Korean HLL. They showed intermediate levels of speaking and writing proficiency, had little experience studying in Korea, and used the Korean language with some frequency in different contexts but mostly with their parents and for entertainment. Cluster 2 can be characterized as prototypical Korean FLLs. These learners also showed intermediate levels of proficiency in speaking and writing. However, they had more experience studying abroad in Korea. As with Cluster 1, they indicated using Korean language with some frequency in various contexts, but those excluded the use of language with family as they had no Korean family connections. Cluster 3 was composed mostly of FLLs with the exception of the six learners who had initially been classified as HLLs. These Cluster 3 learners mostly showed beginner levels of proficiency in speaking and writing. They had no experience studying abroad in Korea and they rarely used Korean in any of the
contexts surveyed. Finally, Cluster 4 was composed of advanced-level HLLs. They were all sequential bilinguals who had come to the United States at a later age (i.e., between the age of 5 to 15). They showed advanced and superior levels of proficiency in speaking and writing. All six reported using Korean with great frequency with multiple family members and across all contexts, even including reading. Four of these six learners had been previously identified by the Rasch model as the highest C-test performers. The remaining two HLLs in this cluster had performed at an advanced but not exceptional level.

When examining the C-test performances across all four clusters, it was found that although the C-test was not included as a variable for the cluster analysis, there was only a small variance in the C-test performance within each cluster. These results suggested that the learners within each cluster showed similar language proficiency as measured by the C-test. Furthermore, the reliability measures for each cluster were also high, which meant that the C-test was able to consistently and reliably measure each cluster’s general language ability. The only exception for high reliability coefficients was Cluster 4. This group was composed of six HLLs (four at exceptionally high levels and two at advanced levels) who had performed at high levels of proficiency in both ACTFL tests. Based on the descriptive statistics, there was almost no variance in C-test performance among these learners, with very low standard deviations and a very small range between the minimum and maximum scores. Therefore, the low reliability coefficient for this group indicated not only that the sample size was small but also that

---

8 Unfortunately, one of the exceptionally high-performing HLL on the C-test could not be classified in any of the clusters, as there were missing responses in her background questionnaire. This learner indicated being born in Korea and staying only until finishing her kindergarten education before moving to the UK.
the Korean C-test had not been able to reliably distinguish the general language ability of learners within this group. This outcome illustrates the ceiling effects of the C-test and agrees with the results of the Rasch model and the lack of an item with higher difficulty level that would target these advanced C-test performers. It is, however, important to reiterate that while the C-test was not highly reliable in distinguishing language abilities within advanced proficiency level learners, it was highly-reliable when distinguishing language abilities of learners with a wide range of ability levels.

In addition, it is worth noting again that for the HCA not only were factors describing learners’ Korean language learning experience and the frequency of Korean language use taken into account, but also variables related to their language proficiency (i.e., scores on the ACTFL OPIc and WPT). For this reason, the clustering of learners revealed that there were two HLL groups, one with mostly intermediate-level proficiency (Cluster 1) and one with advanced-level proficiency (Cluster 4), and two FLL groups, one with mostly intermediate-level proficiency (Cluster 2) and one with mostly beginner-level proficiency (Cluster 3). In other words, there were no groups of learners identified as mostly beginner-level HLLs and advanced-level FLLs.

In turn, these ‘missing’ groups suggested that any inferential statistics results based on the direct comparison between the HLL group (Cluster 1 and 4 combined) and FLL group (Cluster 2 and 3 combined) that does not take into account proficiency differences (like IRT measures do) should be interpreted with caution. The only direct comparison where the proficiency factor could be controlled for was between Cluster 1 and Cluster 2 (in the four-cluster solution), as both group of learners mostly showed intermediate-level proficiency in speaking and writing. As reported in Chapter 7, there
were no statistically significant differences in the C-test performance between these two intermediate groups, suggesting that the C-test consistently assessed both learner groups.

Another important finding of the HCA was the unexpected classification of some HLLs. There were six learners in Cluster 3 who had originally been categorized as HLLs but were found to be more similar to FLLs on the variables used for the HCA. These six learners were ethnically Korean. Five of them had been born in the United States and one indicated her age of arrival to be one-month old. While they exhibited different proficiency levels that ranged from novice-high to intermediate-mid, they shared a very similar learner profile in terms of Korean language use with their parents. Unlike the other HLL-clusters (Cluster 1 and Cluster 4), they indicated minimal use of Korean when considering the frequency of Korean language use with both parents combined. That is, these learners described the interaction with their parents as (1) talking to and being talked to minimally in Korean, (2) talking minimally but being talked to with some frequency in Korean, and (3) talking to and being talked to with frequency in Korean but only by one parent. The majority of these learners indicated using Korean very minimally to talk to other family members, such as siblings and grandparents. Furthermore, they reported using Korean with little frequency mostly just for watching Korean programs and listening to music but almost no use in any other context. They also had no experience studying Korean in Korea.

Despite having Korean in their family heritage, these characteristics set the six learners apart from other HLLs. This is in line with bilingualism research that describes the role of parent language input on the language development of bilinguals. De Houwer (2007, 2009, 2011), for example, described bilinguals who grew up with frequent
minority language input from both parents as different from those bilinguals who grew up with minimal or with only one parent input. Moreover, it is worth pointing out that these six learners did not form their own cluster as in the case with the exceptionally good HLLs in Cluster 4. Instead, they were grouped together with the FLLs in Cluster 3, which suggested that these six learners behaved in very similar ways to these FLLs.

Overall, in addition to providing partial backing for the rebuttal in the generalization inference, the results of the HCA also raise important questions about the definition and classification of HLLs for research in general. First, these results underscore the importance of considering multiple factors when defining or classifying learners as HLLs or FLLs, beyond age or onset of exposure to the majority language. These factors include not only the language use in the family but also learners’ ability in speaking and writing, as well as the frequency of language use across different contexts. Second, the results also point out the importance of measuring language use or language contact as a continuous variable rather than as a dichotomy. For both HLLs and FLLs, the cluster analysis found different types of learners according to the frequency of Korean language use. This challenges some of the common methodologies of classifying learners into different learner types. For instance, as was described in section 2.3.1, one quick way to group learners into HLLs and FLLs for placement purposes reported in some studies (e.g., Beaudrie & Ducar, 2012; Thompson, 2015) has been through yes/no statements on family background and family language use. While these statements might provide an accurate classification of learners in a broad sense, they might not be as precise when researching HLLs and trying to investigate their language development. Third, the different clusters found in this dissertation study also call for a more empirically-based
methodology for describing language learners, in particular HLLs. As was shown with
the HCA in the present study, a more empirically-based methodology using multiple
variables that can describe HLLs can yield groupings of HLLs that are homogenous and
that help explain some of the underlying factors influencing HL development.

8.1.4. Explanation

Explanation was mainly focused on investigating whether learners’ performance
on the Korean C-test could be attributed to their general language proficiency. For this
reason, it was important to find backing that would demonstrate that the items on the C-
test were able to target different proficiency levels and that learners’ performance on
these items could be explained by their general language abilities, particularly their
literacy-based language skills. Accordingly, there were three assumptions underlying the
explanation warrant: (1) the difficulty of the C-test items is systematically influenced by
the item characteristics regardless of learner type; (2) performance on the C-test is closely
related to performance on other literacy-based tests of language proficiency as would be
expected theoretically; and (3) performance on the C-test can be explained by learners’
speaking and writing proficiency. Figure 32 displays the warrant, three assumptions and
backing for explanation. It also illustrates the possible rebuttal to the main claim about
attributing C-test scores to learners’ general language abilities. As will be explained later
in this section, no substantial backing was found to support this rebuttal.
Figure 32. Representation of explanation with three assumptions, backing, and rebuttal

Backign for the first assumption in explanation came from the detailed analyses of C-test item characteristics. Results indicated that the difficulty of C-test items was systematically influenced by the number of sentences per item (i.e., passage), number of clauses per sentence, number of function and content words within the item, and vocabulary level of the words within the item. More specifically, based on both HLLs’ and FLLs’ C-test scores, it was found that more difficult items had a smaller number of
sentences but more clauses per sentence, more content words deleted, and that the words within the passage were higher in vocabulary level according to the NIKL vocabulary list. These item characteristics holistically represented a range of language features in morphology, syntax, and vocabulary. In addition, it was explained in Chapter 7 that the difficulty level of each micro-item (i.e., blank) generally agreed with the difficulty of the whole item. That is, when mapping out all micro-items on a single logit scale, it was found that micro-items within the easiest C-test passage were usually the easiest and the micro-items within the most difficult passage were usually the most difficult. These results were in line with previous research which described C-test micro-items to be locally dependent (e.g., Eckes, 2011; Eckes & Grotjahn, 2006; Klein-Braley, 1985; Lee-Ellis, 2009; Norris, 2006). At the same time, the results also suggested that in order for learners to successfully complete each gap, they needed to process and comprehend the grammar, morphology, and vocabulary of the C-test passage as a whole. Additionally, it was also found that these item characteristics systematically affected item difficulty in similar ways for both HLLs and FLLs. A high Spearman rank order correlation coefficient of item measures was found for both learner groups. While research comparing HLLs and FLLs on their performance on tests targeting discrete language features (e.g., anaphors, relative clauses, sentence agreement, tense-aspect/mood, among others) has suggested that HLLs outperform FLLs (Montrul, 2008, 2009, 2010), particularly on structures acquired before the development of metalinguistic skills, like tense-aspect/mood (Montrul & Perpiñán, 2011), plural nouns (Albirini & Benmamoun, 2015), or nominative markers (Laleko & Polinsky, 2013), the difficulty of the C-test items appears to have affected both groups in similar ways given their holistic nature.
Furthermore, strong positive correlation coefficients between learners’ C-test performance and their performance on the ACTFL WPT provided support for the second assumption. That is, C-test performance was found to be closely related to learners’ literacy-based skills. Learners with advanced levels of writing proficiency were able to achieve high scores on the C-test. One caveat to this finding is that only one indicator of productive literacy was used in the current study (i.e., the ACTFL WPT scores). It is recommended that future studies also include a measurement of reading proficiency, which would add information on the relationship of C-test performance and a receptive literacy skill. Moreover, it is also important to discuss the results of the comparison between the HLLs and FLLs on the ACTFL WPT because a difference in learners’ performance on this test could indicate that the test was behaving differently towards these learner groups. Even though an independent samples t-test found that there was a statistically significant difference between the HLLs and FLLs on the writing test, these results might have been conflated with the effects of the different sample sizes in each proficiency-level subgroup. When the analysis included only the intermediate-level clusters (Cluster 1 and Cluster 2 in the four-cluster solution), no significant difference could be found between the learner groups in their writing test performance, suggesting that the test was overall assessing both learner types in similar ways.

Finally, a multiple regression analysis found that both speaking and writing proficiency significantly predicted learners’ C-test performance, providing backing for the third assumption. It is important to note that in the current study the construct of general language proficiency was defined as a global measurement of learners’ language ability in both spoken and written modalities, and these language skills were assessed by
the ACTFL OPIc and the ACTFL WPT. The results of the multiple regressions indicated that 80% of the variance in learners’ C-test performance closely related to their performance on both speaking and writing tests. When examining each learner group separately, it was found that for HLLs, writing ability was somewhat more important in explaining C-test performance than speaking ability was. However, for FLLs, both speaking and writing were almost equally important when explaining C-test performance.

Before further discussing the results of the multiple regression, it seemed important to also reexamine the descriptive statistics of the ACTFL proficiency tests for each learner group in order to understand the differences in learners’ language abilities according to modality. First, as with the C-test, HLLs performed better than the FLLs in both ACTFL tests. Second, a comparison between the ACTFL OPIc and the ACTFL WPT for each learner group indicated contrasting trends. While the HLLs performed better on the speaking test than on the writing test, the FLLs showed the opposite trend, scoring higher on the writing test. This trend has been commonly reported in previous literature, where HLLs have been found to have asymmetric proficiency levels in speaking and writing (Friedman & Kagan, 2008; Kondo-Brown, 2003, 2004, 2010; Montrul, 2008, 2016; Montrul, Davidson, De La Fuente, & Foote, 2014; Montrul et al., 2008). Given that the majority of the HLLs acquire their HL in naturalistic settings, they are usually more exposed to the spoken modality of the language when compared to the written modality. This in turn contributes to the “disassociation in the linguistic knowledge by skills” (Montrul, 2016, p. 48). For example, HLLs are usually described as having advanced-levels of listening skills more than any other language skill. On the other hand, when FLLs start learning the language as adults, they encounter it in a
classroom context, and therefore are exposed to the language in both modalities, or even more often in the written modality than the spoken modality.

Along the same lines, the results of the multiple regression also described the HLLs’ unbalanced levels of proficiency by skill. As was mentioned before, although it was found that speaking and writing proficiency were both significantly predictive of C-test performance, the relative importance of each skill differed according to learner group. For the HLLs writing proficiency was more closely related to the variance found in the C-test scores whereas for the FLLs both writing and speaking were equally important to explain this relationship. A closer look into the characteristics of ACTFL OPIc test scores revealed that the majority of the HLLs scored between the Advanced-Low and the Advanced-High levels with a mode of 2.1 (i.e., score equivalent to the Advanced-Low proficiency level). In contrast, the scores on the ACTFL WPT showed a wider spread of scores across different proficiency levels. Similarly, the ACTFL scores for the FLLs were more spread across different proficiency levels, particularly Intermediate-Low to Intermediate-Mid for both speaking and writing tests, with a mode of 1.1 (i.e., score equivalent to Intermediate-Low proficiency level) and 1.3 (i.e., scores equivalent to the Intermediate-Mid proficiency level), respectively. This difference in variability of scores could help explain why writing proficiency turned out to be a more significant predictor of C-test performance for HLLs but not for FLLs. Given that the majority of HLLs were advanced-level speakers of Korean, this skill contributed less in explaining the variability of the scores in other tests, such as the C-test. In other words, not all HLLs with advanced-level speaking proficiency were likely to be good C-test performers, whereas those with advanced-level writing proficiency were also good C-test
performers. In the case of FLLs, there was similar wide variation in both speaking and writing tests scores, and thus both helped predict C-test performance. Nevertheless, as with HLLs, writing proficiency was a slightly better predictor than speaking proficiency in explaining the C-test performance.

These findings suggest that although the Korean C-test is accurate and reliable in distributing learners into different levels of proficiency and in providing a global measurement of language proficiency, the underlying discrete skills being measured may vary depending on the learner type. This phenomenon is analogous to Sigott’s (2004, 2006) concept of construct fluidity. He explained that the nature of language proficiency itself is fluid, and not all dimensions of language develop linearly, continually, and equally across all language learners. In the same manner, the discrete language skills being measured by the Korean C-test could differ across different learner types. For example, for HLLs, the C-test may be better at assessing grammatical structures that target their metalinguistic skills (e.g., clause markers, spacing rules). Verbal protocol studies might help clarify what underlying strategies HLLs and FLLs are using and whether there are any differences in the way they approach the completion of C-test passages.

8.1.4.1. Appropriateness of ACTFL tests for assessing heritage language learners

Another important component of explanation to consider was the rebuttal. Figure 32 displays a possible rebuttal to the claim that C-test performance could be attributed to learners’ general language proficiency. As aforementioned, in the current study the indicators of general language proficiency were a combination of learners’ performance
scores on the ACTFL OPIc and ACTFL WPT, given than these tests targeted both learners’ speaking and writing skills, respectively. Nevertheless, using ACTFL proficiency tests to measure HLLs is still a contended practice. Some researchers have questioned whether these standardized tests, which were originally designed to assess FLLs, would be appropriate tools to assess HLLs. Due to the Language Testing Institute’s (LTI) restricted access to test-takers’ speech or written samples for each test item, however, evidence in support of or against this rebuttal could not be examined directly.

Without access to these data, one alternative method to explore the claim made in the rebuttal is to review previous studies that have looked directly into the speech samples of ACTFL OPIc test-takers. Three such studies, Kagan and Friedman (2003), Martin et al. (2013), and Swender et al. (2014), have analyzed Russian and Spanish HLLs’ speech samples on the ACTFL OPI and OPIc and have indicated that the test scores accurately represent HLLs’ language skills at each proficiency level. The researchers analyzed the speech sample characteristics at each ACTFL proficiency level and specifically examined the language functions that prevented HLLs from receiving the next highest proficiency level. They found, for instance, that what prevented Intermediate-Mid HLLs from scoring at an Intermediate-High level was their lack of structural control and limited vocabulary (Swender et al., 2014, p. 432). Moreover, they also noted that even though HLLs tended to have better fluency and pronunciation than FLLs, such characteristics were not enough to “compensate for lack of sustained functional ability as defined by the ACTFL Proficiency Guidelines 2012” (Martin et al., 2013, p. 220). Accordingly, these studies found no evidence that indicated that the
ACTFL proficiency tests were unable to assess HLLs with accuracy. They asserted that, for purposes of assessing learners’ functional ability, the ACTFL proficiency guidelines are appropriate for both HLLs and FLLs.

Moreover, it is worth noting that the same standards and principles are used to assess learners and train raters for ACTFL tests across different languages. In addition, in order to become a certified ACTFL rater in any of the languages, the aspiring rater must undergo rigorous training, which not only includes workshops to understand and internalize the ACTFL proficiency guidelines but also a certification process where raters are trained by rating real learner samples (ACTFL, n.d.). Raters also participate in norming sessions to ensure score accuracy and reliability. Therefore, the studies mentioned above, although focused on Russian and Spanish, provide a partial counterclaim to the rebuttal, given that they describe the situation with the ACTFL proficiency guidelines in general. Unlike these studies that have empirically looked into the language features in learners’ speech sample, however, there have been no empirical studies that have provided evidence of the inappropriateness of these guidelines for assessing HLLs. In other words, although some scholars have descriptively examined the ACTFL proficiency guidelines to argue against their use for assessing HLLs, there have been no studies that have provided direct evidence of learners’ speech samples showing the inappropriateness of the tests, guidelines, or rater-training to assess HLLs. This work remains to be done.

Having said this, therefore, and as can be observed in Figure 32, the results of previous studies only represent a partial counterclaim to the rebuttal given that there was no direct evidence to reject the claim made in the rebuttal. A descriptive analysis of the
statistics in the current study found that both ACTFL tests were able to distribute Korean language learners, both HLLs and FLLs, into a wide range of proficiency levels. The distributions of scores for both tests, however, were slightly negatively skewed, particularly for the ACTFL WPT. For example, no learner (HLLs or FLLs) scored at a Novice-Low or Novice-Mid level. Nevertheless, these results might have also been affected by the fact that one of the requirements for participants in this study was to have been enrolled in a Korean language course higher than the first level of Korean. In other words, there might have been no true beginners in the sample. While the results of the descriptive statistics of both tests contribute to the partial counterclaim of the rebuttal, an analysis of the data describing individual learner’s speaking and writing test samples would be necessary to provide direct support or opposition to the rebuttal. Without these data, it would be difficult to make any inferences on the appropriateness of the ACTFL tests to assess Korean HLLs. For this reason, future studies using ACTFL tests that get direct access to learners’ responses should try to examine the differences between Korean HLLs’ and FLLs’ performance and whether these differences affect the scores that they receive.

8.1.5. Extrapolation

The extrapolation part of the validity argument was mainly concerned about the extent to which the construct of general language ability as measured by the C-test was related to other language proficiency criteria reflective of general language proficiency. Generally, in validity studies that implement the ABA, extrapolation examines the connection between test performance and sample performances of the target domain. Nevertheless, given that the Korean C-test was not design to be domain-specific, the
relationship between C-test performance and other criterion measures of language proficiency was examined. The language proficiency tests used as criterion measures included the ACTFL OPIc, ACTFL WPT, and the Korean Elicited Imitation Test (EIT; Kim et al., 2016). Furthermore, learners’ self-assessment ratings of Korean language proficiency were used as an additional criterion measure. The ACTFL tests were chosen as criterion measures because they represented standardized measurements of language proficiency that are commonly used for language teaching and research. The EIT was also included as another criterion measure because it provided a global measurement of learners’ oracy skills. Underlying the extrapolation were three assumptions: (1) test-takers’ performance on the C-test has a positive relationship with other tests of language proficiency; (2) C-test scores can be extrapolated and interpreted in terms of other language criteria, namely the ACTFL proficiency guidelines; and (3) C-test performance has a positive relationship with learners’ ratings of their own language proficiency. Accordingly, backing for the assumptions came from criterion-related correlations calculated between the C-test and the criterion measures mentioned above as well as C-test and self-assessment ratings. For all correlational analyses performance scores were examined for HLLs and FLLs combined as one group as well as separately. Figure 33 represents the extrapolation inference with its warrant, assumptions, and backing.
Strong positive Spearman rank order correlation coefficients between C-test performance and the ACTFL OPic, ACTFL WPT, and EIT provided support for the first assumption. C-test performance was overall strongly correlated (at .81 or higher) with all criterion measures including those that reflected learners’ Korean language proficiency in speaking and oracy skills. This pattern is in line with previous C-test research that has found that C-test scores can be strongly associated with measures of other discrete language skills (e.g., Harsch & Hartig, 2015; Karimi, 2011). In addition, similar to the
results of the multiple regression (see section 7.4.2 and 8.1.4), correlation coefficients suggested that although both speaking and writing skills were strongly correlated with C-test performance, the latter skills were the most strongly associated with the C-test outcomes. This finding provided further evidence that C-tests are more closely related to literacy-based skills rather than oracy-based skills. Furthermore, while the C-test and the EIT were both shortcut measures of language proficiency, they were found to correlate closely to different criterion measures. In the case of the C-test, it correlated more strongly with the writing test than the speaking test. The EIT, on the other hand, showed the opposite trend, correlating more strongly with the speaking test than the writing test. Interestingly, when compared to the correlation coefficient between C-test and ACTFL OPIc, C-test were found to be more closely associated with learners’ performance on the EIT. This is likely because, despite the differences between the C-test and the EIT in assessing different language skills in terms of modality (i.e., literacy-based skills and oracy-based skills, respectively), they were both intended to measure more global aspects of language proficiency. The strong positive correlations between these shortcut measures suggest that they share some underlying constructs, such as understanding and reconstructing the passage or stimulus.

The most important and the strongest support for the main claim being made in extrapolation was the association between C-test scores and the average of the two ACTFL test scores. Among all correlation coefficients between C-test and criterion measures, it was found that C-test performance was the most strongly correlated with the mean of the ACTFL OPIc and ACTFL WPT. Similar to the findings of previous C-test studies that have suggested that C-test measure a global construct of language proficiency
(e.g., Babaii & Ansary, 2001; Babaii & Jalali Moghaddam, 2006; Grotjahn, 1987; Klein-Braley, 1994; Eckes & Grotjahn, 2006; Raatz, 1984), it was found that the Korean C-test was most strongly associated with the combined mean scores of speaking and writing proficiency. In other words, learners who scored high on the Korean C-test, also scored high on the speaking and writing tests combined.

Similar correlational trends were found for both learner groups when analyzing them separately. This finding indicated that regardless of the learner type, C-test performance was strongly correlated with the mean scores of the speaking and writing skills combined. Also, for both HLLs and FLLs, C-test performance was more highly correlated with writing proficiency than with speaking proficiency. In addition, in order to control for proficiency effects, correlation analyses between only the intermediate-level HLLs and FLLs (i.e., cluster 1 and cluster 2 in the four-cluster solution) were also conducted. Results indicated overall very similar trends, for both HLLs and FLLs. One difference, however, was that for the HLLs C-test performance was more strongly correlated to the mean of ACTFL tests and the ACTFL WPT scores than it was for the FLLs. Furthermore, as indicated in the multiple regression analysis, for FLLs, C-test performance correlated very similarly in magnitude to the ACTFL OPIc and ACTFL WPT, whereas for the HLLs, the correlation was much higher between the C-test and the ACTFL WPT. These findings also suggested that the HLLs have disassociated language skills, with more learners being better at oracy skills than literacy skills (Friedman & Kagan, 2008; Kondo-Brown, 2010; Montrul, 2016; Montrul, Davidson, De La Fuente, & Foote, 2014).
Another important part of the extrapolation inference was to find the direct relationship between C-test scores and other proficiency guidelines that are commonly used in Korean language learning research. Concurring with the results of the correlation analyses, it was found that learners’ C-test performances followed the same trend as the ACTFL proficiency guidelines, providing backing for the second assumption in this inference. That is, as learners’ ACTFL proficiency levels increased so did their C-test scores. When classifying learners into the main ACTFL proficiency level groups (i.e., Novice, Intermediate, and Advanced), their average C-test scores were also found to be statistically significantly different. In turn, these findings helped set up C-test cutoff scores for easier interpretation.

Finally, another critical aspect explored in the extrapolation inference was the relationship between C-test performance and learners’ perceptions of their Korean language proficiency. Based on previous analyses of L2 self-assessments (Ross, 1998) and L2 correlational studies (Plonsky & Oswald, 2014), it was found that there were overall moderate to strong correlation coefficients between C-test performances and Korean learners’ language proficiency self-assessment ratings, which provided backing for the third assumption in the extrapolation inference. Learners were asked to self-rate their proficiency on reading, listening, speaking, and writing according to a four-point Likert scale. Correlational analyses indicated that learners’ C-test performances were the most strongly correlated with the average of the self-ratings for all four skills. Like the results of the correlational analyses between C-test and the mean of both ACTFL tests combined, these results indicated that the Korean C-test could be associated with a combination of multiple language skills. Furthermore, when examining the correlation
coefficients for learners’ self-perception of each language skill separately, it was found that reading had the strongest correlation coefficient with C-test performance followed by listening and speaking. The weakest correlation, on the other hand, was with learners’ self-assessed proficiency on writing. Given that there was a strong positive correlation between C-test performance and learners’ actual writing proficiency as measured by the ACTFL WPT, it appears that learners were inaccurate in assessing their own writing ability.

In addition, self-assessment ratings were also analyzed separately for HLLs and FLLs. While HLLs self-rated their oracy skills as their strongest and literacy skills as their weakest, FLLs rated their literacy skills as their strongest and oracy skills as their weakest. This trend is in line with numerous studies that have reported that HLLs generally feel more confident about their speaking and listening skills, whereas FLLs feel more confident about their reading and writing skills (Carreira & Kagan, 2011; Kagan & Dillon, 2012; Hedgcock & Lefkowitz, 2011; Montrul, Bhatt, & Girju, 2015). These findings also concur with previous research that has found that HLLs are greatly concerned about improving their writing skills, whereas FLLs are mostly concerned about improving their speaking skills (Callahan, 2010; Colombi & Roca, 2003; Hedgcock & Lefkowitz, 2011). In addition, when comparing self-ratings across different language skills, HLLs appeared to have underestimated their writing abilities and FLLs their speaking abilities. These perceptions could be influenced by the language educators’ views on learners’ performance in the classroom. Hedgcock and Lefkowitz (2011) pointed out that language instructors’ methods and procedures of teaching writing are focused on learners’ grammatical and lexical accuracy, particularly for HLLs.
Furthermore, they also indicated that instructors seemed to hold HLLs to higher standards and presuppose HLLs are more proficient in the HL. These types of educator beliefs might reinforce HLLs’ lack of confidence on their writing and literacy skills in general. In regard to the FLLs, their lack of confidence in oracy skills might be due to performance anxiety (see Horwitz & Young 1991; Young 1999) and perhaps to comparing their speaking and listening skills with those of their HLL-cohorts’ skills in mixed classrooms. As most of the HLLs have been exposed to the HL in natural settings, they tend to be more fluent when speaking and show a native-like pronunciation of the HL (e.g., Au et al, 2002; Oh et al., 2003). For this reason, FLLs might set themselves higher standards when it comes to speaking and listening.

In terms of the correlation analyses between C-test scores and self-assessment ratings by learner group, results suggested that HLLs were slightly more accurate in their ratings than FLLs. In particular, there was a much stronger correlation between C-test performance and overall language ability (i.e., mean scores of self-ratings on all four language skills) for HLLs than for FLLs. One of the reasons for this difference could be the nature of the items and scale in the self-assessment questionnaire, which were very broad and might have been interpreted differently by the two learner groups. A closer look at individual responses indicated that FLLs tended to rate themselves similarly with 1 or 2 for speaking and writing. Meanwhile, HLLs made use of the full range of the scale with more variation from 0 to 3. Future studies should gauge learners’ perception of their language proficiency with more detailed items in the questionnaire, like can-do statements, and use a more continuous scale, perhaps using sliding scales instead of discrete four-point scales.
Moreover, although more research should be conducted on the ability of HLLs in self-assessing their language abilities (Donovan et al., 2012; Kang & Kim, 2012; Martin et al., 2013; Swender et al., 2014), some studies on Korean HLLs have pointed out that learners’ perceived HL ability is generally strongly correlated to their actual HL proficiency. Kang and Kim (2011), for example, found strong to moderate correlation coefficients between Korean HLLs’ perceived and actual proficiency in speaking and writing. They added that the correlations seemed to be moderated by learners’ degree of ethnic identify. That is, those who identified themselves closer to the Korean culture appeared to overestimate their language abilities whereas those who identified less with the culture underestimated their abilities. Along similar lines, Gatti and O’Neil (2017) found strong positive correlations between Korean HLLs’ self-perceived proficiency in writing and their actual performance on the ACTFL WPT. Additionally, they indicated that when compared to Chinese and Spanish HLLs, Korean HLLs had the strongest correlation between perceived and actual writing proficiency.

8.2. Limitation and Suggestions for Future Studies

In addition to summarizing the most important results of the Korean C-test validity argument, this chapter also highlighted several methodological limitations in the study. In this section, these limitations will be briefly explained as suggestions for future studies that intend to focus on validating the uses of shortcut measures of language proficiency for research purposes. First, despite efforts to recruit a large number of Korean HLLs and FLLs, with a wide range of proficiency levels, the sample size was small for the purposes of a large-scale validation study. In terms of the multiple
subgroups formed based on learner type and learner proficiency, the sample size within each subgroup was rather small. For example, as was described in Chapter 7 and later in the current chapter, there was a lack of beginner-level HLLs and advanced-level FLLs. For more robust statistical results, it is recommended that subsequent studies include a larger sample size, particularly of those specific learner groups.

Second, when analyzing the relationship between the Korean C-test constructs and other constructs of Korean language proficiency, the only literacy-based measure included in this study was learners’ writing proficiency. Although writing proficiency as assessed by the ACTFL WPT provided a reliable measurement of learners’ productive literacy skills, an additional measurement of receptive literacy skills, like reading, could have contributed with additional backing for the explanation and extrapolation inference. Unfortunately, the logistics and costs of administering an additional standardized test impeded the inclusion of a reading test in the current study. It is suggested that future studies also include a measure of reading as well as of listening in order to evaluate learners’ receptive literacy and oracy skills. These tests results will provide a more comprehensive view of what the Korean C-test actually measures.

Third, despite adequate backing for the explanation inference obtained through criterion measures of language proficiency, conducting verbal protocols might bring further clarity in regard to C-test constructs. This type of data would help examine the strategies that both HLLs and FLLs use when completing C-test items. At the same time, it would provide a more accurate view of the extent to which different learner types approach the completion of the mutilated words in distinct or similar ways.
Fourth, as was described beforehand, the present study was not able to determine with certainty whether ACTFL tests were appropriate for assessing HLLs. Overall, based on score distributions, the tests seemed to be appropriate for indicating the functional language ability of both learner types. Nevertheless, future studies should also examine the linguistic features within learners’ speech and writing samples from the ACTFL tests. With these language samples, it will be possible to go beyond learners’ functional language ability and examine the relationship between C-test performance and discrete linguistic competencies (e.g., vocabulary knowledge, grammatical accuracy, sentence complexity).

Fifth, another point of improvement is the design of the self-assessment questionnaire. As described in Chapter 6, the self-assessment questionnaire was a small part of a larger background questionnaire, and thus included only four items targeting four language skills. A more comprehensive self-assessment questionnaire that includes detailed can-do statements, which are more specific in describing discrete abilities within a language skill (e.g., speaking, writing) will offer a more precise overview of how learners perceive their own language proficiency. In addition, more detailed statements will reduce the error introduced by learners’ different interpretation of the statements. In other words, the more specific the statements in the self-assessment questionnaire are, the more consistent the interpretation of those statements will be across all learners.

As was mentioned in Chapter 5, this study was not intended to be a standalone study but rather one that would open the path for more research on shortcut measures of language proficiency. It is hoped that the Korean C-test will be widely used in Korean language learning research and that further validation efforts will continue to improve
upon the test proposed in the current study. While several suggestions were offered in this section to guide future research, test evaluators and applied linguistics researchers are encouraged to challenge the claims made in the validation argument and provide additional backing for rebuttals to the argument.

8.3. Conclusion

The current dissertation study sought to validate the use of a Korean C-test (Son et al., 2018) for providing a quick measurement of Korean language learners’, both HLLs’ and FLLs’, general language proficiency for applied linguistics research. Following Kane’s (2006, 2011, 2013) Argument-Based Approach to validation, the study evaluated the strength and coherence of the validity argument by examining the claims made in five different inferences, namely, theoretical grounds, evaluation, generalization, explanation, and extrapolation. As discussed throughout this chapter, generally adequate backing was found to support the claims underlying the warrants in the inferences. Therefore, the present dissertation study recommends the interpretation of Korean C-test scores as indicators of Korean general language proficiency of both HLLs and FLLs for research purposes.

Given that the validation efforts in this study were undertaken from a test evaluator’s perspective, rebuttals and backing (or lack of backing) for the rebuttals were also provided. These arguments were examined in terms of the how they could threaten the validity argument and whether evidence supporting these counterclaims could be found. While the Korean C-test validity argument included rebuttals that challenged the warrants in the evaluation, generalization, and explanation inference, no backing or only
minimal backing was found to support these rebuttals. Through the evaluation of these rebuttals, several caveats were identified for the use of the Korean C-test. First, it was discussed in the evaluation inference, that the Korean C-test use is limited to the assessment of a global construct of Korean language proficiency for distributing learners with a wide range of proficiency levels. The Korean C-test was found to be inappropriate for distinguishing learner abilities within a particular proficiency level, such as the very advanced proficiency level. Second, as discussed in the generalization inference, the classification of HLLs used in this study was based on empirical data relevant to learners’ Korean language learning experience and language use with family members and in different contexts. While there was evidence to support a different classification of learners into HLLs and FLLs based on results of a cluster analysis, the re-analyses of the reliability of items and generalizability of results across different learner groups remained the same. That is, the Korean C-test was able to assess both HLLs and FLLs in similar ways without noticeable biases. Third, throughout this dissertation study, it was emphasized that the interpretation of the Korean C-test scores should be limited to literacy-based skills due to the nature of the C-test items, which require learners to have basic reading and writing proficiency skills. In the explanation inference, evidence supported this assertion as it was found that the constructs measured by the C-test were more closely related to the literacy-based indicators of proficiency, like the ACTFL WPT scores. Fourth, the present study could not find evidence that could directly support or reject the rebuttal of using ACTFL proficiency tests to assess HLLs. While the ACTFL tests were able to distribute learners, both HLLs and FLLs, into a wide range of proficiency levels, some skewness was detected particularly for the HLLs when taking
the ACTFL OPIc. Without the analyses of the speech and writing samples, it was not possible to determine whether the HLL group represented in the present study included more learners with advanced proficiency in speaking, or whether the test itself was not able to reliably distinguish their abilities. Thus, findings based on the ACTFL tests should consider this caveat when examining the relationship between the Korean C-test performance and other language proficiency constructs.

Importantly, in Chapter 4, it was noted that one key advantage of Kane’s validation model is that it promotes a clear articulation of warrants, assumptions, and backing underlying the validity argument. In this way, it facilitates the identification of areas of agreement and disagreement that critics might have with the argument. Accordingly, future studies using the Korean C-test or evaluating the validity of its uses will be able to easily identify the limitations of the test or the claims in need of additional backing. Furthermore, building on the validity argument proposed in the current study, other evaluators of the Korean C-test could propose other rebuttals and present evidence to support them.

Finally, it is important to note that the validity argument presented in this dissertation study is incomplete given that the last step of the argument, the utilization inference, has yet to be evaluated. This last inference, represented as the last step of the staircase in Figure 29, is extremely important because it is at this stage where the Korean C-test washback is examined and where the appropriateness of the C-test score use and interpretation can be evaluated. However, this last step can only be assessed once the Korean C-test becomes fully operational. When a sufficient number of studies on Korean language learners use the Korean C-test as an indicator of Korean general language
proficiency, it will be possible to collect and synthesize data about test use and consequences of the decisions made based on test results. At the same time, these data will allow further improvement of the Korean C-test.

To conclude, it is hoped that this study will be an initial step toward more discussion on the importance of validating and using language assessment instruments in applied linguistics research. In particular, this study calls for more scholarly attention to the validation of instruments for assessing HLLs. While several studies have already taken steps to use test scores as indicators of HL proficiency, it is crucial that these studies also evaluate the appropriateness of these tests for measuring this specific learner group. Lastly and most importantly, this study also seeks to encourage a more consistent use of test instruments across studies in order to reach more generalizable and cumulative results.
Appendix A

Korean C-test - Texts and Answers

TEXT 1
안녕하세요. 우리 가족은 네 명이에요. 우리는 중국 사람이에요. 지금 한국에 살아요. 엄마는 학교에서 영어 선생님이에요. 그리고 아빠는 은행에서 일해요. 동생은 한국 영화를 좋아해요. 그래서 극장에 자주 가요. 동생은 도 한국어를 배워요. 우리 가족은 불고기하고 비빔밥을 좋아해요. 한국 음식은 맛도 좋고 비싸지 않아요. 그리고 건강에 아주 좋아요. 그래서 한국 요리를 배우고 싶어요.

TEXT 2
사랑하는 우리 딸 지원이에게
지금 엄마는 지원이가 준 편지를 보고 있어. 그동안 지원이가/는 엄마에게 여러 번 편지를 써서 지만, 엄마가 답장을 쓸 적이 없는데? 엄마가 미안해. 앞으로는 지원이에게 답장을 꼭 써야해. 엄마는 어제/런 지원이가 설거지를 해줘서 정말 고마웠어요. 지원이가 편지에서 '앞으로는 집안일을 도와 드릴게요'라고/는 약속을 하고. 엄마는 참 행복해. 다음에는 답장을 더 빨리 보낼게. 엄마가

TEXT 3
서울은 현재 한국의 수도예요. 조선 시대부터 지금까지 약 600년 동안 한국의 수도였어요. 지금 서울의 모습은 옛날과 많이 다르지만 아직도/은 옛날의 아름다운 모습을 여기저기에서 볼 수 있어요. 서울에는 대학교가 많고, 여러 가지 박물관과 옛날 건물도 많이 있어요. 서울은 아주 큰 도시예요. 현재 서울에 살고 있는
사람이 천만 명이 넘어요. 서울에는 산이 많아 있고 한강도 있어요.
다리가 스무 개쯤 있어요. 그래서 교통이 복잡해요.

TEXT 4
우리 속담 중에 '티끌 모아 태산'이라는 말이 있어요. 아무리 작은 물건이라도 조금씩 쌓아도 나중에 커진다는 뜻이지요. 돈도 그래요. 적은 돈이라고 아껴 모으면 큰돈이 될 수 있어요. 어서부터 저축하는 습관을 가지야 하는 이유에요. 그런데 최근 용돈의 대부분을 인터넷 게임 사로 머니를 구입하는 데 써 버리는 어린이가 늘고 있다고 해요. 용돈을 받으면 미리 어떻게 사용할지 계획을 세우고, 적은 금액이라도 사용 뒤에는 기록을 남기면 것이 중요해요. 이렇게 하면 꼭 필요한 곳에만 돈을 쓸 수 있게 되기 때문에 용돈을 함부로 사용하지 않고 저축할 수 있어요.

TEXT 5
자동차가 없을 때는 먼 거리를 걸어 다니야 했고 편의점, 김밥집, 패스트푸드점이 생기기 전까지는 집에서 도시락을 싸서 다녀야 했다. 그러나 자동차가 발명된 이후에로는 걷는 피곤함이 없어졌고 편의점, 패스트푸드점 덕분에 쉽게 쏙쏙 배고픔을 없앨 수 있게 되었다. 문제는 우리 가/는 편하게 되면서도, 전용, 환경오염이 더 심각해졌다는 점이다. 자동차의/는 배출한/과 공기 오염의 주원인이 되었고, 일회용품, 비닐 등은 오랫동안 잘 찢지 않아서 칭구를 오염시켰다. 그렇다면 이제 우리는 어떻게 해야 할까?

TEXT 6
구글뿐 아니라 페이스북, 야후, 애플 등 세계적인 기업들이 ‘집보다 더 쾌적한 사무실’을 내세우고 나서면서 새로운 경쟁을 하고 있다.
회사에 머무르는 것 자체를 즐겁게/기게 만들겠다는 뜻이다. 무료로 식사를/도 제공하고, 직원 전용 마사지센터와 피트니스센터도 설치했다. 이 같은 최고급/의 시설은/이 직원들의 창의성과 업무 효율을 높이는 데 효과가 있다고 한다. 그러나 반대의 의견도 거세다. 최고급 시설로 회사에 오래 머물게 해서 가정에서 보내야 하는 시간을 빼앗는 것이 아니냐는 의견이다. 한 인사관리 전문가는 “건물을 지어놓고 직원 모두를 사무실 안에 머물게 하는 것은 어처구니없는 일”이라며 보이지 않는 족쇄를 채우는 것이라고 비난했다.

**TEXT 7**

# Appendix B

## Elicited Imitation Test Rubric with Examples

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
</table>
| **Score 4** 1. Perfect repetition without any discrepancy  
2. Exact repetition: String matches stimulus exactly. Both form and meaning are correct without exception or doubt.  
3. Responses containing one word with slightly inaccurate pronunciation which is difficult to discern whether correct or incorrect should be scored as 4 (e.g., 그가 pronounced slightly like 그거; ~길 바라다 vs. ~기를 바라다; 장(위)에 있다.)  
4. Exact repetition of the pronunciation in the stimulus should be scored as 4 (e.g., 넣다 instead of 낮다). |
| **Score 3** 1. Accurate content repetition with some changes of form as long as the meaning is preserved. This includes contractions or alteration of contractions (e.g., 올핸 작년 보다 따뜻해지길 바라; 것을 instead of 걸).  
1.1. Clear mistakes in pronunciation that result in change of form should be scored as 3 (e.g., 하울 것 instead of 할거; 밀가능하다 instead of 불가능하다).  
2. Original, complete meaning is preserved as in the stimulus.  
3. Strings which are ungrammatical can get a 3 score, as long as exact meaning is preserved.  
4. Some synonymous substitutions are acceptable (e.g., 원해 instead of 바래).  
4.1. Anything with or without 너무 should be considered synonymous and should be scored as 3 (e.g., 집들이 참 좋지만 (너무) 비싸다).  
5. Changes in grammar that don’t affect meaning should be scored as 3. (Ambiguous changes in grammar that could be interpreted as meaning changes from a NS perspective should be scored as 2. That is, as a general principle in case of doubt about whether meaning has changed or not, score 2.)  
• Small meaning differences from using subject markers such as 는 or 를 instead of 가 (or vice versa) is not counted as a significant meaning difference so should be scored as 3 (e.g., 그는 고기 요리만 시키고 아채를 안먹는다).  
• Some sentences missing subject markers that don’t affect meaning should be scored as 3 (e.g., 시험은 당신이 말한 만큼(은) 어렵지 않였습니다.)  
• Or missing plural particle 들 like in the case of 아침에 아무것도 먹지 않는 사람(들)이 많습니다.  
• Missing the polite form –시 should be scored as 3 (e.g., 책상 위에 있는 책을 저에게 건네주겠어요?) |
| **Score 2** Changes in content and/or form that affect meaning.  
1. When content of string preserves at least more than half of the idea units in the original stimulus  
1.1. E.g., 그녀는 집 내부 레인트처럼 하는 거를 XXX |
Description

2. String is meaningful, and the meaning is close or related to original, but it departs from it in some slight changes in content, which makes content inexact, incomplete, or ambiguous (e.g., changes present tense to past tense)

2.1. E.g., 올해 작년보다 따뜻해지길 바래

Score 1
Repetition of half of the stimulus or less.

1. When only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out sometimes the resulting meaning is unrelated (or opposed) to stimulus.

2. When string doesn’t in itself constitute a self-standing sentence with some (related or not to stimulus) meaning (This may happen when only 2 of 3 content words are repeated and no grammatical relation between them is attempted).

- 그는 집 페인트를 내부를
- 키우던 개를 … 슬프다

Score 0
1. Nothing (Silence)
2. Laugh
3. Garbled (unintelligible speech)
4. Minimal repetition, then item abandoned:

- Only 1 word repeated
- Only 1 content word plus function word(s)
- Only function word(s) repeated
- Only 1 or 2 content words out of order plus extraneous words that weren’t in the original stimulus (e.g., 신호등 직진)
**Appendix C**

**Elicited Imitation Test Example Matrix**

<table>
<thead>
<tr>
<th>Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>• 머리로 따따를 한다.</td>
<td>• 머리로 닦아야 한다.</td>
<td>• 머리로 닦아야 한다.</td>
<td>• 머리는 깨어야 한다.</td>
<td>Perfect Repetition</td>
</tr>
<tr>
<td>Item</td>
<td>• 그 책은 XXX상위에 있다.</td>
<td>• 그 책은 상위에 있다.</td>
<td>• 그 책은 상위에 있다.</td>
<td>• 그 책은 상위에 있다.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>• 그 책은 상위에 있다.</td>
<td>• 그 책은 XXX상위에 있다.</td>
<td>• 그 책은 XXX상위에 있다.</td>
<td>• 그 책은 XXX상위에 있다.</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 이 도시에 돌아가야?</td>
<td>• 이 도시에 돌아가야?</td>
<td>• 이 도시에 돌아가야?</td>
<td>• 이 도시에 돌아가야?</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>• 뒤로 빨라 닦다.</td>
<td>• 뒤로 빨라 닦다.</td>
<td>• 뒤로 빨라 닦다.</td>
<td>• 뒤로 빨라 닦다.</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 그는 내일 사위를 해야 한다.</td>
<td>• 그들은 내일 사위를 해야 한다.</td>
<td>• 그들은 내일 사위를 해야 한다.</td>
<td>• 그들은 내일 사위를 해야 한다.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>• 오랜 허리가 움직였으나</td>
<td>• 오랜 허리가 움직였으나</td>
<td>• 오랜 허리가 움직였으나</td>
<td>• 오랜 허리가 움직였으나</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 그는 운전하는 다음에 조심해야 한다.</td>
<td>• 그는 운전하는 다음에 조심해야 한다.</td>
<td>• 그는 운전하는 다음에 조심해야 한다.</td>
<td>• 그는 운전하는 다음에 조심해야 한다.</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>• 그가 운전할 때 슬파다.</td>
<td>• 그의 운전은 하기 어렵다.</td>
<td>• 그의 운전은 하기 어렵다.</td>
<td>• 그의 운전은 하기 어렵다.</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 나는 저녁 식사가 좋아</td>
<td>• 나는 저녁 식사가 좋아</td>
<td>• 나는 저녁 식사가 좋아</td>
<td>• 나는 저녁 식사가 좋아</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>• 나는 저녁 식사가 좋아</td>
<td>• 나는 저녁 식사가 좋아</td>
<td>• 나는 저녁 식사가 좋아</td>
<td>• 나는 저녁 식사가 좋아</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 내일 비가 오며</td>
<td>• 내일 비가 오며</td>
<td>• 내일 비가 오며</td>
<td>• 내일 비가 오며</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>• 내일 비가 오며</td>
<td>• 내일 비가 오며</td>
<td>• 내일 비가 오며</td>
<td>• 내일 비가 오며</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td>• 나는 빨래를 하고</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 집안이 좀 비투한</td>
<td>• 집안이 좀 비투한</td>
<td>• 집안이 좀 비투한</td>
<td>• 집안이 좀 비투한</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>• 집안이 좀 비투한</td>
<td>• 집안이 좀 비투한</td>
<td>• 집안이 좀 비투한</td>
<td>• 집안이 좀 비투한</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td>• 귀안은 개는 XXX (개를 찾아)</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>• 그 식사는 음식을 좋아</td>
<td>• 그 식사는 음식을 좋아</td>
<td>• 그 식사는 음식을 좋아</td>
<td>• 그 식사는 음식을 좋아</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>• 그 식사는 음식을 좋아</td>
<td>• 그 식사는 음식을 좋아</td>
<td>• 그 식사는 음식을 좋아</td>
<td>• 그 식사는 음식을 좋아</td>
<td></td>
</tr>
</tbody>
</table>
Score | 0 | 1 | 2 | 3 | 4
--- | --- | --- | --- | --- | ---
**Item 13** | 대 동물들이 산...좋은 집을 원해. | | | | |
| | 대 동물들이 살...좋은 집을 원해. | 대 동물들이 살...좋은 집을 원해. | 대 동물들이 살...좋은 집을 원해. | |
| | 대 동물들이 살...좋은 집을 원해. | | | |
| | 대 동물들이 살...좋은 집을 원해. | | | |

**Item 14** | 그렇너 XXX 그렇지? | | | | |
| | 교과 음악 들어는 거 좋아하지 그렇지? | 교과 음악 들어는 거 좋아하지 그렇지? | | |
| | 교과 음악 들어는 거 좋아하지 그렇지? | 교과 음악 들어는 거 좋아하지 그렇지? | | |
| | | 교과 음악 들어는 거 좋아하지 그렇지? | | |
| | | | | |

**Item 15** | 그렇면 점은 끝했다 | | | | |
| | 그녀는 잡사매트 패인지 | 그녀는 잡사매트 패인트를 연루할 거 tục다 패인트 할 거야. | | |
| | 그녀는 잡사매트 패인지 | 그녀는 잡사매트 패인트 할 거 XXX | | |

**Item 16** | 신호등 전부 | 신호등 전부를 거두는 것이하세요. | 신호등 전부를 거두는 것이하세요. | | |
| | 신호등 전부를 거두는 것이하세요. | 신호등 전부를 거두는 것이하세요. | | |
| | 신호등 전부를 거두는 것이하세요. | | | |
| | 신호등 전부를 거두는 것이하세요. | | | |

**Item 17** | 주택이 가족 | 주택이 가족이 (차우는) 원하십시오. | 주택을 가족이 지배하는 것을 원한다. | | |
| | 주택은 가족이 지배하는 것을 원한다. | 주택을 가족이 지배하는 것을 원한다. | | |
| | 주택은 가족이 지배하는 것을 원한다. | | | |
| | 주택은 가족이 지배하는 것을 원한다. | | | |

**Item 18** | 내 집 따뜻한 바람 | 올해는 작년보다 더 따뜻해지기 바란다. | 올해는 작년보다 더 따뜻해지기 바란다. | | |
| | 올해는 작년보다 더 따뜻해지기 바란다. | 올해는 작년보다 더 따뜻해지기 바란다. | | |
| | 올해는 작년보다 더 따뜻해지기 바란다. | | | |
| | 올해는 작년보다 더 따뜻해지기 바란다. | | | |

**Item 19** | 내 천구는 편안 | 내 천구는 돌봐준다 | 내 천구는 돌봐준다 | | |
| | 내 천구는 돌봐준다 | 내 천구는 돌봐준다 | | |
| | 내 천구는 돌봐준다 | 내 천구는 돌봐준다 | | |
| | 내 천구는 돌봐준다 | 내 천구는 돌봐준다 | | |

**Item 20** | 내가 현재 없는 문을 제쳐가 좋다 | 내가 현재 없는 문을 제쳐가 좋다 | 내가 현재 없는 문을 제쳐가 좋다 | | |
| | 내가 현재 없는 문을 제쳐가 좋다 | 내가 현재 없는 문을 제쳐가 좋다 | | |
| | 내가 현재 없는 문을 제쳐가 좋다 | | | |
| | 내가 현재 없는 문을 제쳐가 좋다 | | | |

**Item 21** | XXX 고기는 아재를 맛이 밥는 | 그는 고기 식터가...아재 밥은다 | 그는 고기 식터가...아재 밥은다 | | |
| | 그는 고기 식터가...아재 밥은다 | 그는 고기 식터가...아재 밥은다 | | |
| | 그는 고기 식터가...아재 밥은다 | 그는 고기 식터가...아재 밥은다 | | |
| | 그는 고기 식터가...아재 밥은다 | 그는 고기 식터가...아재 밥은다 | | |

295
<table>
<thead>
<tr>
<th>Item</th>
<th>Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>• 내 개를 고양이랑 하테 군 응석을 내고는 개가 쫓겨 갔다.</td>
<td>• 내가 개를 고양이랑 하테 군 응석을 내고는 개가 쫓겨 갔다.</td>
<td>• 내 개를 고양이랑 하테 군 응석을 내고는 개가 쫓겨 갔다.</td>
<td>• 내가 개를 고양이랑 하테 군 응석을 내고는 개가 쫓겨 갔다.</td>
<td>• 내가 개를 고양이랑 하테 군 응석을 내고는 개가 쫓겨 갔다.</td>
<td>• 내가 개를 고양이랑 하테 군 응석을 내고는 개가 쫓겨 갔다.</td>
</tr>
<tr>
<td>23</td>
<td>• 사람 죽이지 않으니</td>
<td>• 나는 사람이 죽이지 않으니</td>
<td>• 나는 사람이 죽이지 않으니</td>
<td>• 나는 사람이 죽이지 않으니</td>
<td>• 나는 사람이 죽이지 않으니</td>
<td>• 나는 사람이 죽이지 않으니</td>
</tr>
<tr>
<td>24</td>
<td>• 내가 오리의 개를 사고 싶다.</td>
<td>• 나는 오리의 개를 사고 싶다.</td>
<td>• 나는 오리의 개를 사고 싶다.</td>
<td>• 나는 오리의 개를 사고 싶다.</td>
<td>• 나는 오리의 개를 사고 싶다.</td>
<td>• 나는 오리의 개를 사고 싶다.</td>
</tr>
<tr>
<td>25</td>
<td>• 경찰이 체포한 도둑은 키고 크고 말했다.</td>
<td>• 경찰이 체포한 도둑은 키고 크고 말했다.</td>
<td>• 경찰이 체포한 도둑은 키고 크고 말했다.</td>
<td>• 경찰이 체포한 도둑은 키고 크고 말했다.</td>
<td>• 경찰이 체포한 도둑은 키고 크고 말했다.</td>
<td>• 경찰이 체포한 도둑은 키고 크고 말했다.</td>
</tr>
<tr>
<td>26</td>
<td>• 책상 위에 있는 책은 책상 위에 있는 책을 지에게 주었어요?</td>
<td>• 책상 위에 있는 책은 책상 위에 있는 책을 지에게 주었어요?</td>
<td>• 책상 위에 있는 책은 책상 위에 있는 책을 지에게 주었어요?</td>
<td>• 책상 위에 있는 책은 책상 위에 있는 책을 지에게 주었어요?</td>
<td>• 책상 위에 있는 책은 책상 위에 있는 책을 지에게 주었어요?</td>
<td>• 책상 위에 있는 책은 책상 위에 있는 책을 지에게 주었어요?</td>
</tr>
<tr>
<td>27</td>
<td>• 담배를 피우는 사람들은,</td>
<td>• 담배를 피우는 사람들은,</td>
<td>• 담배를 피우는 사람들은,</td>
<td>• 담배를 피우는 사람들은,</td>
<td>• 담배를 피우는 사람들은,</td>
<td>• 담배를 피우는 사람들은,</td>
</tr>
<tr>
<td></td>
<td>• 담배를 피우는 사람</td>
<td>• 담배를 피우는 사람</td>
<td>• 담배를 피우는 사람</td>
<td>• 담배를 피우는 사람</td>
<td>• 담배를 피우는 사람</td>
<td>• 담배를 피우는 사람</td>
</tr>
</tbody>
</table>

296
<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>열의도는 기차가 XXX서서 모르겠다</td>
<td>열한 시간 있다 기차를 뺏나지게 몰랐다</td>
<td>열한 시 something 기차에 역에 돌아간지 모르겠다</td>
<td>열 بواس기 기차가 이미 빠났는지 모르겠다</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>시험은 당신 열... 열어도 열이 멀었지... 열어도 열이 멀었지</td>
<td>시험은 당신이 (말하지) 저랑... 나쁘지 않았습니다</td>
<td>시험은 당신을 열어도 열이 멀었지</td>
<td>열해시 당신이 이미 열어도 열이 멀었지</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|  | 아침에 아무것도 안... | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상합니다 | 아침에 아무것도 안 먹는 사람이 불상한다...
Appendix D

Background and Self-Assessment Questionnaire

Korean C-test Project

Thank you very much for participating in this study. Your responses to this survey will help us understand a little bit more about you and how you have learned Korean.

This survey is composed of four parts:
Part I - Personal background
Part II - Linguistic history and language use
Part III - Family language background
Part IV - Your perception of how good you are in Korean.

Please respond to all the questions to the best of your ability.

---

**PART I - Personal Background**

Q1 What is your name?

Q2 What is your age?

Q3 What is your gender?
   (1) Male
   (2) Female

Q4 What is your academic year?
   (3) Freshman
   (4) Sophomore
   (5) Junior
   (6) Senior
   (7) Graduate student.
   Indicate whether MA, PhD or other ____________________

Q5 What Korean course are you currently enrolled in (e.g., Intensive Korean I)?

Q6 Where were you born?
   (8) United States
   (9) Korea
   (10) Other ____________________
   If United States is selected, then skip to end of block
Q7 At what age did first you come to the United States?  

Q8 How long have you lived in the United States (e.g., 3 years, all my life)?  

Q9 Have you lived in any other country/ies outside of the United States?  
   (11) Yes  
   (12) No  

Answer: If Have you lived in any other country/ies outside of the United States? Yes is selected  

Q10 In which country/ies have you lived before coming to the United States? For how long have you stayed in that/those country/ies? (e.g., Mexico, 3 years; Korea, 6 months; Canada, 1 year).  

Q11 Do you consider yourself a heritage language learner of Korean? Korean heritage language learners can be defined as learners whose home language is Korean, who were raised in a home where Korean is spoken, and who have Korean ancestral or familiar ties.  
   (13) Yes  
   (14) No  
   (15) I don't understand what heritage language learner is  

PART II - Linguistic history and language use  

Q12 Please list all languages you know in order of dominance (the ones that you feel most comfortable with first).  
   1  
   2  
   3  
   4  
   5
Q13 Please list all the languages you know in the order of acquisition (in the order you learned them).
   1
   2
   3
   4
   5

Q14 Complete the following sentence: English is my ________________ language.
   (16) Native language
   (17) Second language
   (18) Third language
   (19) Fourth language

Q15 At what age did you first...
   begin acquiring English __________
   begin reading in English __________

Q16 Complete the following sentence: Korean is my ________________ language.
   (20) Native language
   (21) Second language
   (22) Third language
   (23) Fourth language

Q17 At what age did you first...
   begin acquiring Korean? __________
   begin reading in Korean? __________

Q18 Have you studied Korean language in the US?
   (24) Yes
   (25) No

If No Is Selected, Then Skip To Have you attended a school in Korea?
Q19 Where did you study Korean (in the US)?
1. Elementary school
2. Middle school
3. High School
4. Weekend school
5. University courses
6. Korean Language Institutes
7. Private tutoring
8. Church program
9. Other ____________________

Q20 How long have you studied Korean (in the US)? Please state your answer in years and months (e.g., Elementary school, 2 years; University courses, 1 year; Private tutoring, N/A; etc).

Elementary school
Middle school
High School
Weekend school
University courses
Korean Language Institutes
Private tutoring
Other

Q21 Have you attended a school or studied in Korea?
(26) Yes
(27) No

If No Is Selected, Then Skip To In general, which language do you fee...

Q22 When did you attend school or studied in Korea?
10. Elementary School (1st - 6th grade)
11. Middle School (7th - 9th grade)
12. High-School (10th - 12th grade)
13. University (Undergraduate)
14. University (Graduate)
15. Study Abroad
16. Other ____________________

Q23 How long did you attend school or studied in Korea (e.g., elementary 2 years; study abroad 2 months)?
Q24 In general, which language do you feel most comfortable with when:

<table>
<thead>
<tr>
<th>Activity</th>
<th>English (1)</th>
<th>Korean (2)</th>
<th>Both (3)</th>
<th>N/A (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watching TV programs</td>
<td>(28)</td>
<td>(29)</td>
<td>(30)</td>
<td>(31)</td>
</tr>
<tr>
<td>Watching the news</td>
<td>(32)</td>
<td>(33)</td>
<td>(34)</td>
<td>(35)</td>
</tr>
<tr>
<td>Watching movies</td>
<td>(36)</td>
<td>(37)</td>
<td>(38)</td>
<td>(39)</td>
</tr>
<tr>
<td>Listening to the radio</td>
<td>(40)</td>
<td>(41)</td>
<td>(42)</td>
<td>(43)</td>
</tr>
<tr>
<td>Reading novels</td>
<td>(44)</td>
<td>(45)</td>
<td>(46)</td>
<td>(47)</td>
</tr>
<tr>
<td>Reading the news</td>
<td>(48)</td>
<td>(49)</td>
<td>(50)</td>
<td>(51)</td>
</tr>
<tr>
<td>Reading magazines</td>
<td>(52)</td>
<td>(53)</td>
<td>(54)</td>
<td>(55)</td>
</tr>
<tr>
<td>Taking notes</td>
<td>(56)</td>
<td>(57)</td>
<td>(58)</td>
<td>(59)</td>
</tr>
<tr>
<td>Surfing the Web</td>
<td>(60)</td>
<td>(61)</td>
<td>(62)</td>
<td>(63)</td>
</tr>
<tr>
<td>Using Facebook</td>
<td>(64)</td>
<td>(65)</td>
<td>(66)</td>
<td>(67)</td>
</tr>
<tr>
<td>Using Twitter</td>
<td>(68)</td>
<td>(69)</td>
<td>(70)</td>
<td>(71)</td>
</tr>
<tr>
<td>Using Snapchat</td>
<td>(72)</td>
<td>(73)</td>
<td>(74)</td>
<td>(75)</td>
</tr>
<tr>
<td>Writing text messages</td>
<td>(76)</td>
<td>(77)</td>
<td>(78)</td>
<td>(79)</td>
</tr>
<tr>
<td>Talking on the phone</td>
<td>(80)</td>
<td>(81)</td>
<td>(82)</td>
<td>(83)</td>
</tr>
</tbody>
</table>
Q25 How often do you use Korean when you:

<table>
<thead>
<tr>
<th>Activity</th>
<th>Never</th>
<th>2</th>
<th>3</th>
<th>A great deal</th>
<th>N/A (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talk to your parents</td>
<td>(84)</td>
<td>(85)</td>
<td>(86)</td>
<td>(87)</td>
<td>(88)</td>
</tr>
<tr>
<td>Talk to your siblings</td>
<td>(89)</td>
<td>(90)</td>
<td>(91)</td>
<td>(92)</td>
<td>(93)</td>
</tr>
<tr>
<td>Talk to your friends</td>
<td>(94)</td>
<td>(95)</td>
<td>(96)</td>
<td>(97)</td>
<td>(98)</td>
</tr>
<tr>
<td>Talk to your spouse</td>
<td>(99)</td>
<td>(100)</td>
<td>(101)</td>
<td>(102)</td>
<td>(103)</td>
</tr>
<tr>
<td>Talk to your relatives</td>
<td>(104)</td>
<td>(105)</td>
<td>(106)</td>
<td>(107)</td>
<td>(108)</td>
</tr>
<tr>
<td>Talk to Korean adults</td>
<td>(109)</td>
<td>(110)</td>
<td>(111)</td>
<td>(112)</td>
<td>(113)</td>
</tr>
<tr>
<td>Surf the Web</td>
<td>(114)</td>
<td>(115)</td>
<td>(116)</td>
<td>(117)</td>
<td>(118)</td>
</tr>
<tr>
<td>Use Facebook</td>
<td>(119)</td>
<td>(120)</td>
<td>(121)</td>
<td>(122)</td>
<td>(123)</td>
</tr>
<tr>
<td>Use Twitter</td>
<td>(124)</td>
<td>(125)</td>
<td>(126)</td>
<td>(127)</td>
<td>(128)</td>
</tr>
<tr>
<td>Use Snapchat</td>
<td>(129)</td>
<td>(130)</td>
<td>(131)</td>
<td>(132)</td>
<td>(133)</td>
</tr>
<tr>
<td>Write text messages</td>
<td>(134)</td>
<td>(135)</td>
<td>(136)</td>
<td>(137)</td>
<td>(138)</td>
</tr>
<tr>
<td>Talk on the phone</td>
<td>(139)</td>
<td>(140)</td>
<td>(141)</td>
<td>(142)</td>
<td>(143)</td>
</tr>
<tr>
<td>Write emails</td>
<td>(144)</td>
<td>(145)</td>
<td>(146)</td>
<td>(147)</td>
<td>(148)</td>
</tr>
<tr>
<td>Take notes in class</td>
<td>(149)</td>
<td>(150)</td>
<td>(151)</td>
<td>(152)</td>
<td>(153)</td>
</tr>
<tr>
<td>Talk informally (e.g., small talk)</td>
<td>(154)</td>
<td>(155)</td>
<td>(156)</td>
<td>(157)</td>
<td>(158)</td>
</tr>
<tr>
<td>Talk formally (e.g., presentation)</td>
<td>(159)</td>
<td>(160)</td>
<td>(161)</td>
<td>(162)</td>
<td>(163)</td>
</tr>
</tbody>
</table>

Q26 How often do you watch or listen to Korean:

<table>
<thead>
<tr>
<th>Source</th>
<th>Never</th>
<th>2</th>
<th>3</th>
<th>A great deal</th>
<th>N/A (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV shows</td>
<td>(164)</td>
<td>(165)</td>
<td>(166)</td>
<td>(167)</td>
<td>(168)</td>
</tr>
<tr>
<td>Dramas</td>
<td>(169)</td>
<td>(170)</td>
<td>(171)</td>
<td>(172)</td>
<td>(173)</td>
</tr>
<tr>
<td>Movies</td>
<td>(174)</td>
<td>(175)</td>
<td>(176)</td>
<td>(177)</td>
<td>(178)</td>
</tr>
<tr>
<td>News</td>
<td>(179)</td>
<td>(180)</td>
<td>(181)</td>
<td>(182)</td>
<td>(183)</td>
</tr>
<tr>
<td>Music</td>
<td>(184)</td>
<td>(185)</td>
<td>(186)</td>
<td>(187)</td>
<td>(188)</td>
</tr>
</tbody>
</table>
Q27 How often do you read Korean:

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>2</th>
<th>3</th>
<th>A great deal</th>
<th>N/A (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspapers (1)</td>
<td>(189)</td>
<td>(190)</td>
<td>(191)</td>
<td>(192)</td>
<td>(193)</td>
</tr>
<tr>
<td>Internet news (2)</td>
<td>(194)</td>
<td>(195)</td>
<td>(196)</td>
<td>(197)</td>
<td>(198)</td>
</tr>
<tr>
<td>Novels (3)</td>
<td>(199)</td>
<td>(200)</td>
<td>(201)</td>
<td>(202)</td>
<td>(203)</td>
</tr>
<tr>
<td>Magazines (4)</td>
<td>(204)</td>
<td>(205)</td>
<td>(206)</td>
<td>(207)</td>
<td>(208)</td>
</tr>
<tr>
<td>Emails (5)</td>
<td>(209)</td>
<td>(210)</td>
<td>(211)</td>
<td>(212)</td>
<td>(213)</td>
</tr>
<tr>
<td>Internet sites (6)</td>
<td>(214)</td>
<td>(215)</td>
<td>(216)</td>
<td>(217)</td>
<td>(218)</td>
</tr>
</tbody>
</table>

PART III - Family History

Q28 Do your parents speak Korean with you?
(219) Yes (1)
(220) No (2)

Answer If Do your parents speak Korean with you? Yes Is Selected

Q29 Who speaks Korean with you?
(221) Mother (1)
(222) Father (2)
(223) Both (3)

Q30 In what language(s) do you speak to your parents?
(224) English (1)
(225) Korean (2)
(226) Mixed (Korean and English) (3)
(227) Other (4) ____________________

Q31 Do you have siblings?
(228) Yes (1)
(229) No (2)

If No Is Selected, Then Skip To Do you have a Korean spouse?

Q32 What language(s) does/do you use when speaking to your siblings?
(230) Korean (1)
(231) English (2)
(232) Mixed (3)
(233) Other (4) ____________________
Q33 What language(s) does/do your sibling/s use to speak with you?
   (234) Korean (1)
   (235) English (2)
   (236) Mixed (3)
   (237) Other (4) ____________________

Q34 Do you have a Korean spouse, partner, girl/boyfriend?
   (238) Yes (1)
   (239) No (2)

If No is Selected, Then Skip To Do you speak Korean with other family...

Q35 What language(s) do you use to speak to your spouse, partner, girl/boyfriend?
   (240) English (1)
   (241) Korean (2)
   (242) Mixed (3)
   (243) Other (4) ____________________

Q36 What languages does your spouse, partner, girl/boyfriend use to speak with you?
   (244) English (1)
   (245) Korean (2)
   (246) Mixed (3)
   (247) Other (4) ____________________

Q37 Do you speak Korean with other family members or relatives?
   (248) Yes (1)
   (249) No (2)
Answer If Do you speak Korean with other family members or relatives? Yes Is Selected

Q38 Which family members or relatives do you speak Korean to? How often do you meet with them?

<table>
<thead>
<tr>
<th></th>
<th>Less than once a month (1)</th>
<th>Once a Month (2)</th>
<th>2-3 Times a Month (3)</th>
<th>Once a Week (4)</th>
<th>2-3 Times a Week (5)</th>
<th>Daily (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>(250)</td>
<td>(251)</td>
<td>(252)</td>
<td>(253)</td>
<td>(254)</td>
<td>(255)</td>
</tr>
<tr>
<td>2.</td>
<td>(256)</td>
<td>(257)</td>
<td>(258)</td>
<td>(259)</td>
<td>(260)</td>
<td>(261)</td>
</tr>
<tr>
<td>3.</td>
<td>(262)</td>
<td>(263)</td>
<td>(264)</td>
<td>(265)</td>
<td>(266)</td>
<td>(267)</td>
</tr>
<tr>
<td>4.</td>
<td>(268)</td>
<td>(269)</td>
<td>(270)</td>
<td>(271)</td>
<td>(272)</td>
<td>(273)</td>
</tr>
<tr>
<td>5.</td>
<td>(274)</td>
<td>(275)</td>
<td>(276)</td>
<td>(277)</td>
<td>(278)</td>
<td>(279)</td>
</tr>
<tr>
<td>7.</td>
<td>(286)</td>
<td>(287)</td>
<td>(288)</td>
<td>(289)</td>
<td>(290)</td>
<td>(291)</td>
</tr>
<tr>
<td>8.</td>
<td>(292)</td>
<td>(293)</td>
<td>(294)</td>
<td>(295)</td>
<td>(296)</td>
<td>(297)</td>
</tr>
<tr>
<td>9.</td>
<td>(298)</td>
<td>(299)</td>
<td>(300)</td>
<td>(301)</td>
<td>(302)</td>
<td>(303)</td>
</tr>
<tr>
<td>10.</td>
<td>(304)</td>
<td>(305)</td>
<td>(306)</td>
<td>(307)</td>
<td>(308)</td>
<td>(309)</td>
</tr>
</tbody>
</table>

Q39 Do you have Korean friends?
   (310) Yes
   (311) No

If No Is Selected, Then Skip To End of Block

Q40 Do you speak Korean with them?
   (312) Yes
   (313) No

Q41 How often do you speak Korean with your friends?
   (314) Daily
   (315) 2-3 times a week
   (316) Once a week
   (317) 2-3 times a month
   (318) Once a month
   (319) Less than once a month
   (320) Other ____________________

PART IV - Self-assessment of Korean language proficiency
Q42 How would you rate your proficiency in Korean?

<table>
<thead>
<tr>
<th></th>
<th>Very poor (1)</th>
<th>Needs work (2)</th>
<th>Good (3)</th>
<th>Native-like (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>(321)</td>
<td>(322)</td>
<td>(323)</td>
<td>(324)</td>
</tr>
<tr>
<td>Listening</td>
<td>(325)</td>
<td>(326)</td>
<td>(327)</td>
<td>(328)</td>
</tr>
<tr>
<td>Speaking</td>
<td>(329)</td>
<td>(330)</td>
<td>(331)</td>
<td>(332)</td>
</tr>
<tr>
<td>Writing</td>
<td>(333)</td>
<td>(334)</td>
<td>(335)</td>
<td>(336)</td>
</tr>
</tbody>
</table>

Q43 How would you rate your proficiency in English?

<table>
<thead>
<tr>
<th></th>
<th>Very poor (1)</th>
<th>Needs work (2)</th>
<th>Good (3)</th>
<th>Native-like (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>(337)</td>
<td>(338)</td>
<td>(339)</td>
<td>(340)</td>
</tr>
<tr>
<td>Listening</td>
<td>(341)</td>
<td>(342)</td>
<td>(343)</td>
<td>(344)</td>
</tr>
<tr>
<td>Speaking</td>
<td>(345)</td>
<td>(346)</td>
<td>(347)</td>
<td>(348)</td>
</tr>
<tr>
<td>Writing</td>
<td>(349)</td>
<td>(350)</td>
<td>(351)</td>
<td>(352)</td>
</tr>
</tbody>
</table>
Q44 Please indicate how well you can do the following tasks in Korean based on the scale provided.

<table>
<thead>
<tr>
<th>Task</th>
<th>Quite easily (1)</th>
<th>With some difficulty (2)</th>
<th>With great difficulty / Not at all (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can introduce myself</td>
<td>(353)</td>
<td>(354)</td>
<td>(355)</td>
</tr>
<tr>
<td>I can talk about familiar topics with basic vocabulary</td>
<td>(356)</td>
<td>(357)</td>
<td>(358)</td>
</tr>
<tr>
<td>I can describe my family and friends</td>
<td>(359)</td>
<td>(360)</td>
<td>(361)</td>
</tr>
<tr>
<td>I can understand Korean language at a sentence-level</td>
<td>(362)</td>
<td>(363)</td>
<td>(364)</td>
</tr>
<tr>
<td>I can understand Korean language at a paragraph-level</td>
<td>(365)</td>
<td>(366)</td>
<td>(367)</td>
</tr>
<tr>
<td>I can understand informal phone call conversations</td>
<td>(368)</td>
<td>(369)</td>
<td>(370)</td>
</tr>
<tr>
<td>I can read and understand informal emails</td>
<td>(371)</td>
<td>(372)</td>
<td>(373)</td>
</tr>
<tr>
<td>I can read and understand basic narratives or stories</td>
<td>(374)</td>
<td>(375)</td>
<td>(376)</td>
</tr>
<tr>
<td>I can read and understand a journal entry</td>
<td>(377)</td>
<td>(378)</td>
<td>(379)</td>
</tr>
<tr>
<td>I can understand advertisement</td>
<td>(380)</td>
<td>(381)</td>
<td>(382)</td>
</tr>
<tr>
<td>I can address Korean adults with the appropriate respect forms</td>
<td>(383)</td>
<td>(384)</td>
<td>(385)</td>
</tr>
<tr>
<td>I can talk about unfamiliar topics</td>
<td>(386)</td>
<td>(387)</td>
<td>(388)</td>
</tr>
<tr>
<td>I can read and understand public notices</td>
<td>(389)</td>
<td>(390)</td>
<td>(391)</td>
</tr>
<tr>
<td>I can read novels</td>
<td>(392)</td>
<td>(393)</td>
<td>(394)</td>
</tr>
<tr>
<td>I can read articles about culture and society</td>
<td>(395)</td>
<td>(396)</td>
<td>(397)</td>
</tr>
<tr>
<td>I can read newspaper articles</td>
<td>(398)</td>
<td>(399)</td>
<td>(400)</td>
</tr>
<tr>
<td>I can understand news broadcast</td>
<td>(401)</td>
<td>(402)</td>
<td>(403)</td>
</tr>
<tr>
<td>I can read and understand research reports in Korean</td>
<td>(404)</td>
<td>(405)</td>
<td>(406)</td>
</tr>
</tbody>
</table>
Q45 Have you taken a Korean language proficiency test, like TOPIK, OPI, OPIc, WPT, SAT II?
   (407) Yes (1)
   (408) No (2)

Answer If Have you taken a Korean language proficiency test? (E.g., TOPIK) Yes Is Selected
Q46
   What was/were the test(s)? (1)
   What was your scores(s)? (2)

Thank you very much for your responses! 감사합니다!
Appendix E

Item Characteristic Curves for Seven Items
Appendix F

Seven-Item C-test Item Map

MEASURE Student - MAP - Text

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>T+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XX</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>XXXX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>XX</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XX</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>T</td>
<td>Text 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<less><freq>
Appendix G

Cluster Analysis Dendrogram

*Cluster labelling is based on the four-cluster solution
Appendix H

Descriptive Statistics for 10 Variables across Four Clusters

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Std. Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>28</td>
<td>1.77</td>
<td>0.49</td>
<td>1.58</td>
<td>1.96</td>
<td>0.80</td>
<td>2.80</td>
</tr>
<tr>
<td><strong>OPIc</strong></td>
<td>2</td>
<td>29</td>
<td>1.53</td>
<td>0.48</td>
<td>1.35</td>
<td>1.72</td>
<td>0.10</td>
<td>2.30</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>0.84</td>
<td>0.34</td>
<td>0.71</td>
<td>0.97</td>
<td>0.10</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>2.43</td>
<td>0.29</td>
<td>2.12</td>
<td>2.74</td>
<td>2.10</td>
<td>2.80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>1.45</td>
<td>0.63</td>
<td>1.32</td>
<td>1.58</td>
<td>0.10</td>
<td>2.80</td>
</tr>
<tr>
<td><strong>WPT</strong></td>
<td>1</td>
<td>28</td>
<td>1.56</td>
<td>0.45</td>
<td>1.38</td>
<td>1.73</td>
<td>0.80</td>
<td>2.30</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29</td>
<td>1.69</td>
<td>0.44</td>
<td>1.52</td>
<td>1.86</td>
<td>1.10</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>1.14</td>
<td>0.20</td>
<td>1.07</td>
<td>1.22</td>
<td>0.80</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>2.53</td>
<td>0.36</td>
<td>2.15</td>
<td>2.91</td>
<td>2.30</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>1.53</td>
<td>0.51</td>
<td>1.43</td>
<td>1.64</td>
<td>0.80</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Length of Stay</strong></td>
<td>1</td>
<td>28</td>
<td>2.45</td>
<td>5.29</td>
<td>4.40</td>
<td>4.50</td>
<td>0.00</td>
<td>17.40</td>
</tr>
<tr>
<td><strong>Studying in Korea</strong></td>
<td>2</td>
<td>29</td>
<td>21.65</td>
<td>28.81</td>
<td>10.69</td>
<td>32.61</td>
<td>0.00</td>
<td>117.33</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>2.41</td>
<td>5.85</td>
<td>1.09</td>
<td>4.63</td>
<td>0.00</td>
<td>26.10</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>243.33</td>
<td>120.79</td>
<td>49.31</td>
<td>116.57</td>
<td>370.09</td>
<td>104.28</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>24.20</td>
<td>67.39</td>
<td>7.03</td>
<td>10.24</td>
<td>38.15</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Parents</strong></td>
<td>1</td>
<td>28</td>
<td>2.18</td>
<td>0.83</td>
<td>1.86</td>
<td>2.50</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>0.24</td>
<td>0.52</td>
<td>0.04</td>
<td>.44</td>
<td>0.00</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>2.63</td>
<td>0.59</td>
<td>2.01</td>
<td>3.24</td>
<td>1.50</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>0.91</td>
<td>1.18</td>
<td>0.67</td>
<td>1.16</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Siblings</strong></td>
<td>1</td>
<td>28</td>
<td>0.71</td>
<td>0.60</td>
<td>0.48</td>
<td>.95</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29</td>
<td>0.03</td>
<td>0.19</td>
<td>-0.04</td>
<td>.11</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>0.14</td>
<td>0.44</td>
<td>0.08</td>
<td>.31</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>2.17</td>
<td>0.98</td>
<td>1.13</td>
<td>3.20</td>
<td>1.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>0.41</td>
<td>0.73</td>
<td>0.26</td>
<td>.56</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Relatives</strong></td>
<td>1</td>
<td>28</td>
<td>2.36</td>
<td>1.03</td>
<td>1.96</td>
<td>2.76</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29</td>
<td>0.03</td>
<td>0.186</td>
<td>-0.04</td>
<td>.11</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>0.14</td>
<td>0.35</td>
<td>0.00</td>
<td>0.27</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>2.50</td>
<td>1.23</td>
<td>1.21</td>
<td>3.79</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>1.41</td>
<td>1.23</td>
<td>0.93</td>
<td>1.89</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Grandparents</strong></td>
<td>1</td>
<td>28</td>
<td>0.30</td>
<td>0.35</td>
<td>0.26</td>
<td>0.81</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29</td>
<td>0.03</td>
<td>0.19</td>
<td>-0.04</td>
<td>0.11</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>0.14</td>
<td>0.35</td>
<td>0.00</td>
<td>0.27</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>1.75</td>
<td>1.13</td>
<td>0.57</td>
<td>2.94</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>0.60</td>
<td>1.01</td>
<td>0.39</td>
<td>0.81</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td><strong>Computer Use</strong></td>
<td>1</td>
<td>28</td>
<td>0.76</td>
<td>0.54</td>
<td>0.55</td>
<td>0.97</td>
<td>0.14</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29</td>
<td>1.17</td>
<td>0.61</td>
<td>0.94</td>
<td>1.41</td>
<td>0.29</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29</td>
<td>0.29</td>
<td>0.25</td>
<td>0.19</td>
<td>0.38</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>1.36</td>
<td>0.33</td>
<td>1.02</td>
<td>1.71</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>92</td>
<td>0.78</td>
<td>0.61</td>
<td>0.65</td>
<td>0.91</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>Std. Error</td>
<td>95% Confidence Interval for Mean</td>
<td>Minimum</td>
<td>Maximum</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----</td>
<td>------</td>
<td>-----</td>
<td>------------</td>
<td>---------------------------------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Upper Bound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watch &amp; Listen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>1.17</td>
<td>0.77</td>
<td>0.15</td>
<td>0.87</td>
<td>1.47</td>
<td>0.00</td>
<td>2.80</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>1.73</td>
<td>0.61</td>
<td>0.11</td>
<td>1.49</td>
<td>1.96</td>
<td>0.80</td>
<td>3.00</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>1.42</td>
<td>0.70</td>
<td>0.13</td>
<td>1.15</td>
<td>1.69</td>
<td>0.00</td>
<td>2.60</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2.03</td>
<td>0.56</td>
<td>0.23</td>
<td>1.45</td>
<td>2.62</td>
<td>1.60</td>
<td>3.00</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>1.48</td>
<td>0.73</td>
<td>0.08</td>
<td>1.33</td>
<td>1.63</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Read</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>0.46</td>
<td>0.57</td>
<td>0.11</td>
<td>0.24</td>
<td>0.68</td>
<td>0.00</td>
<td>1.75</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>0.85</td>
<td>0.48</td>
<td>0.09</td>
<td>0.67</td>
<td>1.03</td>
<td>0.00</td>
<td>2.17</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>0.18</td>
<td>0.21</td>
<td>0.04</td>
<td>0.11</td>
<td>0.26</td>
<td>0.00</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>1.14</td>
<td>0.41</td>
<td>0.17</td>
<td>0.70</td>
<td>1.57</td>
<td>0.50</td>
<td>1.67</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>0.54</td>
<td>0.54</td>
<td>0.06</td>
<td>0.43</td>
<td>0.65</td>
<td>0.00</td>
<td>2.17</td>
</tr>
</tbody>
</table>
Appendix I

Numeric Values for the ACTFL Proficiency Levels (from Vande Berg, Connor-Linton, & Paige, 2009)

<table>
<thead>
<tr>
<th>ACTFL Proficiency Levels</th>
<th>Numerical Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superior</td>
<td>3.0</td>
</tr>
<tr>
<td>Advanced High</td>
<td>2.8</td>
</tr>
<tr>
<td>Advanced Mid</td>
<td>2.3</td>
</tr>
<tr>
<td>Advanced Low</td>
<td>2.1</td>
</tr>
<tr>
<td>Intermediate High</td>
<td>1.8</td>
</tr>
<tr>
<td>Intermediate Mid</td>
<td>1.3</td>
</tr>
<tr>
<td>Intermediate Low</td>
<td>1.1</td>
</tr>
<tr>
<td>Novice High</td>
<td>0.8</td>
</tr>
<tr>
<td>Novice Mid</td>
<td>0.3</td>
</tr>
<tr>
<td>Novice Low</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Appendix J

Boxplots of Korean C-test, ACTFL Oral Proficiency Interview-computer, ACTFL Writing Proficiency Test, and Elicited Imitation Test
Appendix K

Residual Scatterplots

Residual Scatterplot for both HLLs and FLLs

Residual Scatterplot for HLLs

Residual Scatterplot for FLLs
Appendix L

Bivariate Correlation Scatterplots

ACTFL OPIc and WPT

C-test and ACTFL WPT

C-test and ACTFL OPIc

ACTFL WPT and ACTFL Tests Mean Scores

ACTFL OPIc and ACTFL Tests Mean Scores

ACTFL OPIc and EIT

ACTFL WPT and EIT

C-test and EIT

EIT and ACTFL Tests Mean Scores
References


AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.


Fremdsprache’ (TestDaF): Erste Forschungsergebnisse. [C-tests within the ‘Test of German as a foreign language’ (TestDaF): preliminary research findings]. In R. Grotjahn (Ed.), Der C-test: Theoretische Grundlagen und praktische Anwendungen [The C-test: Theoretical foundations and practical applications] (pp. 175–209). Bochum: AKS-Verlag.


Baghaei, P. (2011a). Do C-tests with different number of gaps measure the same construct? *Theory and Practice in Language Studies, 1*(6), 688-693.


Univeristy.


Draper, J., & Hicks, J. (2000). Where we’ve been; what we’ve learned. In J. Webb, & B. Miller, Teaching Heritage Language Learners: Voices from the Classroom (pp. 15-35). Yonkers, New York: ACTFL.


Harsch, C., & Hartig, J. (2015). Comparing C-tests and Yes/No vocabulary size tests as


Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the con-struct validation of


Martin, C. (2010). Assessing the oral proficiency of adult learners, "heritage" and


Montrul, S. (2002). Incomplete acquisition and attrition if Spanish tense/aspect


http://www.nhlrc.ucla.edu/nhlrc


Polinsky, M., & Kagan, O. (2007). Heritage languages: In the "wild" and in the
classroom. *Language and Linguistics Compass*, 1, 368-395.


Thompson, G. L. (2015). Understanding the heritage language student: Proficiency and


