COMPARING NATIVE AND NON-NATIVE Raters of US FEDERAL GOVERNMENT SPEAKING TESTS

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

By

Rachel Lunde Brooks, M.A.T.

Washington, DC
August 7, 2013
COMPARING NATIVE AND NON-NATIVE RATERS OF US FEDERAL GOVERNMENT SPEAKING TESTS

Rachel Lunde Brooks, M.A.T.

Thesis Advisor: Jeffrey Connor-Linton, Ph.D.

ABSTRACT

Previous Language Testing research has largely reported that although many raters’ characteristics affect their evaluations of language assessments (Reed & Cohen, 2001), being a native speaker or non-native speaker rater does not significantly affect final ratings (Kim, 2009b). In Second Language Acquisition, some researchers conclude that performance and perception differences exist between native and non-native speakers, while others contend that there is little conclusive evidence to support end state differences. The US Federal Government requires speaking test raters to be both native and high-proficiency speakers of the test language (FBI, 2009). An exploration of how the native speaker is defined in research reveals a lack of common understanding, referring both to an ideal speaker and a native acquirer of language.

This study built on previous research by expanding the breadth of proficiency levels rated to include highly articulate examinees, regrouping the raters to represent three ideas of nativeness (native/non-native speakers, speaking proficiency, and first language), and examining final and linguistic category ratings to reveal the raters’ scoring construct. Thirty FBI speaking testers, native and non-native speakers of English, rated 25 English Speaking Proficiency Tests. They assigned ratings for the overall test and linguistic categories, including functions, organization, structures, vocabulary, fluency, pronunciation and social/cultural appropriateness. Using ANOVAs and MANOVAs, the results indicated no significant difference between the native and non-native speaker groups. When raters were grouped by English proficiency level,
lower proficiency raters gave significantly lower ratings, both in the final and in many linguistic
category ratings, although with a small effect size.

Non-native speakers rated comparably to native speakers in the results, but significant
differences occurred between rater groups when they were arranged by rater English speaking
proficiency and first language. The results suggested that rater training organizations should
consider rater proficiency level rather than whether or not they are native speakers. Additionally,
they supported the theory that non-native speakers can demonstrate language acquisition
equivalent to native speakers, at least when evaluating language. Finally, it was recommended
that researchers and testing practitioners that use native speakers should clearly define and justify
their use, or avoid the native speaker term altogether.
DEDICATION

To Dave Brooks, my favorite.
ACKNOWLEDGEMENTS

There are many people who contributed to the completion of this dissertation. To each of them, I am extremely grateful. I have come to understand that the journey of a dissertation is one that is a team effort and I very thankful to those who made this effort possible. This dissertation would not have been possible without the assistance of my family, mentors, colleagues and friends. The credit for this degree is shared, but the faults in it are all mine.

First and foremost, I must thank my heavenly father, for whom I am completing this degree. I pray that any accolade that I receive be used for your service and to further your will. I certainly could not have completed it without the strength, endurance, and intelligence you have given me. I give you all the glory for this degree.

I give many thanks to my wonderful, thoughtful and enduring husband, the Honorable David Steven Brooks. Your encouragement, love, and support have gotten me through these seven years, and for that I could not thank you enough. You have done more than your share of housework and babysitting to give me the time to complete this goal. Moreover, you have helped to code data, transcribe, and revise my writing. For almost all the time we have been married, I have been pursuing this degree, so I look forward to devoting my new found free time to adventures with you.

I am grateful to Jeff Connor-Linton, my advisor, Elana Shohamy, and Pardee Lowe, Jr. for helping me to see this project through. I appreciate the long hours of consultations and feedback that you have offered. Any balance that this study shows is reflective of the diverse perspectives offered by the committee members and their commitment to a quality product. I would also like to thank Alison Mackey, who gave me sound advice through most of the process, but was not able to serve as a final committee member.

Maria Brau, Unit Chief of the Language Testing and Assessment Unit at the FBI was also instrumental in seeing this dissertation completed. I am grateful for her support and flexibility for seven years during the course of my studies. She helped me secure the participants and the exams they rated, as well as supported my studies while I was working. Most of all, she has been a trusted friend and mentor both academically and personally and I have learned so much from her both in my career and in life.

I greatly appreciate my colleagues who served as participants. It has been a pleasure working with the speaking testers program for nearly a decade, as there has hardly ever been such a group of diverse, intelligent, articulate, and caring people assembled. Thank you to my colleagues that helped organize data as well as give me feedback on the project.

I could not give enough thanks to my parents, Ronald and Jennifer Lunde, who have loved and supported me all my life. My accomplishments are your accomplishments. Thanks also to my family members and friends, who have offered countless prayers and words of encouragement on my behalf. Thank you to my siblings, my mentor, Karen, and the members of my bible study. In particular, I am grateful to women of my prayer circle, 10-40. Deb, Erin, Gayleen, Jen, Lanier, Lindsay, Melissa, Selina and Stephanie, your support has blessed me beyond measure and I pray that you are blessed for your dedication.
# Table of Contents

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

1.1 Overview .................................................................................................................................. 1

1.2 The ubiquitous and ambiguous native speaker ....................................................................... 3
  1.2.1 Comparing native speakers ................................................................................................... 4
  1.2.2 Native speakers in as language arbiters ................................................................................. 5
  1.2.3 Contributions of the current study ......................................................................................... 8

1.3 Research questions ..................................................................................................................... 9

1.4 Research summary ...................................................................................................................... 13

1.5 Implications of research ............................................................................................................ 15

Chapter 2: Literature Review ......................................................................................................... 18

2.1 Overview ................................................................................................................................... 18

2.2 Native speaker perspectives ...................................................................................................... 20
  2.2.1 History of the ‘native speaker’ ............................................................................................... 20
  2.2.2 Native speaker definition reconsidered ................................................................................... 22
  2.2.3 The native speaker in Second Language Acquisition ............................................................ 29
  2.2.4 The native speaker in Language Testing ................................................................................. 36
  2.2.5 The native speaker in the United States Federal Government ............................................. 40

2.3 Social and political implications of the native speaker concept ................................................. 45
  2.3.1 Identity and language test scores ......................................................................................... 47
  2.3.2 Which English? ....................................................................................................................... 49
  2.3.3 Correctness .............................................................................................................................. 51
  2.3.4 The power of a score .............................................................................................................. 55
Chapter 5: Results ................................................................................................................................. 110

5.1 Overview ........................................................................................................................................ 110

5.2 Normality of the data ...................................................................................................................... 111

5.3 Exam level redistribution .............................................................................................................. 114

5.4 Native and non-native speaker rater comparison: Final ratings .................................................. 117
    5.4.1 Comparison of NS and NNS raters’ final ratings: Introduction ........................................... 117
    5.4.2 Comparison of NS and NNS raters’ final ratings: Inter-rater reliability .............................. 117
    5.4.3 Comparison of NS and NNS raters’ final ratings: All exams .............................................. 119
    5.4.4 Comparison of NS and NNS raters’ final ratings: Analysis of exams by level of proficiency
    ...................................................................................................................................................... 121

5.5 Rater English proficiency comparison: Final ratings ................................................................ 128
    5.5.1 Comparison of raters’ final ratings by English proficiency: Introduction .......................... 128
    5.5.2 Comparison of raters’ final ratings by English proficiency: Inter-rater reliability ............. 129
    5.5.3 Comparison of raters’ final ratings by English proficiency: All exams ............................. 130
    5.5.4 Comparison of raters’ final ratings by English proficiency: Analysis of exams by level of proficiency .............................................................................................................................................. 133

5.6 Rater native language comparison: Final ratings ......................................................................... 139
    5.6.1 Comparison of raters’ final ratings by native language: Introduction ............................... 139
    5.6.2 Comparison of raters’ final ratings by native language: Inter-rater reliability .................. 139
    5.6.3 Comparison of raters’ final ratings by native language: All exams .................................... 140
    5.6.4 Comparison of raters’ final ratings by native language: Analysis of exams by level ........ 144

5.7 Native and non-native speaker rater comparison: Linguistic category ratings .......................... 149
    5.7.1 Comparison of NS and NNS raters’ category ratings: Introduction .................................. 149
    5.7.2 Comparison of NS and NNS raters’ category ratings: All exams ....................................... 150
5.7.3 Comparison of NS and NNS raters’ category ratings: Analysis of exams by level of proficiency ................................................................. 153

5.8 Rater English proficiency comparison: Linguistic category ratings ................................................................. 163

5.8.1 Comparison of raters’ category ratings by English proficiency: Introduction ................. 163

5.8.2 Comparison of raters’ category ratings by English proficiency: All exams................. 164

5.8.3 Comparison of raters’ category ratings by English proficiency: Analysis of exams by level of proficiency .................................................................................................................. 168

Chapter 6: Discussion .................................................................................................................................................. 179

6.1 Overview ....................................................................................................................................................... 179

6.2 Explanation of findings.................................................................................................................................. 181

6.2.1 RQ 1: Native speaker and non-native speaker rater comparison .................................................. 181

6.2.2 RQ 2: Comparison by rater English speaking proficiency .................................................................. 187

6.2.3 RQ 3: Comparison by rater first language ......................................................................................... 192

6.2.4 RQ 4: Linguistic category differences among rater groups .......................................................... 195

6.3 Application to previous literature ........................................................................................................... 201

6.3.1 Application to Language Testing research ......................................................................................... 202

6.3.2 Application to Second Language Acquisition ................................................................................. 207

6.3.3 Application to native speaker construct .......................................................................................... 209

6.3.4 Application to social and political issues ......................................................................................... 213

6.4 Implications ............................................................................................................................................. 217

Chapter 7: Conclusions ...................................................................................................................................... 222

7.1 Overview ................................................................................................................................................. 222

7.2 Conclusions ............................................................................................................................................. 224

7.2.1 (Re)defining the native speaker ....................................................................................................... 224
6.2.2 Justifying the native speaker ................................................................. 231
7.2.3 It does not take one to know one .......................................................... 236
7.2.4 Differences in holistic rating versus analytical rating ......................... 238
7.3 Local implications .................................................................................... 240
7.4 Limitations and future research .............................................................. 244

Appendix A: Interagency Language Roundtable Language Skill Level Descriptions -
Speaking ........................................................................................................ 247

Appendix B: Speech Evaluation Guide ............................................................ 256

Appendix C: Individual Tester Report (ITR) .................................................... 261

References .................................................................................................... 263
List of Figures

Figure 1. ILR Skill Level Descriptions Summary Statements (FBI, 2009)................................. 87
Figure 2. Selected Exams by Level and Examinee Native Speaker Group................................. 112
Figure 3. Original Distribution of Ratings ................................................................................ 114
Figure 4. Revised Distribution of Ratings ................................................................................ 114
Figure 5. NS Raters’ Final Ratings ......................................................................................... 120
Figure 6. NNS Raters’ Final Ratings ......................................................................................... 120
Figure 7. NS Raters’ Final Ratings of L 2/2+ Exams ................................................................. 123
Figure 8. NNS Raters’ Final Ratings of L 2/2+ Exams ................................................................. 123
Figure 9. NS Raters’ Final Ratings of L 3/3+ Exams ................................................................. 124
Figure 10. NNS Raters’ Final Ratings of L 3/3+ Exams ............................................................ 124
Figure 11. NS Raters’ Final Ratings of L 4/5 Exams ................................................................. 125
Figure 12. NNS Raters’ Final Ratings of L 4/5 Exams ............................................................... 125
Figure 13. Histogram of L2 Raters’ Final Ratings ................................................................. 131
Figure 14. Histogram of L3 Raters’ Final Ratings ................................................................. 131
Figure 15. Histogram of L4 Raters’ Final Ratings ................................................................. 132
Figure 16. Histogram of L5 Raters’ Final Ratings ................................................................. 132
Figure 17. Rater Native Language Groups’ Final Ratings for All Exams .............................. 142
List of Tables

Table 1. English SPT Exam Distribution ................................................................. 102
Table 2. Redistribution of Ratings from Original ILR Levels to New Exam Groupings .... 116
Table 3. Inter-rater Reliability Statistics: Native Speaker Group ............................. 118
Table 4. Final Ratings for All Exams: Native vs. Non-native Raters .......................... 119
Table 5. Final Ratings for Each Proficiency Level: NS vs. NNS Raters ..................... 122
Table 6. Analyses of Variance for Native Speaker Group by Exam Level ................. 127
Table 7. Inter-rater Reliability Statistics: Rater English Proficiency Level ................. 129
Table 8. Final Ratings for All Exams: Rater English Speaking Proficiency Level .......... 130
Table 9. Final Ratings for Each Exam Proficiency Level: Rater English Proficiency Levels ... 134
Table 10. Analyses of Variance for Rater English Proficiency by Exam Level .......... 137
Table 11. Inter-rater Reliability Statistics: Rater Native Language .......................... 140
Table 12. Final Ratings for All Exams: Rater Native Language ............................... 141
Table 13. Final Ratings for Each Proficiency Level: Rater Native Language Groups .... 144
Table 14. Analyses of Variance for Rater Native Language by Exam Level ............. 146
Table 15. Linguistic Category Ratings for All Exams: Rater Native Speaker Group .... 150
Table 16. MANOVA for Linguistic Category Rating by Native Speaker Group .......... 152
Table 17. Linguistic Category Ratings for Level 2/2+ Exams: Rater Native Speaker Group .... 154
Table 18. MANOVA for Linguistic Category Rating by Native Speaker Group: Level 2/2+ Exams ........................................................................................................ 155
Table 19. Linguistic Category Ratings for Level 3/3+ Exams: Rater Native Speaker Group .... 157
Table 20. MANOVA for Linguistic Category Rating by Native Speaker Group: Level 3/3+ Exams ........................................................................................................ 158
Table 21. Linguistic Category Ratings for Level 4/4+/5 Exams: Rater Native Speaker Group. 160

Table 22. MANOVA for Linguistic Category Rating by Native Speaker Group: Level 4/4+/5
Exams............................................................................................................................................. 161

Table 23. Linguistic Category Ratings for All Exams: Rater English Speaking Proficiency ... 165

Table 24. MANOVA for Linguistic Category Rating by Rater English Proficiency ............ 167

Table 25. Linguistic Category Ratings for Level 2/2+ Exams: Rater English Proficiency .... 169

Table 26: MANOVA for Linguistic Category Rating by Rater English Proficiency: Level 2/2+
Exams............................................................................................................................................. 170

Table 27. Linguistic Category Ratings for Level 3/3+ Exams: Rater English Proficiency ...... 172

Table 28. MANOVA for Linguistic Category Rating by Rater English Proficiency: Level 3/3+
Exams............................................................................................................................................. 173

Table 29. Linguistic Category Ratings for Level 4/4+/5 Exams: Rater English Proficiency .... 176

Table 30. MANOVA for Linguistic Category Rating by Rater English Proficiency: Level 4/4+/5
Exams............................................................................................................................................. 177
Chapter 1: Introduction

1.1 Overview

Most language proficiency exams rely on human raters to evaluate productive skills such as speaking and writing. Despite efforts to minimize the subjectivity inherent in human rating, individual characteristics of the raters introduce error to test scores (Barnwell, 1989b; Eckes, 2008; Galloway, 1980; Kang, 2008; Ludwig, 1982; Reed & Cohen, 2001). The validity of language test results is often of great importance since many standardized language test scores are used to make high stakes decisions (Messick, 1996; Shohamy, 2001), such as course placement, course or program final evaluations, employment suitability, incentive pay eligibility, or assignments abroad (FBI, 2009). Consequently, researchers have made extensive efforts to determine the significant sources of rater variation and to reduce the influence of such sources though improved rater training, rater selection criteria or various other means (Reed & Cohen, 2001). The main objective of such research is to minimize the variation caused by human raters in order to produce test scores that more accurately reflect the examinee’s true language ability (Reed & Cohen, 2001).

Research to date has indicated that raters’ scores are affected to some degree by such rater attributes as age, gender, occupation, international experience, personality, cultural background and opinion (Barnwell, 1989b; Eckes, 2008; Galloway, 1980; Kang, 2008; Ludwig, 1982; Reed & Cohen, 2001). In addition to these, researchers in Language Testing have also focused on whether or not the rater is a native speaker of the test language introduces bias and
affects the final test rating. Originally much of the research that has been conducted on native versus non-native speaker raters concluded that there were fundamental differences between native and non-native speakers produced or perceived language (Barnwell, 1989a; Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979). These early studies based their conclusions on descriptive statistics, such as comparing means or counts, and since then research has evolved to employ statistical tests, such as tests of significance and additional testing contexts. Recent research suggests that whether or not a person is a native speaker does not significantly affect rater reliability or test score accuracy (Barnwell, 1989b; Eckes, 2008; Kim, 2009a; Kim, 2009b; Zhang & Elder, 2011), which challenges the idea that native speakers are preferable raters to non-native speakers.

Comparing native and non-native speaker rater performances is because it is more difficult to clearly define the variable in question, a native speaker, than other rater characteristics, such as age, gender, or occupation, which may be easier to clearly categorize (Davies, 2003). When native speakers have addressed in Language Testing and Second Language Acquisition research in the past, they were often undefined with the assumption that the reader understands what it means to be a native speaker (Escudero & Sharwood-Smith, 2001). Although the precise characteristics associated with the native speaker in a study may be unclear, researchers still draw conclusions about what it means to be a native speaker or what native or non-native speakers can or cannot do. Unfortunately, it is often the case that the social and political impacts of test results are not properly considered. Testing practices can create de facto policy (Hawkey, 2006; Shohamyy, 2001). The unspecified use of the native speaker in Language Testing or Second Language Acquisition research, and consequently results based on such research, leads to assumptions about what native speakers can or cannot do, when the
results may be caused by attributes of the native speaker, such as language proficiency, rather than whether or not a person is a native speaker in itself. Ambiguity in what a researcher or test practitioner means by ‘native speaker’ can lead to misunderstandings about what a test says about a person’s identity (McNamara, 2007). Additionally, the fairness of the test can be jeopardized if conclusions made about native or non-native speakers are over generalized or improperly attributed to a rater’s nativeness (Elder, 1997; Kunnan, 2000, 2004).

1.2 The ubiquitous and ambiguous native speaker

The native speaker concept has been explored in a number of studies in Second Language Acquisition and Language Testing. Thus, Paikeday (1985) proposes two views of the native speaker. Accordingly, the popular definition of the native speaker’s definition is “a person who has a specified language as the mother tongue or first learned language” (pp. 9-10). The second view the idealistic sense of the native speaker as “one who is a competent speaker of a specified language and who uses it idiomatically” (p. 10).

Alan Davies has written several publications about the role of the native speaker in Second Language Acquisition and Language Testing. Davies (2004) identifies some attributes of a native speaker that he claims are commonly accepted by the public at large. Everyone is a native speaker of his/her own code. Everyone adheres to norms of his/her own language. The standard version of the language is favored and advantaged. Native speakers have intuitions about their standard languages. Lastly, native speakers have access to a language faculty. Yet, even with this purportedly common understanding of what it is to be a native, being native means different things to different people. In addition to these common statements about native speakers, a native speaker according to Davies (2004) is defined as a person who acquired
his/her first language by where he/she is in childhood; a person who has intuitions about his/her ideolectal [unique, personal] grammar; a person who has intuitions about standard grammar; a person who is widely fluent, spontaneous, with a large vocabulary and communicative competence; a person who writes creatively; and a person who has a unique capacity to interpret or translate into his/her first language. Davies’ (2004) characterization of the native speaker is complex, but has been widely accepted and referenced by researchers (Muñoz & Singleton, 2011; Zhang & Elder, 2011). Most research that utilizes the native speaker does not specify whether or not all these characteristics are met.

Hyltenstam and Abrahamsson (2000) argue that a person can be classified as a native speaker by any number of avenues. A review of research on maturational constraints in Second Language Acquisition led them to believe that it is justified to categorize a native speaker by birth, by virtue of being a native user, or by being an exceptional learner. Moreover, one can become a native speaker through education in a target language or through residence in an adopted country. This approach to being a native speaker addresses the relationship between labeling someone as a native speaker and identity. In each of these cases, the speaker has encountered a significant amount of exposure to the language, and identifies with it. This approach does not consider the speaker’s proficiency level.

1.2.1 Comparing native speakers

In a number of places in the linguistics literature (Davies, 2003; Maher, 2001; Paikeday, 1985), a native speaker is defined in two primary ways. A native speaker can refer either to the way a person learned the language (e.g. since youth and in an environment where the language was spoken) or to a level of language attainment (e.g. someone who reached a highly articulate
level of language proficiency). When native speakers are used as raters or evaluators, they are often selected because the meet the former definition (they acquired the language natively), and then are assumed to have the latter definition (they have a high level of language acquisition). However, the variation in what is meant to be a native speaker makes it difficult to compare results on native speakers’ rating abilities from study to study.

In the context of Language Testing (Abrahamsson & Hyltenstam, 2009; Barnwell, 2009b; Galloway, 1980; Kang, 2008; Kim, 2009b; Johnson & Lim, 2009; Marinova-Todd, 2003), native speaker judges are selected as participants because they acquired the language in a native setting. They are then assumed to be high-level acquirers of the language and suitable arbiters of language correctness (Escudero & Sharwood-Smith, 2001; Paikeday, 1985). However, this relationship has never been explored and this assumption has not been justified.

1.2.2 Native speakers in as language arbiters

The native speaker arbiter is used both in Second Language Acquisition and in Language Testing. The native speaker concept is prevalent in many acquisition theories. Corder’s Interlanguage Hypothesis (1967, 1981) asserts that second language acquisition is the development of an interlanguage grammar that is based on Universal Grammar (UG) and influenced by the native language. This idea contrasts with the then prevailing theory that learning a second language was a process of restructuring and readapting the rules of the native language (Tarone, 2012). Most relevant to this study here, is the much debated Critical Period Hypothesis, which indirectly asserts that non-native speakers are not likely to be able to attain language to the same extent as native speakers because of the delayed onset of acquisition (Sorace & Robertson, 2001). Based on these hypotheses, native speaker and non-native speaker
comparisons have been prevalent in language acquisition. Some studies conclude that differences exist (Bley-Vroman, 1990; Coppieters, 1987; Hyltenstam & Abrahamsson, 2000; Lenneberg, 1967; White, 2003) and some show no marked differences (Birdsong, 1999; Chiswick & Miller, 2008; Montrul & Slabakova, 2003; White & Genesee, 1996). As a result, many researchers favor the latter opinion, finding it difficult to support the idea that the critical period is any more than a hypothesis.

The concept of the native speaker is as prevalent in Language Testing research as it is in SLA research. The use of the native speaker term is further muddled in the Language Testing context because native speakers are not only used as arbiters, but also as a reference for the top tier of a rating system. Many of the standardized rating scales that are used in government, private and academic contexts refer to the native speaker as an ideal speaker of the language, or a target to acquisition. In the United States government, the native speaker is mentioned in the Interagency Language Roundtable (ILR) Skill Level Description for Speaking Proficiency (ILR, 1985) as a label for the highest rating, Level 5: Functionally Native Proficiency. The native speaker here represents the ideal language user. Since the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines for Speaking Proficiency (Swender, Conrad, & Vicars, 2012) has been developed from the ILR Skill Level Descriptions, it is no surprise that it also uses the native speaker reference as well. The top band of the Common European Framework of Reference (CEFR), C2, also refers to the native speaker (Davies, 2011). Nevertheless, it is typical to select native acquirers of the language to evaluate examinees instead of raters who have demonstrated their language abilities, confounding native speaker constructs.

Although the native acquirer of a language may not be a completely inappropriate yardstick by which to measure a person’s ability (Davies, 1999, 2003, 2011), it is problematic
because the average native acquirer of the language may not demonstrate a skill as high as the examinee and the average native acquirer is not a homogenous concept (Bachman, 1990; Hamilton, Lopes, McNamara, & Sheridan, 1993), leading to variation among a group that is assumed to be a unified concept. Still, Davies states that the idealized native speaker is needed as a model and a goal, although the average or typical native speaker (the native acquirer) is not useful as a measure. For many first and second language learners, the goal is the acquisition of an idealized standard language (Taylor, 2006). So native-born speakers do not automatically acquire it, but have to become language learners themselves in order to reach certain levels of language proficiency (Brutt-Griffler, 2002). Both the native and non-native speakers have the same target language, the standard language (Davies, 2011).

As important as the test results are the social and political consequences of that result. Tests have enormous power. They can define language knowledge; stipulate criteria of language correctness; perpetuate national, hegemonic languages; establish priorities and hierarchies of language policy and status; and, by using simplified rating scales, trivialize the complexity of language (Cheng, Watanabe, & Curtis, 2004; Shohamy, 2001). Social and political ramifications can include issues of score power, correctness, test fairness, identity, and English variants. Perhaps the most relevant of these is the issue of identity, as it is so closely tied with being a native speaker (McNamara, 2007). People define themselves as being native speakers of a particular language, so when they are tested in that language with imperfect results, it may be misconstrued as criticism of their identities. When discussing native speaker of a language in relationship to testing, the issue of “native of which variety of the language” becomes crucial. Although standard varieties of a language have typically been favored (Taylor, 2006), it is important to consider the context of the assessment and the appropriateness of the variety that is
used (Jenkins, 2006b). A determination of which variety or varieties of a language are acceptable then affects what is considered correct in a test. When measuring an examinee according to a native speaker standard, the issues of correctness often derive from an abstract ideal of the language that is rooted in preconceptions about language (Hutton, 1999). All of this reflects back on the powerful impact that a score has, and its wide-reaching effects beyond the test itself (Shohamy, 2001).

1.2.3 Contributions of the current study

What the current study proposes to offer to the current literature and debates is an expansion on what has previously been discovered by investigating differences between native and non-native speakers evaluating high-proficiency examinees, in an operational rating context with highly-trained raters using explicit rating criteria. The US Federal Government Language Testing context offers an opportunity to explore differences between what native and non-native users can notice about language and how they evaluate language. The US Federal Government regularly tests examinees who are highly proficient, articulate speakers of the language. Additionally, they employ highly articulate native and non-native speakers of language who have all undergone the same training in language testing. This scenario provides an opportunity to discover if significant differences appear between native and non-native speakers when dealing with language at levels that are typically higher than what is available at a university setting, where much of such research is conducted.

This study also explicates what is meant by a native speaker in testing and language acquisition contexts. Typically, the native speaker refers to a native acquirer or an ideal speaker of the language. The research then interprets the native speaker to mean these two theoretical
concepts as a) a rater’s first language and b) a rater’s speaking proficiency to reanalyze the data collected under the new construct. Grouping raters by their first language or level of language proficiency instead of just by whether or not they are native speakers explores whether the characteristic that is important in rater selection is the nativeness aspect itself, the first language or the proficiency.

Lastly, this research is interested in looking beyond the singular, final rating that the rater assigns to the process that the rater uses in arriving at and justifying that decision (Zhang & Elder, 2011). This process is essential to rater reliability and test validity (Bachman, 2007). In the present study, the linguistic category ratings are also examined for functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness (FBI, 2009). Significant differences may appear among these rater groups in the different linguistic categories that do not affect the final score. Even though the final rating is not statistically affected, this information sheds light on whether or not raters from the different groups place different value on a linguistic category or perhaps assign more or less accurate ratings in those areas. The linguistic category ratings could reveal that raters from certain backgrounds in terms of nativeness are more or less able to notice certain features of the language. Differences in the linguistic category ratings could impact the reliability of the raters, if they are rating using varying construct. As a result, the validity of the test would be impacted.

1.3 Research questions

At the present time, native speakers play an important role in Second Language Acquisition and Language Testing, but is the emphasis on whether or not a person is a native speaker really warranted? Since being a native speaker or a non-native speaker rater may impact
the final rating assigned (Barnwell, 1989; Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979), the first research question is posed.

1. Do native and non-native speaker raters assign comparable ratings on speaking tests?

If no difference in ratings were found, the results would support the inclusion of non-native speaker raters, and making the categorical exclusion of raters based on nativity alone invalid and discriminatory. A significant difference between the ratings of native and non-native speaker rater groups would not only support a preference for native speaker raters over non-native speakers, but demonstrate some fundamental differences in the way native and non-native raters perceive or evaluate language.

Previous research on native versus non-native speakers has not given full attention to the importance of strong rater training and socialization to a common, explicit set of rating criteria (Barnwell, 1989b; Kang, 2008; Kim, 2009b; Zhang & Elder, 2011). By controlling these variables, it is less likely that raters’ evaluations will be affected by idiosyncratic beliefs (Winke, Gass, & Mayford, 2012). Also, the extensive training that the raters in this study have received will answer whether or not non-native speakers can be trained to rate like native speakers. If so, it would indicate that differences between the evaluative abilities of highly proficient native and non-native speakers can be overcome with proper training (Shohamy & Gordon, 1992).

Additionally, being a native speaker has previously been confounded with speaking proficiency by the assumption that native speakers are fully proficient and justified language evaluators since they are used as the standard in research (Barnwell, 1989b; Coppetiers, 1987; Davies, 2004; Eckes, 2008; Fayer & Krasinski, 1987; Hill, 1996; Kang, 2008; Kim, 2009a;
Native speakers are presumed to have high proficiency (Baron, 1982; Paikeday, 1985); yet, non-native speakers can at times have speaking proficiency that exceeds that of natives (Brooks & Brau, 2007), so nativity is not the best indicator for proficiency. Since being a native speaker is sometimes confused with being a highly-proficient speaker, the raters were regrouped from a native non-native speaker division to a division by proficiency of the language being evaluated.

2. Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?

Moreover, being a native speaker can refer to the fact that a given language was the first language that a speaker learned (Davies, 2003). Instead of grouping all the non-English first language learners as non-native speakers, it seemed appropriate to regroup the raters by their first language learned. The variation that occurs among the non-native speaker raters may be due the fact that they come from a variety of first language backgrounds. They are not just non-native speakers of English, for example, but they are native speakers of Arabic, French, or Mandarin. It is possible that a rater with a native language that is closely related to English, such as another Indo-European language, would more reliably evaluate English than a non-Indo-European language, such as Vietnamese. An analysis of this grouping can determine if having a particular first language impacts a rater’s ability to reliably evaluate.

3. Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?
It is possible that whether or not a person is a native speaker is less of a contributing factor to reliable speaking test rating than the characteristics that compose being a native or non-native speaker, such as speaking proficiency level and first language acquired. Results from research questions two and three that indicate that rater proficiency level determines the reliability of ratings more than being a native speaker would support the consideration of an ability factor, speaking proficiency rating, rather than a categorization based not on an ability, speaking proficiency, but a demographic variable, whether or not a language was natively learned.

If the results of research question two indicate that rater speaking proficiency level has a significant effect on reliable rating, then the raters should be required to have a certain (likely high) level of proficiency. The research results from this study would indicate what the minimum proficiency requirement for a rater would be. An examination of raters grouped by their first languages rather than by native and non-native speakers would indicated whether non-native speakers of some linguistic backgrounds may make more reliable raters than those raters of other linguistic backgrounds, or first languages. Likewise, it could be that the first language acquired by a rater has an effect on the ability to rate accurately.

Comparability of test ratings does not mean equivalence of tests. Not only must test results be comparable, but there must also be comparability of test constructs (Bachman, 2007). Further analysis on how the raters arrived at their ratings is warranted (Zhang & Elder, 2011). Thus, an analysis of the linguistic category ratings assigned gives insight as to the rationale behind the rating and potential differences in what the raters notice about the examinee’s speech. An exploration of rating processes between the two rater groups (the native and the non-native
raters) will be investigated in order to examine the values which raters give to different linguistic components and how those components contribute to the final rating.

4. Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

If non-native speakers assign the same linguistic category ratings as native speakers, non-native speakers would be further justified as raters. However, if non-native speaker raters based their ratings on significantly different evaluations of the subcomponents of speech than native speaker raters, then they would be using a different construct for speaking proficiency than native speakers, affecting the overall validity of the test (Bachman, 2007).

1.4 Research summary

This study compares native and non-native speaker rater groups in a Language Testing context that includes examinees with a high proficiency of English. Such data is found in the operational testing context at the FBI. Thirty current, certified raters volunteered for the study from eight different first language backgrounds: English, Arabic, Farsi, French, German, Mandarin Chinese, Spanish, and Vietnamese. The rater participants are all FBI Speaking Proficiency Test (SPT) raters, and they normally rate tests in their native languages using the same ILR scale, and are all ILR Level 5 speakers of their native languages. Additionally, the rater participants all have an excellent level of English proficiency, with ILR speaking ratings of Level 2+ through 5. These raters evaluated recordings of twenty-five previously administered English SPTs, which follow an oral interview format.
The raters are instructed to conduct reviews of the tests, assigning ratings and completing the accompanying rater reports. The tests sample represents the diverse population of English language examinees that the FBI encounters. Examinees ranged in English proficiency from mid to high level speaking, and were both native speakers and non-native speakers themselves. Each rater assigned ratings for the final, holistic ability of the examinee, as well as separate ratings for the linguistic categories, including functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness.

To address the first research question, the English testers (n=6) compose the native speaker group, and the raters of other language tests (n=24) compose the non-native speaker group. To address the second research question, the raters are further subdivided by language proficiency in English, the test language, resulting in four rater groups: Level 2+, Level 3/3+, Level 4/4+, and Level 5. To address the third research question, raters are not divided by their proficiencies in English, but strictly by first or native language background. The comparisons of the final ratings given by three different rater groupings (native/non-native speakers, English proficiency, and first language) address the first three research questions. The fourth question is answered through an examination of the same rater groupings of linguistic category ratings. All statistical tests are preceded by analyses of the descriptive and inter-rater reliability statistics. When final ratings are examined, Analyses of Variation (ANOVA) are used. When linguistic category ratings are analyzed, Multivariate Analyses of Variation (MANOVA) are used. All twenty-five English SPTs are analyzed together at first, and later they are divided into three groups by final rating (Level 2/2+, Level 3/3+, and Level 4/4+/5 exams).

The results indicated that there was no significant difference between the native and non-native speakers’ final ratings over all tests, or at any of the three test levels. When the raters were
regrouped by rater English speaking proficiency, than there was a significant difference among the groups, largely between the Level 2+ English speakers and the other rater groups. Differences were also present among some of the rater first language groups. Analysis of the linguistic category ratings resulted in significant differences between native and non-native speaker groups as well as English proficiency groups.

1.5 Implications of research

Results of the current study show the impact of whether or not a rater is a native speaker on rater reliability, but also contribute to understanding of larger theoretical issues, including the construct of native speaker, characterizations of learner end states of acquisition, and understanding of social issues such as the power of tests, fairness in test evaluation, the use of standard forms of English, and the impact of testing on identity. Although some claim that the native speaker is only an abstract construct (Lantolf & Frowley, 1985) or a myth (Davies, 2003), the proposed research’s expanded context over previous studies contributes to the debate on the existence of distinctive native speaker intuition (Davies, 2003; Richards, Platt, & Weber, 1985) and the native speaker myth, suggesting language proficiency as the appropriate primary criterion for determining qualified language arbiters (Paikeday, 1985).

The data collected contributes to end state language acquisition theories by examining the perceptive, evaluative abilities of highly proficient non-native speakers of English compared to high-achieving English native speakers. The concept of native-like attainment (Birdsong, 2005) extends beyond the native speakers used in much Critical Period Hypothesis literature, who are not highly proficient speakers, (Johnson & Newport, 1989; Long, 1990) to include speakers at the highest levels of language proficiency. The study builds on current Second Language
Acquisition research (Marinova-Todd, 2003) that gives a comprehensive analysis of adult native-like second language learners across a range of receptive language acquisition domains, including vocabulary, pronunciation, fluency, and pragmatic use, rather than a focus on only one linguistic feature in isolation, such as fluency, pronunciation, or grammar. Results contribute to a refined conceptualization of ‘native-like’ acquisition, with attention paid to the relationship between the learners’ first and second languages (Birdsong, 2005), which in this case are the native language and the language being rated.

The utility and necessity of the native speaker in the US Federal Government context is essential to the Assessment Use Argument (Bachman, 2007) for the validity of speaking proficiency testing, justifying or negating the role of the native speaker in SPT rating. Results from the proposed study are significant at a social level, as language assessments are related to issues of identity (McNamara & Roever, 2006). Any factor that affects a language test’s results, such as whether or not a rater is a native speaker, should be closely scrutinized by Language Testing researchers because of the power of such scores in granting access and authority in a society (Shohamy, 2006). Also, the beliefs and policies that underlie a language test, such as whether or not a native speaker can be the only arbiter of language correctness can lead to de facto language policies, determining notions in a society such as exclusivity and can perpetuate ideologies (Shohamy, 2006).

For the past ten years, various scholars have called for additional research justifying the use of testing procedures and practices within the US Federal Government, one of the United States’ largest investors in language instruction. There is a need for accountability of public funds and an assurance that measures used by government agencies meet professional standards for validity and impact (Bachman, 2007). This research takes a first step toward answering that
call by documenting rater behavior and examining and providing an expanded interpretation of the Federal Government’s native speaker criterion (Chalhoub-Deville & Fulcher, 2003).
Chapter 2: Literature Review

2.1 Overview

For over thirty years, language test research has been concerned with whether or not native and non-native speakers are comparable raters of speaking proficiency tests. There has been no broad consensus on the answer to this question. A number of research studies have found differences between final ratings given by native speaker and non-native speaker raters (Barnwell, 1989a; Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979). These studies claim that the two groups do not give equivalent ratings on oral tests and therefore cannot be used interchangeably. The conclusions of the studies confirmed the practice that native speaker raters should be preferred over non-native speaker raters. However, an increasing number of studies (Barnwell, 1989b; Eckes, 2008; Kim, 2009a; Kim, 2009b; Ludwig, 1982) have found no significant difference between final ratings given by native speaker raters and non-native speaker raters, challenging the idea that native speakers are preferable raters. Recent research has looked beyond the single measure of the final rating in the analysis (Zhang & Elder, 2011) and paid attention to the comparability of the evaluation native and non-native speakers go through in rating, which reveals whether or not raters from the two groups are employing the same construct of proficiency. A review of relevant linguistics research will help to elucidate the main issues surrounding the use of native speakers as language proficiency test raters. These issues include:

- What does it mean to be a native speaker?
- What are the social and political implications of characterizing people as native and non-native speakers?
• Do you need to be a native speaker to perform certain language evaluation tasks accurately?
• What is the standard to qualify as a speaking test rater? Is nativity a sufficient trait for reliable ratings? If not, then, what are the appropriate qualifications?
• Does being a native speaker matter in holistic scoring differently than in analytical scoring of various language features?

In order to unpack the issues surrounding the use of non-native speakers as raters, it is important to first understand what it means to be a native speaker. Hence, definitions and uses of the native speaker term date back to 1859 (Marsh, 1859). Most researchers operate on the assumption that the native speaker concept is clear (Escudero & Sharwood-Smith, 2001), but studies by Paikeday (1985) and Davies (2003, 2004, 2011) point to many different notions of the native speaker features. The native speaker’s definition in Second Language Acquisition, Language Testing, and the US Federal Government is also explored.

A better understanding of how the native speaker is defined is just at the preliminary phases of research. It is equally important therefore to look beyond the test instance at the social and political ramifications of using the native speaker construct in Language Testing (Shohamy, 2001). This is so since the native speaker construct involves issues of identity, both for the rater and for the examinee (McNamara, 2007). It affects the rater, who may be required to be a native speaker, and for the examinee, who is measured against a native speaker-based rating scale (FBI, 2009). The native speaker construct introduces questions about privileges given to certain standardized forms of a language (Jenkins, 2006a), which if shared by the examinee and rater, may bias the rater toward the examinee’s performance. It requires judgments to be made about
what is considered correct in a language (Hutton, 1999; Shohamy, 2006). The native speaker construct reveals the influence a test score can have on an examinee (Shohamy, 2001) and it also affects the overall fairness of the test (Elder, 1997; Kunnan, 2000, 2004).

After exploring how native and non-native speaker raters are compared in Language Testing research, the idea of the native speaker is further investigated by examining literature on additional rater characteristics that are confounded with the native speaker concept. Studies on the effect of the raters’ first languages and the raters’ English proficiency levels on their judgments show some effect on rater bias, but these are not consistent. It appears that the most meaningful differences between native and non-native speaker raters are found in the analyses of the process of how raters came to their final scores (Zhang & Elder, 2011).

The native speaker has historically had a central role in Applied Linguistics, including both Second Language Acquisition and Language Testing. A review of what that role has been and how the native speaker has been defined will enlighten the exploration of the utility and appropriateness of the language testing qualification for raters of being a native speaker and add support to the arguments for or against the existence of the concept in general. A thorough consideration of the native speaker is presented, including its uses in Applied Linguistics and its social and political ramifications beyond that field of study.

2.2 Native speaker perspectives

2.2.1 History of the ‘native speaker’

Before a meaningful discussion about whether or not raters must be native speakers of a language in order to perform evaluation tasks adequately is broached, the concept of what it means to be a native speaker must first be explored. Unfortunately, discussants rarely identify
what they mean by the term, either under the assumption that the reader understands what a
native speaker is or because it hard to agree upon a definition. Establishing a definition for
‘native speaker’ is a troublesome task. The Merriam-Webster Online Dictionary does not list an
entry for the term. However, the Oxford English Dictionary (OED) defines the native speaker as,
“a person who has spoken a specified language since earliest childhood, as opposed to a person
who has learnt it as a second or subsequent language in later life” (OED, n.d.). The OED notes
that ‘native speaker’ was first referred to in George Marsh’s inaugural address at Columbia
College in 1859, in which he advocates the formalized and mandatory study of English in
America. Marsh is quoted as saying, “[T]here is enough of grammatical inflection to familiarize
the native speaker with syntactical principles imperfectly exemplified in French and English”
(Marsh, 1859). Hackert (2009) investigates this early native speaker reference using a discourse-
historical approach, unveiling the political undertones implicit in Marsh’s use of the term during
his Columbia College address. Even in the first recorded reference to the native speaker, Marsh
uses the term as if the audience fully understands its meaning and the societal implications of the
meaning. Marsh seems to apply the common definition for native speakers: people who have
acquired the language in a native manner, so called native-born speakers. Native-born speakers,
as a class, are thought to have unique intuitions about use of the language.

Hackert also examines other early adopters of the term: William Dwight Whitney (1875)
and Henry Cecil Wyld (1906). Whitney associates native speakers with “standard” and “good”
language use, defined by the speech of educated Englishmen. Whitney does not see nativity as a
dichotomous variable, i.e. a proposition in which one is either native or non-native. Instead,
Whitney sees being a native speaker as a matter of degree, with some speakers “more native”
than others. Native speech is, in this case, necessary for upper class membership, and Whitney’s
definition of it marginalizes English dialects that retain what are considered undesirable aspects of the language, such as pronunciation features found in American and Irish accents. Likewise, Wyld advocates the acquisition of the English pronunciation exhibited by the educated citizens of England, whom he refers to as native speakers. Again, the concept of the native speaker is associated with a person who speaks an ideal and desired form of the language. Hackert proposes that the ambiguous and divisive nature of the native speaker term might be due to the fact that its early uses are politically charged, as select ‘native speakers’ carry the burden and privilege of determining the acceptability of the language.

From its first documented uses, the native speaker has been considered a commonly understood definition, someone who speaks as a native and like a native. The term has been generally considered to be desired and favored, carrying with it social status. Therefore, the fullness of what it means to be a native speaker has been understood to mean more than the sum of its component words,—i.e., speaking like a native. It implies a model of speech, a standard speech, and a goal for non-native speakers.

2.2.2 Native speaker definition reconsidered

One of the earliest discussions into understanding the complex and layered role of the native speaker is lexicographer Thomas Paikeday’s 1985 book, *The native speaker is dead!* This book chronicles a discussion with over forty notable linguists about their perceptions of the native speaker. Paikeday interviews these linguists to try to unravel their perceptions of what it is to be a native speaker and whether or not the term has any utility.

Previous references to the native speaker take for granted that ‘native speaker’ had but one understood meaning, which includes both the ideas of being native-born and having ideal
speech. Paikeday (1985) proposes separating these definitions of the native speaker, with one referring to how the language is acquired and the other referring to competent, proficient, or ideal speakers of the language. The popular definition refers to the language acquisition method, “a person who has a specified language as the mother tongue or first learned language” (pp. 9-10). This first definition implies continuity of language use and learning since birth or at least within the critical period or the onset of puberty. (Operational definitions of what this means are still being debated.) Woolf (1973) specifies demographic requirements for an arbiter of language correctness beyond just where a person was born: “a native speaker of English and hav[ing] at least a bachelor’s degree from a reputable college or university” (p. 10, 1973, quoted in Paikeday, 1985).

Paikeday also recognizes a second, idealistic sense of the native speaker as “one who is a competent speaker of a specified language and who uses it idiomatically” (p. 10). Paikeday further explains what he means by idiomatically; “I mean in ‘the usual way in which the words of a particular language are joined together to express thought’ (Webster’s New World, idiom, 2) or in ‘the syntactical, grammatical, or structural form peculiar to a language’ (Webster’s New Ninth Collegiate, idiom, 1b). Such a definition should cover all professional linguistic uses” (p. 24). In this definition, the native speaker is assumed to be a highly competent speaker. Also native speakers are assumed to be qualified language reviewers, as in the statement “The manuscript was read by Iryce Baron, James Hartman, and Joseph B. Trahern, Jr. (himself a native speaker of West Saxon), all of whom made many useful suggestions for its improvement.” (Baron, 1982, p. iii). Paikeday and the other discussants acknowledge that neither definition of native speaker presumes the other.
In an effort to delve further into what each of these definitions means, Paikeday encounters numerous theoretical issues that prevent clear conclusions. As to the latter definition, native speakers are better described as proficient speakers, who are highly capable of precise and crafted elocution, rather than merely competent speakers, who may be able to discern linguistic differences but be unable to display them. However, the level of proficiency needed to be considered a native speaker is not established. Native speakers who are proficient users of the language can “make immediate and intuitive judgments of a linguistic (and related cultural) kind” (Paikeday, 1985, p.74); they are deemed the arbiters of grammatical acceptability. Paikeday also asserts that native-born speakers may be capable of making judgments of some simple grammatical structures, but are not necessarily capable of making evaluations of advanced matters. They must be highly educated (how/how much is undefined) to make such evaluations. Often, errors committed by native speakers are not judged to be as serious as those same mistakes made by non-natives, and instead are considered to be merely creative uses of the language or slips of the tongue. The ambiguity and social baggage attached to the term ‘native speaker’ leads Paikeday to conclude that the native speaker is dead; it is of no relevant use to linguists. Instead, proficiency should be the focus. It is not enough to know a language; a speaker must be able to employ it. Since proficiency is variable, “the proficiency of any speaker of a language at any given point in time could be tested objectively on the basis of criteria we could lay down for achieving the purpose for which the proficiency is required” (Paikeday, 1985, p. 87-88).

Maher (2001) discusses the same basic two notions of the native speaker: the social native speaker (the native-born speaker) and the ideal native speaker. The ideal reflects the grammarian’s notion of correct and perfect usage, aligned with Chomsky’s notion of Universal
Grammar where language usage is a reflection of a uniform innate structure. The native-born (social) native speaker is reflective of the sociolinguist’s idea of competence, which is measured by one who has acquired the language from within a society that utilizes the language; Maher’s ideal native speaker is thought to physically manifest itself in the social context, being fully accepted as native in a native language community through performance. The native achieves his or her own status by being accepted by fellow members of the ideal native speaking community as one of their own, regardless of whether or not a person acquired the language natively. According to Maher’s terminology, native speakers ‘appropriate’ the language, and no one, not even native speakers, who may or may not have been raised in a culture using the language, own the language in any authoritative sense. They have both biological competence and competence stemming from the social relations in the world, giving them access to the operational meanings of language beyond its strict interpretations. Native speaker competence is dependent on membership in a community of knowledge, where acceptance is gained from judgments made on a group scale by the other native speakers.

In Paikeday’s (1985) published discussion with Noam Chomsky and other contemporaries on the definition and existence of native speakers, he raises the issue of whether or not it is up to the native speaker linguist to prove the validity of the native speaker concept. In return, Chomsky proffers the view that the native speaker does not exist in its pure, ideological form. However, experts in a field do not normally require such a level of purity in a concept – an ability to reach an undisputed, black and white conclusion – in order to have a meaningful discussion about the concept and reach consensus about how to categorize the concept. Chomsky also theorizes that language is a genetically determined organ that carries with it misleading philosophical and historical baggage, leading to false assumptions about the nature of
the native speaker. He suggests the possibility that the language faculty’s initial state does not change through the process of puberty until its end state. Philosophical notions and influences from society result in each person being a native speaker not of a standard language, but of his or her own idiolect (Paikeday, 1985).

Alan Davies’ discussions on the nature and usefulness of the native speaker concept have formed the foundation for the understanding and use of the concept in Applied Linguistics. Davies (2003) investigates psycholinguistic, linguistic, and sociolinguistic aspects of the concept of a native speaker, as well as applied aspects of language such as native speaker-specific knowledge, communicative competence, and intelligibility. Davies’ goal is to determine whether being a “native speaker” is solely a self-ascribed term for identity, or an objective, definable reality.

Davies (1991, 2003) describes the linguistic competence of a native speaker in terms of three distinct grammars: the idiolectal grammar (an individual’s personal set of grammatical rules), the shared grammar between a group of native speakers (the standard grammar of a language), and the grammar of a human’s language capacity (Chomsky’s Universal Grammar). In the discussion of whether or not someone is a native speaker, the first and third grammar types are not controversial. Everyone by definition is a native speaker of his or her own idiolect and everyone is a native speaker of Universal Grammar. It remains, therefore, that the definition of a native speaker has most to do with a consideration of whether a not an individual speaks a shared, standard grammar. Identifying the standard form of a language is in itself a thorny issue, as a language may have multiple standards (British, American, Australian, South African and other Englishes) and within each standard type variation exists.
Most definitions of the native speaker either have to do with how a person has acquired or how well a person knows a language. Both of these views of a native speaker involve linguistic knowledge, which in itself is not a uniform concept. Davies (2003) describes four knowledge types related to language ability: metalinguistic, discriminating, communicational (relating to competence), and skills (relating to performance). A native speaker’s metalinguistic knowledge supports intuitions about language. Discriminating knowledge describes the native speaker discernment about language, allowing them to make judgments and evaluations about language ability. Communicational knowledge refers to the ability to understand the language, and skills, the ability to express thoughts orally. All of these abilities influence a person’s ability to administer and evaluate a speaking test. Since knowledges are fundamental to language ability, they are additional aspects that should be considered in exploring whether differences between native and non-native speakers exist in a rating context.

Ultimately, Davies (2003) describes the native speaker as a person who (a) acquires the language as a child, (b) has intuitions of productiveness and acceptability of the grammar of his or her own idiolect, (c) has intuitions of the standard language grammar as distinct from idiolect grammar, and (d) produces fluent, unrehearsed discourse employing a massive array of vocabulary with an extensive ability to display communicative competence. These four descriptors of the native speakers address both the native acquirers (a) and the ideal speakers (d). It is uncertain whether (b) and (c) belong to the former or the latter. When posed with the question as to whether or not a learner can become comparable to a native speaker, Davies states that it is possible to meet all of the above criteria except for (a), childhood acquisition, because it is contrary to the definition of a second language learner. Davies therefore acknowledges that theoretically there is a distinction between the demographic and ideal concepts in the native
speaker, but does not consider the two to be mutually exclusive. He suggests that with enough practice with the language and contact in a native speaking society, most of the criteria can be met, though it is difficult and rare. The criteria themselves mostly describe the differences typically encountered between natives and non-natives, but do not explain why a non-native would be prohibited from being considered a native speaker.

Davies (2004) finds that differences between native speakers and high-achieving non-natives are typically psycholinguistic rather than sociolinguistic. Davies finds that most non-native-born speakers have difficulty with more than just accent, but also with speed and confidence in grammaticality judgments.

Although Davies finds distinctions between native and non-native born speakers by performance measurements such as fluency, grammar, and cohesion, Davies asserts that distinctions between the two groups by psychological measurements are unable to be tested, at least by current testing methods. Therefore, one must rely on autobiographical indications to distinguish the groups. According to this idea, a policy that excludes non-natives based wholly on the assumption that a non-native could not reach the native speaker ideal may be unjustified based on research to date, though he concedes that even high-level second language learners are very unlikely to achieve the native speaker ideal. In conclusion, becoming a native speaker is in many ways a social issue, one of confidence and identity from being raised within a society that uses the language (Davies, 2003).

With such ambiguity in the term “native speaker,” the usefulness of such a term is in question. The continuing usage of the native speaker concept has been challenged by Paikeday (1985), who declared the native speaker dead, and Davies (2003) who considered whether or not the native speaker was a myth or reality. Rampton (1990) suggested that the varying meanings of
native speaker and mother tongue be replaced by other terms, such as language expertise, language inheritance, and language affiliation. Davies (2011) suggests differentiating between a native speaker and a native user, with the native speaker having acquired the language as a child and a native user being someone who uses the language proficiently. Butcher (2005) gives a history of the term’s origins and uses, noting that many meanings for ‘native’ are negative, stemming from colonialism. Moreover, ‘speaker’ is a unilateral term, which places no emphasis on communication, in the way that a term ‘interlocutor’ or ‘communicator’ would. Butcher recommends that ‘native speaker’ be used only to refer to how a language was acquired. Lantolf and Frawley (1985) agreed with Paikieday and Davies that the native speaker exists only as an abstract idea.

Whether or not a native speaker is an abstract ideal or a real concept has not had a large impact on its use: people, both inside and outside linguistics, commonly refer to the native speaker concept. Nevertheless, a better understanding of the role of the native speaker in Applied Linguistics and Language Testing is important because of the layers of assumptions about ability, status in society and competence that are associated with it. This way, researchers and pedagogues can more specifically identify what they mean by the term.

2.2.3 The native speaker in Second Language Acquisition

Research in Second Language Acquisition often contrasts native and non-native speaker performances to examine issues of ultimate attainment: Can a learner of a language reach a native speaker-like level? In this research, evidence that non-native speakers are able to demonstrate competence or performance in the use of various linguistic features comparable to that of native speakers informs language acquisition theory by supporting the idea that there is no
inherent difference between a native speaker and a non-native learner of a language. In effect, this would mean that languages could be learned at any state in life, leading to the theory that there is no fundamental difference between native language and Second Language Acquisition. Differences found between highly proficient language learners and highly proficient native speakers imply that there could be some end state difference between those who learn the language in an immersion context from birth and those who commence language learning later in life. Gaining information on language acquisition supports and improves language learning practices by giving insight to how language is cognitively acquired. In Applied Linguistics, the term ‘native speaker’ is used both to refer to how a person acquired the language as well as a model to whom learners are compared, implying that native speakers use ideal language. Considering that these two definitions are ambiguous and overlapping, much of the confusion around what a native speaker actually is causes murkiness in interpreting Second Language Acquisition research that uses the native speaker concept.

Discussions on whether non-native speakers can attain all aspects of native language acquisition typically interest SLA researchers who focus on the issue of ultimate attainment, which attempts to define the end state of language acquisition for learners. Debates about ultimate attainment can be traced back in part to Corder’s Interlanguage Hypothesis (1967, 1981). Corder asserts that Second Language Acquisition is the development of an interlanguage grammar that is based on Universal Grammar (UG) and influenced by the learner’s native language. The interlanguage grammar is a language in its own right, and it develops along the ‘interlanguage continuum’ as it is constantly restructured by each learner to conform more and more to target language features. Selinker (1972) emphasizes that an interlanguage is its own system, separate from the native or second languages, that includes all linguistic subsystems,
such as phonology, syntax, morphology, semantics, pragmatics, and sociolinguistics. This idea contrasts with the theory at the time that learning a second language was a process of restructuring and readapting the rules of the native language (Tarone, 2012). Therefore, the second language is cognitively constructed in much the same way as the native language, minimizing the potential for differences between first and second language acquisition. However, Corder believes Second Language Acquisition is constrained by fossilization occurring before full native-like competence can be attained, supporting the critical period hypothesis (Sorace & Robertson, 2001).

Many SLA researchers subscribe to the critical period hypothesis, which states that there is a limited developmental period during which it is possible to begin achieving native-like levels of a language (Lenneberg, 1967), although there is no consensus on the age by which a learner must learn a second language in order to be truly considered bilingual (Singleton, 2012). White (2003) discussed the extent to which first language grammars constrain Second Language Acquisition and the effects of these constraints. Such research proves relevant when considering whether a non-native speaker of a language should be tasked with evaluating the intricacies and nuances of appropriate language use. If native language dominance does have a strong influence on second language acquisition that begins after the hypothesized critical period, or even if it constrains the second language acquisition in unpredictable ways, the ability of a non-native to rate language may be compromised.

Research since the Critical Period Hypothesis, Corder’s Interlanguage Hypothesis, and Universal Grammar were introduced suggests that these theories are not as dependable as once believed. Emerging second language grammars are no longer thought to be so distinct from first language grammars; current research’s focus is on the extent of differences between the two,
with only certain subtle features distinguishing natives from non-natives (Sorace & Robertson, 2001). Differences between first and second language acquisition end states may occur because the initial mental state of an adult learner is different from that of a child. Differences may not be found in performance as much as in competence, where knowledge representations of non-natives may ultimately be different from those of native speakers (Sorace, 1993). Even if such differences are only in competence and not in performance, there may be serious implications for the roles native and non-native speakers should play in language learning and testing. Research in this area is complicated because such differences may or may not be evident in speech production, and competence, an internal ability, may be difficult to measure (Sorace & Robertson, 2001).

Another reason for differences in end states of first and second language learning is the differing processes that first and second language acquisition undergo, as, for example, discussed in the Fundamental Difference Hypothesis (Bley-Vroman, 1990). Bley-Vroman posits that the operation of UG partially deteriorates with age. As Hawkins and Chan’s (1997) Failed Formal Features Hypothesis suggests UG may not deteriorate uniformly. Although access to the main principles of UG is not subject to a critical period, access to parameterized formal features are subject to a critical period. Partial deterioration of the subcomponents of language supports the modular nature of the language acquisition device, where areas like phonology, the lexicon, pragmatics, and other areas of language are function independently of each other. The Failed Formal Features Hypothesis predicts that non-natives who began language acquisition post-critical period may be limited in some linguistic features, and perhaps only in competence, but not necessarily evident in performance, where weaknesses can be avoided or disguised.
Research in the Failed Formal Features Hypothesis paradigm explores which features are affected by age, and to what extent they are affected (Montrul & Slabakova, 2003). Further evidence for partial native-like acquisition is demonstrated in Lardiere’s (2006) longitudinal case study of an English learner in the United States, who reaches native-like proficiency in some areas, but plateaus in others. Even if the competence of only some linguistic features is affected, this may be critical to the ability of non-native speakers to be language evaluators. DeKeyser and Larson-Hall (2005) review several studies that show discontinuities in language development that they attribute to the critical period, particularly in grammar and pronunciation. They mention that there is little evidence of a non-native adult learner who has acquired the language perfectly. A lack of such evidence in itself cannot be considered evidence that it is due to maturational constraints, but rather to quality and quantity of input.

Coppetiers (1987) investigates whether learners who appear on the surface to have achieved native performance also have native competence. Observations of interviews with native and non-natives led him to hypothesize that the different groups had varying intuitions of French sentences. Coppetiers chooses non-native speakers on impressionistic grounds (a limitation to his conclusions) and compares subjects’ interpretations of a range of grammatical constructions. Results indicate that although native and non-native groups did not diverge significantly on formal features discussed in UG, there is greater divergence in grammatical aspects related to cognition and function.

Coppetiers’ study has been partially replicated by Birdsong (1992) with different results. Birdsong strengthens his study by having all participants with the same native speaker background and expanded the tasks Coppetiers used. Unlike Coppetiers, Birdsong finds few competence differences between the natives and near-natives, even in some of the same
structures as Coppetiers. Based on Coppetiers’ study alone, we would expect that it is important to use only native speakers as raters, as there are significant differences between their competencies. However, Birdsong’s subsequent research negates many of Coppetiers’ findings, suggesting that there is no difference between using native and non-native speakers as raters.

Many linguists do not subscribe to the critical period hypothesis; some believe instead that ultimate attainment in a second language has the potential to equal that of the first language (Birdsong, 1999). Some attainment studies have shown that second language learners can obtain equal native performance and competence in areas of phonology and syntax (Montrul & Slabakova, 2003). White and Genesee (1996) administered grammaticality judgment and written production tasks to natives and near-natives, concluding that there are no differences in terms of accuracy or speed between the two groups. Furthermore, White and Genesee’s participants show no decline in performance with age of initial exposure to the target language, supporting the notion that ultimate attainment in a second language can equal that of a native speaker, at least in some aspects of language.

Montrul and Slabakova (2003) have improved upon previous research by identifying near-native learners at the end state of acquisition by administering a proficiency test to all participants, natives and non-natives. Those non-natives who scored within the same range as natives were deemed near native. The results indicate that native and near-native competence of preterit and imperfect verb tenses could not be differentiated, supporting the notion that functional categories of universal features remain accessible even through adulthood.

Most current language acquisition researchers do not subscribe to the critical period hypothesis as strictly as it was originally formulated: that after a certain maturational age, learners would not be able to attain native-like levels of another language. Hyltenstam and
Abrahamsson, 2000) have conducted a review of end state related research to determine whether all, some, or no learners can achieve a native-like state, but have found that there was no agreement on the matter. They state that when detailed and systematic research is conducted, differences appear between native speakers and non-native learners of the language. Birdsong (2005) replies to this by saying that if a researcher looks hard enough at the details of a non-native speaker’s performance compared to a native speaker, they are likely to always find some differences.

Previous research (Birdsong, 1999; Coppeliers, 1987; Montrul & Slabakova, 2003; Sorace, 1993; White & Genesee, 1996) has typically used the definition of a native speaker as a native-born speaker, instead of the native speaker as an ideal speaker, as the selection criterion for participants, resulting in a broader range of performance and competence measures within the native speaker group and a greater acceptance of learners as near-native equivalents. Birdsong (2005) notes that, “more studies should look at individual [second language] learners’ end-state attainment across a range of linguistic behaviors, to determine if nativelikeness, when observed, is in fact limited to narrow domains of performance” (p. 324-325). Not only should definitions of natives and near-native speakers better reflect the end state, but they should also measure competence across a number of linguistic features in a wide range of proficiency and processing tasks (Birdsong, 2005). Following Birdsong’s observations, the evaluations of high-level language tasks attempted by high-proficiency speakers can offer data to contribute to our understanding of ultimate language attainment.

The use of native speakers as an end state point of comparison is in itself a complicating factor for SLA research. Non-native speaker learners are bilinguals, and these bilinguals are usually compared to the performance of a monolingual native speaker (Cenoz & Genesee, 1998).
Muñoz and Singleton (2011) discuss the appropriateness of the use of the native speaker as a yardstick for ultimate attainment in age of acquisition literature. They suggest that it is problematic to compare these two different entities. As an alternative, SLA researchers might simply analyze the language that the non-native speaker produces, rather than having a comparison (Cook, 2002). Ortega (2010) proposed that native speakers as an acquisition target be replaced by the bilingual speaker.

Historically, the native speaker concept has had an integral place in SLA research. It has been implicit in basic theories, such as interlanguage and critical period hypotheses. The native speaker, and typically the monolingual native speaker, has enjoyed the place of being the ideal speaker against which non-native speakers are compared. In most cases, SLA research assumes that the reader knows what a native speaker is (Escudero & Sharwood-Smith, 2001). The native speaker is still not consistently defined in SLA studies, and researchers have not found consistent trends when comparing learner performances to native speaker performances. In recent years, perspectives on those theories are evolving, and so are ideas about taking for granted that the native speaker is an appropriate model for learners.

### 2.2.4 The native speaker in Language Testing

As in SLA research, the native speaker concept is often used as the criterion by which speakers are evaluated in language proficiency testing and, as in SLA research, ambiguities with the native speaker concept problematize its understanding and appropriate use. Nevertheless, many of the rating criteria employed by large testing organizations use the concept of the idealized native speaker in rating scales.
In the United States Government, the native speaker is mentioned in two manners in the Interagency Language Roundtable (ILR) Skill Level Description for Speaking Proficiency (ILR, 1985): as a measure of comprehensibility (e.g. “Pronunciation is understandable to natives used to dealing with foreigners” (ILR, 1985); and a label for the highest rating, Level 5: Functionally Native Proficiency. In the former instance, the native refers to the native-born speaker, one who has an innate ability to set an empirical (lower) limit of comprehensibility. The latter instance refers to the native speaker as the ideal language user, and is used at the top level of the scale, ILR Level 5, “Functionally Native Proficiency.”

Since the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines for Speaking Proficiency (Swender et al., 2012) has been developed from the ILR Skill Level Descriptions, it is no surprise that it also uses the native speaker reference as well. The references that appear in the ACTFL Guidelines parallel how the ILR Skill Level Descriptions use it: i.e., as a native-born speaker who makes implicit evaluations of comprehensibility.

The use of the native speaker in language proficiency scales and other forms of language evaluation criteria is not limited to the United States. The highest level of the Common European Framework of Reference (CEFR), C2, is also the native speaker level (Davies, 2011). In this case the native speaker is again used as the idealization of the standard language. The International English Language Testing System (IELTS) rating scale has nine bands, and the top band refers to the native speaker: “ Produces consistently accurate structures apart from ‘slips’ characteristic of native speaker speech” (British Council, 1995). Here again, there is an assumption that it is known what is characteristic speech of native speakers, or idealized standard speech.

Davies (1998, 2003, 2011) examines the utility of the native speaker as a measurement yardstick. A speaker’s performance equivalent to general native speakers (a developmental
measure) demonstrates a certain level of mastery; however, it lacks discrimination between average speakers and exceptional speakers, who are capable of superior levels of language skill (an attainment measure). Still, Davies states that the idealized native speaker is needed as a model and a goal, as the average or typical native speaker is not useful as a measure.

Much of the doubt that arises from using the native speaker concept in Language Testing is there is much variation among the performance abilities of native speakers, yet when they are to uniformly applied as a rating criterion, a yard stick, there is an assumption of a unified native speaker construct. Native speakers cannot be considered homogenous, even though there is often an assumption that they are (Bachman, 1990). Even considering the different types of native speakers, both the native-born speakers and the ideal speakers, there is variation in each of them.

In Hamilton et al. (1993), the homogeneity of native speakers is explored by reviewing three studies examining native speaker performance on the International English Language Testing System (IELTS) exam. In Hamilton (1991), no significant difference is found between native-born students from three different concentrations within a vocational school, but native-born speakers’ scores ranged between 10 and 27 (with a maximum possible score being 37). This result not only shows that native speakers do not automatically achieve that highest test score in an assessment of their native language, but that there is a great amount variation among the native speaker group. Sheridan (1991) conducted a similar study with a writing test that showed comparable results. Lopes (1992) replicated Hamilton’s study, but with three more highly educated groups of native speakers (junior barristers, post-graduate students, and academic staff). In this instance there was a significant difference between the groups of native speakers, though no group had perfect scores.
Brooks and Brau (2007) examine the final scores of native speaker performances on Arabic, Mandarin and Spanish speaking assessments and found wide variation between the performances, ranging from ILR 2+ to ILR 5. The researchers examined hundreds of native speaker examinees who took the Speaking Proficiency Test as a part of their hiring process and looked at their overall outcomes. The examinees generally had at least one university degree and range in age from the mid-20s to the 70s. Even though these examinees were all “educated native speakers” (meaning native-born speakers who have completed higher-level education) of the language in which they were tested, they scored throughout the top half of possible ratings, with many in the ILR Level 3 range. These results support the idea that the concept of the ILR Level 5 native speaker concept is really an embodiment of the idealized standard language, which is the form of a language typically used in large-scale assessments, and not the simply a native acquirer with a high level of education, as the native acquirers did not receive a Level 5 rating.

This acquisition of an idealized standard language is the goal for both first and second language learners, so native-born speakers have not naturally acquired it (Brutt-Griffler, 2002). Both native and non-native speakers have the same target language, the standard language (Davies, 2011), and both need to become language learners to achieve it.

Davies (2011) explains that the role of the native speaker in Language Testing is to exemplify the idealized standard language. The native speaker’s language, in this sense, would be the standard community language, not the native speaker’s idiolect, or the human capacity for language or Universal Grammar. The standard language is instead the convergence of a community of people’s idiolects that are similar enough to communicate. In Language Testing, the community’s standard language is the one that matters, and this language is a goal for both native and non-native speakers. Davies points out that language proficiency testing is about the
acquisition of the standard form of the language, justifying the idealized native speaker is useful as a yardstick.

Native speakers are in many ways the same as non-native speakers, learners of the standard language. Since neither native non-native speakers are homogenous groups, than there is no clear logical argumentation for the preference of native speakers over non-native speakers as raters. Analyses both of the final ratings given of native and non-native speakers as well as how the native and non-native rater speaker groups arrived at their final rating is therefore require to determine whether there is a basis to preference one group over the other.

2.2.5 The native speaker in the United States Federal Government

Even though the US Federal Government tries to be specific about what it means by a native speaker, confusion about the term still results from general misunderstandings and misuses of the term by examinees, particularly when it is used to refer to a Level 5 speaker. The US Federal Government uses the idealized notion of a native speaker in its language testing practices. This ideal speaker serves as a practical model against which language proficiency is measured, allowing raters to make judgments as to what is correct and appropriate in language use. The commonly understood native speaker as a native acquirer of the language is confused with the US Federal Government’s idealized use of the term. Further explication of the Federal Government’s native speaker is needed to avoid such confusion.

When the “native speaker” is mentioned in reference to the ILR Skill Level Descriptions, it typically refers to the speaker that appears at the top of the ILR scale in ILR Level 5. The use of “native speaker” in this context is really a shortened version of the full reference in the ILR: “speaking proficiency [which] is functionally equivalent to a highly-articulate native speaker”
This native speaker is not an average person who has learned the language in a native context, but is an ideal user of the language. ILR’s Level 5 native speaker is often referred to in US Federal Government materials as the well-educated native speaker (WENS). The concept of the ILR Level 5 WENS is also referred to by the condensed level title “Functionally Native Proficiency.” “Functionally Native Proficiency” may appear to suggest that the speaker needs only to function as well as a typical “native” would to receive this rating, but in fact the typical native-born speaker does not function at this high a level. Out of the approximately 3,000 native-born speakers the FBI tests every year, there are only a handful of people who are rated ILR Level 5 (Brooks & Brau, 2007).

In the ILR Level 5 Description, the phrase “functionally equivalent” is interpreted to mean that a person does not need to have acquired the language as a first language to receive a rating of ILR Level 5; one doesn’t need to be a native-born speaker, one needs only to function in a manner equivalent to a WENS. Although it is extremely rare, there are cases where a non-native speaker has received an ILR 5. “Functionally equivalent” also refers to the functional nature of the ILR scale. It does not mean that the examinee needs only to be able to complete the functions that a WENS can complete to receive an ILR Level 5; to the contrary, the examinee must complete the functions in a manner and of a quality equivalent to how a WENS would complete the functions, including all aspects of WENS speech.

The ILR Level 5 native speaker needs also to be “highly articulate.” By “highly articulate,” the ILR means a speaker who not only does not make any errors or mistakes, but also consistently and appropriately employs the language to its fullest capacity. Speaking in a Level 5 manner involves conveying complex ideas in a clear and precise manner and using sophisticated strategies to tailor the message to a wide variety of audiences, doing so in an effortless and
natural manner. The Level 5’s vocabulary is nuanced and precise, employing terms that fit the meaning exactly. The structures employed are complex when appropriate, and do not include mistakes often made even by native speakers. The fluency has no noticeable hesitations or restarts, and the pronunciation is standard for the language. The message that the speaker conveys, including the direct and meta-messages, suits the situation perfectly and is highly effective for achieving the speaker’s purpose. This type of language is one that few normal native speakers ever attain, and one that is not typically attained though routine life interactions. The development of “highly articulate” speech must be extensively practiced and cultivated (FBI, 2009). The profile of the person that most likely would attain an ILR Level 5 is a “highly articulate well-educated native speaker,” but this does not mean that all well-educated native speakers are highly articulate. Raters recognize that ILR Level 5 is a range. There are a variety of ways that “highly articulate” speech may manifest itself.

Much of the misperception about the ILR’s definition of WENS comes from ILR’s reference to “well-educated.” Often, “well-educated” is taken to mean someone who has completed higher education, perhaps a Bachelor’s or Masters degree, as it would be understood in today’s educational system. When the developers of the ILRs thought about education, they imagined the equivalent of a high-level liberal arts education in the 1950s. Indeed, the ILR Skill Level Descriptions, which were begun in the 1960s and adopted by the government’s Office of Personnel Management in 1985, refer to education:

"Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education; however, in cultures where formal higher education is common, the language-use abilities of persons who have had such education are
considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language. (ILR, 1985)

Although the ILR defines “well-educated” in the context of the language-use abilities of educated individuals, it does not establish a metric for how well-educated the speaker must be or how much weight must be given to the “well-educated” aspect of the “Highly Articulate Well-Educated Native Speaker” (HAWENS). In a literal sense, no weight is given to the education of the speaker; instead, the manner in which Level 5 speakers convey themselves gives the impression of a high level of education. The overarching requirement of ILR is that the examinee be “functionally equivalent” to a HAWENS. Most university graduates receive between ILR Levels 2+ and 4. However, in Language Testing practitioner circles it is frequently opined that the focus of higher-level education has devolved from the 1960s to the present day, and has led to a shift in the speaking ability of graduates. Fifty years ago, university students pursuing a liberal arts education focused on higher-level thinking, oratory and written skills, with the study of philosophy, rhetoric, and debate. Nowadays, university degree programs focus on preparation for a profession, which often means little opportunity to develop extemporaneous speaking skills. It is therefore possible that the context in which the above-quoted language of the ILR was written has changed.

Even though the native speaker concept is politically and socially charged, as well as practically fraught with difficulty, simply eliminating it might not be the right solution. Raters need criteria against which to rate a performance. If the “ideal” definition of a native speaker is not the appropriate model for test criteria, then what is?
According to Bachman’s (2007) Assessment Use Argument, a test must be considered valid for a particular use, and along with that, the test user should consider “the interpretations and decisions that are to be based on assessment performance as well as the consequences of using an assessment and of the decisions that are made.” The US Federal Government context offers a valid use for such a thing as the WENS. In some cases, the US Federal Government wants an employee that is entirely indistinguishable from a highly articulate well-educated native speaker: diplomats must be able to interact with high-level officials from foreign lands without causing offense; complex treaty negotiations are dependent on the nuances and implications of words; and undercover agents must maintain a covert status with no foreign accent or other linguistic traits revealing their identity. The purpose of US Federal Government language testing is distinct from most other language testing contexts, where the goal is a communicative ability and where it is acceptable to preserve some characteristics from the first language. Since the US Federal Government needs employees who communicate at the ILR Level 4 or 5 (levels which include functions that academic scales do not reach), it is appropriate that the WENS, the ideal speaker that embodies all of the characteristics at that level, is used as the model. When considering the use of the native speaker in the SPT, we must consider its use in the US Federal Government context as well as the larger sociocultural context of the test, and its implications (McNamara & Roever, 2006).

Not all native-born speakers are exemplary speakers of the language, as the distribution of outcomes from language testing makes plain. The typical high school or college learner in the United States receives two or three years of classroom foreign language study, and reaches an ILR proficiency level of 0+ to 1+ in the second language. Language majors in college, particularly those who have included immersion experiences as a part of their studies, commonly
reach between the 1+ and 2+ Levels in the second language; it is less typical to attain L3 or higher. Heritage speakers have a wide range of outcomes as well, depending on how much the language was used at home, travel experiences, and formal education in the language in addition to home use. Heritage speaker outcomes typically range between Levels 2 and 3. Even native-born speakers who are raised and educated through all of their developmental years in a native-speaking environment typically reach only Levels 2+ to 3+. In order to receive a rating beyond this level, even the native speaker must become a conscious adult learner of his or her own native language, which involves developing a sophisticated and precise lexicon, cultivating skill in employing pragmatic and sociolinguistic strategies that are culturally appropriate and highly-effective, and practicing this skill in order to improve the fluidity of speech across a broad range of contexts. Opportunities to develop linguistic skills to this level are rare and are not emphasized even in higher-level education. This is part of the reason that so few individuals ever attain the functionally equivalent highly articulate well-educated native speaker level.

2.3 Social and political implications of the native speaker concept

Tests can define language knowledge; stipulate criteria of language correctness; perpetuate national, hegemonic languages; establish priorities and hierarchies; and, by using simplified rating scales, trivialize the complexity of language (Cheng, Watanabe, & Curtis, 2004; Shohamy, 2001). Research to date has largely concluded that rater characteristics, such as whether or not a rater is a native speaker of the language being rated, can impact score reliability and therefore test validity, but the ramifications of test results can reach far beyond the testing context itself. The social and political implications of decisions made as a consequence of test results on examinees, score users, and testing organizations are equally relevant. It is therefore
important when designing a study focused on the effect of rater characteristics to think of the consequences of the different possible conclusions of the research at the outset in order to frame the design of study, as well as what the use of the native speaker concept in testing means for the examinee as well as the test construct.

The issues of identity, English standard variants, correctness, score power, and test fairness have been broadly discussed in Language Testing, Word Englishes, language policy, and sociolinguistics literature, and are relevant to the present study’s context. The native speaker concept is obviously more than a linguistic concern; it is a concern of identity and community. Complications arise not only with the discussion of native speakers, but in this study’s specific context English is the medium of the research. Defining a native speaker if English brings about a problem that may be unique to the language, as English is not only a language of its native speakers, but also has ownership by its second language users, those who use English as a Lingua Franca (ELF). Coupled with which variety of English to use is the problem of which standard should be used to evaluate correctness. Decisions about which English to use will impact what raters consider as correct and the overall rating, which has a considerable amount of political power. All together, the overall fairness of a test can be jeopardized if appropriate decisions are not made. Consideration of the social and political implications of the definition of the English standard and an English native speaker have the potential to take on meaning beyond what is originally intended by an examination, increasing the responsibility of those designing and implementing language testing policy and procedures.
2.3.1 Identity and language test scores

The language that a person speaks natively is often associated with their nationality, and plays a significant role in their identity. The relationship between an individual’s native language and identity does not cause much concern when a non-native speaker is tested. On the other hand, when native speakers are tested in their first and primary language, many perceive that their identity is being evaluated. Furthermore, many examinees find it disheartening when they do not receive the top rating in their native language, as if it signifies imperfections in their identities. When the top score is labeled “native speaker” level, what does it mean for native-born speakers who do not achieve it? The examinee may take it to mean that their identity is somehow flawed, when it is only intended to mean that the examinee has not yet fully acquired the idealized form of the language.

Davies (2011) discussed why a person’s language is so closely related to the concept of identity. Most speakers don’t recognize the distinction between Davies’ three grammars (personal, community, and Universal Grammar) and merge them all together into one concept of language. When a person’s native language is assessed, the standard used is of the community grammar, or the standard language, not the personal grammar. Without that distinction being salient to the examinee, they might take the rating as a reflection of their idiolect rather than their acquisition of the standard grammar. Davies points out that native and non-native speakers are equal in their pursuits of the standard grammar: both must become learners.

Davies believes that native speaker issues are fundamentally identity issues. When a person identifies the language(s) of which he or she is a native speaker, then that person is establishing an identity with a community where that language is spoken. It is a sociolinguistic decision about who the person is (Davies, 2011). Davies exemplifies this concept with the story
of Irish novelist John Banville, who writes in English. Banville is quoted as saying “we are not writing in our own language” (Davies, 2011, p. 299). Banville considers himself, however, a native speaker of Irish Gaelic even though he does not speak that language, as it is his identity. He does not identify with English, his first and primary language, as it is associated with the British community.

McNamara (2007) wrote about the complicated relationship between language testing and identity. Identity is closely related to language assessments (McNamara & Roever, 2006). McNamara noted that these issues were first addressed in Foucault’s (1977) works. He described that social identities become available as a function of dominant discourse. Since exams evaluate and prescribe what is acceptable in a particular language, they define the language, and therefore impact identities. Therefore, tests act as mechanisms for the definition of subject positions and for the recognition of subjects; tests create the identities they measure.

McNamara (2007) goes on to explain that all language tests are about identity both at the micro and macro levels. At the micro level is the concept of language proficiency. Language proficiency tests often cycle between context and ability: the target language use situation is conceptualized in terms of functions of communicative language ability, which in turn are generalized to represent the ability to handle the target language use situation. McNamara recommends that testers engage in a theory of the social context of testing, considering how testing decisions impact issues outside the test, such as identity.

Identity issues do not just apply to examinees, but also to raters. Being a native speaker is a part of a person’s identity, associating them with a language, culture and tradition. Raters are currently grouped into native and non-native speaker categories, a reflection on their identities. When being a native speaker is a requirement for a position, a piece of their identity, not their
qualifications, permits or denies them access to a position. It is therefore vital that the native speaker definition is empirically linked to rater performance to justify this policy.

2.3.2 Which English?

English is one of the most widespread languages in the world today; consequently, it is one of the most tested languages. Because English is spoken as a native, standard, or official language in countries on most continents, the question of which English is tested is of great interest and concern to many stakeholders. The field of World Englishes has paid particular interest to this matter, defending against practices where a standard American or British English would be the only form acceptable as correct, particularly in parts of the world where non-natives more often speak English than natives, where English is used as a lingua franca.

Jenkins (2006a) addresses this issue directly in her criticism of large-scale testing organizations’ preference for standard forms, described as “Inner Circle” Englishes, with the exclusion of “Outer Circle” Englishes, such as those spoken in multilingual societies. The goal of World Englishes is to gain acceptance of the diversity of Englishes offered in the Outer Circle—as valid as traditional forms, such as American or British English. Jenkins argues that English accepts the influences from the local language and culture in these areas, giving Outer Circle Englishes their characteristic accents, syntactic structures, varied lexicon, and unique pragmatic features. This is a part of natural evolutionary processes of language contact (Mufwene, 2001). Such influences have brought about a range of educated second language English varieties whose differences from standard native speaker English are considerable. Therefore, Jenkins asserts that to apply traditional English standards and ‘conform to norms which represent the

The arguments for and against which variant of English should be used in testing contexts are generally concerned with the power given to a variant by its association with the test. However, the fact that variants such as American and British English have been historically favored does not mean that there are not still situations where they are most appropriate variants. There have been efforts to create language tests using a local variety of English. In one example, an English test was created for Indonesian teachers of English (Hill, 1996). Unfortunately, this effort has its limitations, as the majority of the general public wants testing in the standard form of the language (Taylor, 2006), implying that they value the standard. As a result, this test was largely unused.

Is it possible to speak of one standard variety of a language? Davies (2011) notes that the traditional linguistic view is that the variation across standard English(es) (American, English, Australian, South African, etc.) is minimalistic enough that it may be considered one standard. He points out a comment made by Quirk (1990) that those who try to promote an individual or local variety of a language find themselves subject to the opinions of those who have power over them, for example employers. Language is community driven. The community in turn creates rules and establishes a standard to facilitate interaction, so the communities themselves become the gatekeepers (Canagarajah, 2005). By this estimation, it seems reasonable that a community should decide the variant of a language that is useful to them and administer tests according to that language standard.

Arguments for the acceptance of all variants of English are important, but more important is that the test uses the variant of English appropriate for the test’s context. In the case of this
study, that would mean that it is most appropriate for US Federal Government institutions to use the American English standard variety. In the case of this study, standard forms of English are appropriate for the SPT’s purpose within the US Federal Government.

2.3.3 Correctness

Another social issue that is relevant to definitions of native speakers and high level proficiency is the issue of correctness. Given the wide variety in speech among native speakers, deciding what is or is not correct in the language test is not only difficult, but it may be perceived as subjective by examinees. After the construct of a test has been well established, it is important for test procedures, including judgments of quality, to be well defined. Test procedures, and in particular rating procedures, may often be as complex as the construct of the test itself. Correctness should reflect the goals and context of the test; this may include notions of correctness that are flexible, yet reliable, and may vary from level to level. This moves correctness beyond the idea of being just grammatically correct, using the prescribed structure, to politically correct, using language that is appropriate for the situation (Ohlander, 2012).

It is accepted that there is no single correct way of expressing oneself in a particular context. It follows that there is not a singular way to be a correct or standard language user, and rigid concepts of being a standard language user perpetuate ideas of standard language dominance. Therefore, language evaluations that insist on inauthentic standard grammar and accent may not appropriately acknowledge the variety of language that exists. At the same time, supporting language ideals can create opportunities to discriminate against undesirable cultural groups from being equally accepted within a society, for example immigrant groups seeking citizenry, higher-level education, or employment. Shohamy (2006) comments that tests, by
having one criterion for correctness, are capable of perpetuating uniformity and standardization according to pre-determined and defined criteria, which constitute the native variety. Tests can be used as mechanisms employed by central authorities to control language and people.

Hutton (1999) explores that connection in yet another way when he claims that the idea of language boundaries and homogeneous ways of defining and using languages, as promoted by linguists, are based on views of both languages and peoples as “pure” and “correct” in opposition to “different,” “impure,” and “incorrect.” Just as there is no single correct way of using language in terms of grammar, lexicon, accent, and all other dimensions of language, there also is no one way to define or to be a correct language user. Procedures, including judgments of quality, need to be as complex as the construct itself. Accordingly, multiple levels of correctness as well as flexible criteria of correctness that reflect the specific context and goals are appropriate. Worded negatively, we could say that rigid “native speaker” correctness criteria no longer apply; worded positively, an explicit focus on content and message intentions is appropriate.

Some critics of how native speakers are used in speaking proficiency assessment criticize the fact that the term “native speaker” appears in testing criteria or test constructs without the testing practitioners’ closer examination of what it means (or does not mean) in regards to the evaluation of correctness in a test and what its role is in the construct of the test (Taylor, 2006). Jenkins also expressed concern that there would be a different standard held for the acceptability of errors displayed by native versus non-native speakers if standard forms were used as the benchmark. She asserted that there is more correctness demanded of non-native speakers on tests than native speakers in life. Instead, language testers should avoid the practice of linguicism: “the valuing of native speaker English language forms above those of non-native speakers even though the former do not lead to greater communicative efficiency for the majority in
international contexts of use” (Ammon, 2000). As a solution, Jenkins offers the acceptance of English as an international language, not owned or governed by any one standard, since most of the meaningful interaction occurs between non-native speakers rather than including native speakers. Language testers should adopt a sociocultural perspective, measuring only success in communication rather than linguistic form.

Taylor (2006) argues that wide-scale language testing needs to reflect the expectations and attitudes of teachers and learners. She elaborates that teachers and learners desire to learn the standard, widely adapted form of the language. Therefore, it is important for the standard form of the language to be used as a model for evaluation. If the focus of standardized testing were completely without regard to accuracy of output, then tests would not reflect what learners want to learn; language testing should be an accurate reflection of the expected output of language learning. Measures of correctness are not focused entirely on grammatical accuracy. Grammatical accuracy cannot be fully ignored however. Without grammatical accuracy, communication is often impaired. Language testing can balance the need for measuring correctness and the expectation for the use of standard language against the need to be fair to the examinee and accommodate the flexibility of language.

Advanced language speakers, or correct language speakers in measurement terms, have been labeled as such through multiple means. In various studies (Bartning, 2000; Bongaerts, 1999; Cobb, 2003; Freed, Segalowitz, & Dewey, 2004; Geeslin, 2003; Kotz & Elston-Güttler, 2004; Liskin-Gasparro, 1998, 2000; Montrul & Slabakova, 2003; von Stutterheim, 2003), advanced language ability has been defined in many ways: as having between 4 1/2 and 6 years of instruction prior to University, a C-1 or C-2 rating on the common European framework of reference (which can be given as an exam result or upon completion of a course), graduate
student status in Spanish, an advanced rating on the ACTFL Oral Proficiency Interview, enrollment in an advanced grammar course, at least two years of college for language instruction, an ILR Level 3 rating, or an accent that is indistinguishable from that of native speakers. An advanced speaker as defined by years of instruction or completion of a course or program is insufficient for government employment purposes because there is no direct correlation between extent of language exposure and language proficiency. What is needed is research that will determine measures for requisite vocabulary size and depth, fluency and phonology, social and pragmatic sensitivity, syntactic complexity, competence and discourse, along with other language features that are appropriate for different contexts, thereby defining what in an exam is correct or not correct.

The issue of language correctness is certainly relevant to the ratings assigned by both native and non-native speaker raters, and relevant to this research. Variation in ratings potentially reveals differences in the ideologies of what the rater groups consider correct language use. The ILR Skill Level Descriptions and their accompanying tools guide raters’ notions of language correctness, appropriateness, and acceptability at the different levels. Raters’ interpretations of the scale are influenced by their perceptions of what is correct for the language. Uniformity or divergence in the evaluations given by native and non-native speaker raters is a commentary on whether there is an agreement as to the rules of a standard language. The US Federal Government needs to evaluate examinee correctness. Categorizing examinees by the ability to accomplish functional language tasks is essential to the context, and so what is correct in this context may differ from academic language acquisition and testing needs, aligning with Hutton’s ideas (1999). Just as the ‘client’ in academia, namely the students, wants to acquire standard, correct language (Taylor, 2006), government agencies as clients need speakers who have the
standard forms of language with varying levels of accuracy for its specific purposes. The results should show that there is a unified concept of correctness among the SPT raters and that definition is appropriate for the US Federal Government context.

2.3.4 The power of a score

A test score is a representation of not only an examinee’s ability, but of policy and values in the testing system, including which standard of a language was used and how correctness was measured. Shohamy’s 2001 book, *The power of tests*, explores the social and political implications of large-scale, high stakes testing, such as those in large academic and government contexts. Shohamy recognizes the impact of tests beyond the decision made by the score. As Messick (1981, 1989, 1994, 1996) states, examinations make determinations and separate people into groups. Such determinations embody values, linked to psychological, social, and political variables, which are often unrecognized and unevaluated. Shohamy also points out that tests have the power to stigmatize people, and make them regard themselves as failures. Tests control access to education, careers, and other economic benefits; they can open and close doors.

Shohamy advocates ‘use-oriented testing,’ a view of testing whereby tests are not seen as isolated from but rather as connected to psychological, social, and political variables that have effects on curriculum, ethicality, social classes, bureaucracy, politics, and knowledge (Messick 1981, 1994, 1996, 1998). The danger of tests is that they can be used in subtle ways to manipulate language and create de facto language policies (Shohamy, 2006). Tests should be recognized as potentially powerful devices that are imposed by groups in power to affect language priorities, language practices, and criteria of correctness leading to inclusion of desirables, exclusion of undesirables, and perpetuation of ideologies. Tests have the power to
manipulate languages by determining prestige and status of languages, standardizing and perpetuating language correctness, and suppressing language diversity. Research that considers the use of native speakers or non-native speakers as raters must deal with the impact of labeling groups: is it fair, appropriate, or necessary to categorize some people as native speakers and some people as not?

According to Shohamy, in order to have full consideration of the power of a test, one must give consideration to the different dimensions of the test. To begin, the influence of the testing body, the organization that developed and/or administered the test, must be examined. The testing body is responsible for the test items. The testing body will inevitably influence how the test is administered and how scores are interpreted. Another important element is the composition of the final score, the extent to which the score represents a single measurement or combination of different measurements across different contexts or measurement sessions. The contents of the test reflect the breadth and depth of subjects covered in test material. Compromises in content are inevitable; however, the content should still reflect a range of information representative of the final score.

The next dimension of testing is the sample of examinees that are evaluated in the testing or in test development. If a sub-sampling of examinees is used, maintaining an accurate representation of the entire population is imperative. It is important to identify and recognize the purpose of the test, whether it is formative, summative, or predictive. The test purpose guides the sampling of test items, the mode in which results are delivered, the stakes of the test, and how test results should be interpreted. Another dimension is the format of the test; the test may be uniform or flexible. Each test format has its place, as uniformity may not consistently results in fairness when sampling of test content is involved. Only after these other dimensions are
examined should one turn to psychometric accuracy. The validity of a test depends on how it is used and how its scores are interpreted. Only after these other dimensions are established can one conduct informed research of test instruments, whose conclusions are applicable to the public at large. Likewise, measures of item or rater reliability are dependent on the other test dimensions.

Similar to Shohamy's framework of test dimensions, Johnson (2009) is also concerned with language measurement through broadly used testing instruments and the implications of the scores and broader issues of language policy. Johnson describes her framework for exploring the impacts of language policy as the ethnography of language policy. Language policy ethnography examines language policy and all of its layers as if one were to interpret a test score by removing the layers of an onion. Johnson describes the test agents, the creators of the test, and those responsible for the test’s interpretation and perpetuation. Test agents often have overt or covert agendas that may bleed into how tests are administered and evaluated. Similarly, agents of language policy perpetuate ideas of the nature of language and language construct through test scores.

Interpretation of language policy can be elucidated only by a thorough examination of the goals of that policy. Language policy goals should be explicitly stated so that the intentions are clearly understood by all stakeholders who are interpreting them and testing organizations can be held accountable for their practices. The processes behind creating a language policy as well as the policy’s implementation and appropriation should be transparent to all stakeholders in the same way that test development should be transparent. It is also important to examine discourse, which may perpetuate or engender the policy. These types of discourse may be present within the policy itself or in other contexts. All of these layers of the ethnography of language policy should
be taken into context with the dynamic, social, and historical aspects that surrounded the policy as it was created, interpreted, and appropriated.

The concerns expressed by Shohamy, Messick, and Johnson are relevant to this study. SPT ratings are part of multiple test batteries that help group FBI applicants into groups such as qualified and unqualified, or label them by ILR rating subtitles such as elementary proficiency (Level 1), general professional proficiency (Level 3), or functionally native proficiency (Level 5) (ILR, 1985). Often examinees remove the word “proficiency” from the description name and use the subtitle to apply more broadly, thinking that their score implies that they are not professional or not a native. However, discriminating between job candidates is a necessary evil. It is challenging to balance the need to select qualified personnel for demanding positions with the responsibility to consider the effects of language policy within the government. As per Shohamy’s recommendations for use-oriented testing (2006), the research design and subsequent result, discussion, and conclusions are framed in the US Federal Government operational testing context, including the test’s uses, construct, administration and rating procedures. The results of the study are discussed balancing the needs of the US Federal Government and the implications of government policy, and considering the ethnography of such policy, as recommended by Johnson (2009).

2.3.5 Test fairness

Issues of test fairness have been receiving increased attention in Language Testing literature in recent years (Elder, 1997; Kunnan, 2000, 2004). Test fairness examines how test scores are used and the implications of these results on examinees and other stakeholders. Furthermore, test fairness is interested in the broad societal ramifications that come about as both
intended and unintended consequences of large-scale testing. Bachman (2007) ties together some of these philosophies expressed by other authors by defining test fairness as the impact and consequences of assessment use (e.g., Hawkey, 2006; Shohamy, 2001), the societal values that underlie such use (McNamara & Roever, 2006) and the larger sociocultural contexts in which language tests are used (e.g., McNamara & Roever, 2006). These ideas combine to support the idea of test validity being measured through an Assessment Use Argument (AUA). An AUA explicitly states the interpretations and decisions that are to be based on assessment performance as well as the consequences of using an assessment and of the decisions that are made (Bachman, 2007).

In order for the SPT to be considered a valid test through an AUA, an explicit argument must be made for its policies and procedures in the framework for it operational constraints and needs, including the criteria for and selection of raters. Aspects of test administration and evaluation such rater reliability and qualifications are vital to such an argument, and should be defended with rational argumentation and supporting research. An examination of the raters’ qualifications, including whether or not they are native speakers and English proficiency, is directly related to its AUA and the overall fairness of the test.

Addressing the myriad of issues surrounding a test takes multiple research projects and regular evaluation of test data. Shohamy, Reeves, and Bejarano (1986) support the idea of including multiple measures of proficiency assessment, including discussions, reports, interviews, conversations, instead of including merely a single discourse. While this initiative is far from complete, it has begun within the US Federal Government context. Lowe (1983) notes
that the context of the ILR is specific to the US Federal Government context, and not directly applicable outside that context.\(^1\) As is appropriate, this discussion of social and political implications of language testing will continue as long as language tests are in use, and help to provide checks and balances on test usage, including the current study of SPT rater characteristics.

With consideration of all of the social and political implications of language proficiency testing, scholars often return to the fact that all of these elements must be weighed in consideration of the local use of the test (Chalhoub-Deville & Wigglesworth, 2005). Test users need to undertake local validation research to help make sure that their interpretation and use of test scores are appropriate in their local academic and professional contexts. The use of native or non-native speakers as raters in the FBI’s speaking tests touches on all of these issues. The SPT’s score has significant power within the organization; raters are tasked with being the arbiters of language correctness. Decisions made about who judges the examinees’ proficiencies and how those judgments are made affect test fairness. The ratings issued from the test and the qualifications of the raters touch upon the issue of whether or not they are equivalent to native speakers. The variant of English considered as standard for the test impacts the overall rating. Indeed, all of these factors need to be weighed carefully to ensure that tests are administered, scored, and the results are used ethically and appropriately.

\(^1\) Lowe concluded this after studies conducted through a 1978 ETS grant for “The Common Metric Project” (1983).
2.4 Native and non-native speaker raters in Language Testing

The field of Language Testing has largely been interested in the comparison of native and non-native speakers as language evaluators. Previous research on the equivalence of native born and non-native born speaker raters in Language Testing has yielded mixed results (Reed & Cohen, 2001). Early research comparing the ratings of native born and non-native born speakers has found differences between the rater groups (Barnwell, 1989b; Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979). However, research both during and following that time period did not always find significant differences between the groups’ final ratings (Brown, 1995; Hill, 1996; Kim, 2009b; Ludwig, 1982; Zhang & Elder, 2011).

Although some of the studies have focused specifically on whether or not native-born speaker raters and non-native born speaker raters gave different evaluations, there is much focus on what was different about the way that the two groups assessed language. Early research in this area looks at teachers evaluating students’ oral productions as more of a question of whether non-native speakers are well suited to be teachers. Data collected is observational and usually descriptive, with emphasis on what native and non-native teachers noticed or emphasized. The foci of the studies are varied, making it difficult to identify trends across varying groups. Observations of teachers give way to studies analyzing differences in raters in larger scale or summative assessments involving written performances, where differences in rater performance can lead to issues of test fairness. By the 21st century, Language Testing research has become more focused on language tests themselves with research techniques becoming more refined, and emphasis more on formal test raters and high-stakes exams. The focus has returned to ratings of oral ability as well as written ability.
The seminal study on native and non-native rater comparisons of speaking ability is Galloway’s 1980 work, which has as its original intent to examine what types of student errors impaired communication, and to look at whether or not classroom-trained students would be effective communicators in the country where the language was spoken. In the study, Galloway has 32 raters evaluate 10 high school students’ video segments in Spanish on informational communication, errors, and overall affective communicative ability. The raters are divided evenly into four groups: non-native speaker Spanish teachers in the US, native speaker Spanish teachers in the US, non-teaching native speakers in the US, and non-teaching native speakers in Spain, who had poor to no ability to understand English.

In the quantitative analysis, no real differences could be seen between the mean ratings of the groups. However, a study of the qualitative data, the raters’ comments, reveals an interesting trend. Overall, the native speaker groups seem more interested in content over form. Non-teaching native speakers exhibit rapport with students who struggled to communicate, but are critical of students who speak with ease and grammatical accuracy.

In his 1989 work, Barnwell challenges the methodology used previously in native versus non-native speaker research where native speaker raters are typically affiliated with universities or other educational institutions. To correct for this, he compares ‘naïve’ native speakers of Spanish with trained American Council for the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) raters’ evaluations of Spanish. In the ACTFL scale levels, there are many references to communication with native speakers of a language, yet typical native speakers are rarely, if ever, used as the raters. Barnwell posits that if a non-native speaker rater (for example a native English speaker rating a Spanish test) is to act as a surrogate for a native speaker judge, then non-native speaker ratings should be comparable to the judgments given by
naïve (untrained) native speakers, who are the actual target interlocutors. In this study, fourteen native speakers of Spanish evaluate four Spanish OPI exams according to the ACTFL scale and these ratings are compared to the exams’ original ACTFL ratings. In general, Barnwell finds similar patterns between native and non-native speaker raters, in that the two groups agree which examinees gave the best and worst performances. However, native speakers are harsher than the non-native speaker raters in some cases. Native speaker rater severity is exaggerated for high-level non-native speaker examinees (Barnwell, 1989b).

In the mid 1990s, many published studies on native and non-native speaker differences begin to shift focus. Earlier research conducted on differences between native and non-native speaker evaluators has taken place in the academic context, with the evaluators being teachers or graduate students, rather than in a formal Language Testing context using trained raters. The shift from classroom-oriented evaluation begins with Hill’s (1996) examination of rater differences in an institutionalized language exam, using trained raters. Hill finds that Indonesian raters were as suitable to rate English exams as native speaker raters from Australia. Moreover, Indonesian raters proved to be more consistent in their ratings than the native English-speaking raters. Hill found the native speaker raters to be harsher in rating overall.

Considering the significant amount of research that has indicated that native speaker and non-native speaker raters apply rating criteria differently—even at times leading to unreliable test scores, as there is no consistency in the criteria that the raters are basing scores on—the question remains as to which group is better, native or non-native speaker raters. Brown’s research (1995) suggests that non-natives have a tendency to follow rating criteria more explicitly, whereas natives use more intuition. Hill’s (1996) study suggests that in contexts where the highest level of proficiency on an assessment is the native speaker ideal, it is more
appropriate to use a non-native speaker. Nevertheless, favoring non-native speaker raters over native speaker raters is still not fully supported, as the conclusions from many of the native and non-native rater comparisons are limited.

Hill’s (1996) foray into formal testing research of native and non-native raters marks a transition in academic research, where differences between native and non-native rater outcomes are determined through statistical tests of significance rather than descriptive statistics. For example, early research uses a comparison of means to represent group differences, which later become analyses of variation and other statistical tests. Analysis of ratings of 220 French student compositions by four native and four non-native speakers teaching assistants in Jolivet’s (1997) study reveals no difference between rater group evaluations. Both native and non-native evaluators tend to focus more on form than on meaning or content of the responses. Porte’s (1999) investigation of error-gravity perception in English essays written by intermediate level university students in Granada, Spain and graded by native and non-native speaker faculty reveal little difference between the groups.

Shi (2001) takes a more in-depth look into written test evaluations by native and non-native speakers. She examines the ratings of ten English compositions given by 46 raters, including both native and non-native speakers. Raters assigned a holistic score on a 10-point scale, with no explicit criteria, and then gave three reasons for each of their ratings. Shi concludes that there is no significant difference between the ratings of the two groups, but they differ markedly in the criteria they used to justify their responses. Native speakers focused on content and remarked positively, focusing on what the examinee was able to do well; non-native speakers commented negatively on organization and length, focusing on what the examinee was not able to do. Shi’s research informs the issue of whether holistic scoring reflects analytical
judgments, consequentially affecting the construct validity of the exam. In this case, the holistic rating does not reflect the analysis behind the ratings, since there is no difference in the final ratings, yet differences appear in the analysis. Differences in how the raters arrived at the final rating may indicate that the raters are using different constructs for the exams. Shi’s research also starts a trend in analyzing and comparing whether evaluative comments given by raters on examinees’ performances are particularly positive or negative. Did they reflect the rater noticing what the examinee could do, or could not do? Following this model, this study also collected information on rater linguistic category sub-ratings and supporting evaluative comment in order to gain insight into the rater evaluation construct.

Lazaraton (2008) reviews many of the studies that examine the use of non-native speakers as evaluators, concluding that there is no justification to exclude non-natives from becoming test raters and supporting the emerging trend to abandon the term “native speaker” altogether. In her recommendations, she notes that rating scales used for speaking and writing might contain native speaker bias; therefore, the roles native speakers play selecting test input, evaluating test responses, selecting test administrators and evaluators, and considering the overall test construct should be carefully considered. According to Taylor (2002), the questions that are left unanswered regarding native and non-native speakers as raters involve: how the examiner’s own language proficiency affects test delivery; how the relationship between the examiner and examinee’s native languages affects test delivery; and whether or not native and non-native speaker raters’ behavior are comparable with each other.

In the latter part of the 2000s, an increasing amount of the research conducted on the native/non-native speaker raters in Language Testing returned to oral production. In a 2008 study, Kang investigated measures of proficiency, comprehensibility, instructional competence,
and accentedness in native and non-native speakers. Both native and non-native speakers evaluated four-minute segments of fourteen native and non-native speaker teaching assistants’ lectures. She finds, in part, that non-native speakers gave significantly harsher scores on ratings of comprehensibility and proficiency, but not on accentedness and instructional competence. Kang speculates that perhaps this difference is due to a native speaker’s more global view of language proficiency. Kang also finds that the more contact that raters had with non-native speakers; the more lenient they were in rating, which lead to higher accentedness scores.

Rossiter (2009) explores differences in native and non-native speaker evaluations of fluency, noting again no significant differences between the groups. Furthermore, Rossiter finds that the rater’s assessments concur with measures of pauses per second and pruned syllables per second in the samples. Most of the negative remarks made by the native born / non-native born raters had to do with fillers, speech rate, pausing, and self-repetition, but Rossiter also finds many comments in reference to pronunciation, grammar, and vocabulary.

Kim’s (2009b) research addresses the question of native and non-native speaker comparability in English speaking tests. In her study, she contrasts the evaluations of twelve native English speaker raters from Canada with twelve non-native speaker raters from Korea, using both quantitative and qualitative analyses. The participant raters evaluate ten examinees, each performing eight different speaking tasks. The tasks produce about one minute of speech each and are rated on a four-point scale ranging from almost always successful to almost always unsuccessful. Kim’s Rasch analysis reveals that the raters are mostly internally consistent and the native and non-native groups assign comparable scores, with only a couple of exceptions. The rater groups also display similar patterns of rating severity across tasks. However, Kim notes that the native and non-native speaker groups do not handle the rating samples equivalently. Native
speakers are found to give more detailed and elaborate ratings, particularly in respect to pronunciation, grammar, and accuracy. To follow up on the second finding, Kim (2009a) conducts a subsequent study to further investigate differences in how native and non-native speaker raters evaluate the speaking performances of students in English as a Second Language courses. The second study focuses on the effect of being a native speaker on score reliability and how such reliability might be affected by adjusting the number of native and non-native speaker raters used. Kim finds that most of the variation in the scores was attributable to examinee ability rather than rater effect. Again, the rater groups show similar severity patterns, though native speaker raters contribute less variance to the overall score.

Most recently, Zhang and Elder (2011) reexamine the comparability of native and non-native speaker ratings of oral language proficiency in the context of a high-stakes, institutionalized language exam, the College English Test-Spoken English Test (CET-SET) in Australia. Twenty non-native raters and nineteen native raters from various rater training backgrounds evaluated ten speech samples of about twenty minutes length each. Even though this study uses an established test, the raters were asked to assign a holistic score to the responses using a 1-5 rating scale that does not include descriptions at each level beyond very poor (1) to excellent (5). Additionally, the raters provided written comments on a subset of 6 tests and the comments were coded for content. Like other recent research, Zhang and Elder do not find significant overall differences in the holistic ratings assigned by the two groups; however, the commentary provided to support those scores does deviate significantly in the frequency in mention of some linguistic categories. The non-native speakers focused on some of the linguistic resources (which included phonology, grammar, and other general linguistic resources) considerably more than the native speakers. Native speakers pay significantly more attention to
interaction, demeanor, and compensation strategies. An examination of the subcomponents of the linguistic resources group revealed that native speakers mentioned vocabulary significantly more frequently than non-native speakers. Overall, the native speakers seem more communication-focused and the non-native speakers seemed more form-focused. These findings, as well as findings of other Language Testing studies comparing native and non-native speaker raters are related to the current study below.

### 2.5 Other concepts of nativeness in Language Testing

#### 2.5.1 Rater native language

Even though researchers have considered the impact that many characteristics of test raters would have on ratings, they have not looked too often at the relationship between the rater’s first language and the language that the rater is assessing. It is hypothesized that raters whose native language is similar to the language being assessed (e.g. a German tester rating English) would be more reliable than those whose first language is quite different (e.g. a Korean tester rating English). This hypothesis is based on evidence seen in research outside Language Testing that listeners tend to favor speakers with similar accents (Gass & Varonis, 1984; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). Further, there is some evidence to support that raters may favor familiar accents (Carey, Mannell, & Dunn, 2011; Fayer & Krasinski, 1987).

Elder and Davies (1998) look at language distance effects for learners of English. In the study, students are grouped into seven different language families according to their home languages. These groups are ordered as being closest to English to furthest from English. After the examinees were assessed in English, it was then determined if their first language’s distance from English had a general effect on their test outcome. The researchers conclude that first
language distance from target language exists, but it can’t be separated from other learner factors.

In 2008, Hamp-Lyons and Davies build on this concept by evaluating the language distance of raters’ first languages from the language they assessed. They evaluate written English assessments of examinees from six different backgrounds, evaluated by both speakers of their own variety as well as raters from a different linguistic background. The authors appear to find some trends to suggest that language distance did affect ratings; however, their sample was too small to be conclusive.

Following up on this research, Johnson and Lim (2009) also look at the influence of a rater’s first language on their scores of examinees of various language backgrounds. In this study of English exams, most of the raters have English as a native language, with the exception of four raters (Spanish, Korean and Filipino/Amoy). Examinees come from 21 different language groups. Here, the effect of bias is minimal overall, so there is no pattern of bias in the ratings.

The most recent investigation of this type is Winke, Gass, and Mayford (2012), who hypothesize that accent familiarity is a cause of rater bias. This study examines Spanish, Chinese, and Korean native-speaking examinees taking the TEOFL iBT™, which is an English test. Native speakers of Spanish, Chinese, and Korean rated 72 examinees. Results indicate that both Spanish and Chinese raters were more lenient with examinees that share their first language, raising the need to address rater first language effects in rater training.

Most of the research that investigates rater first language bias looks at the relationship between the rater’s first language and the examinee’s first language, no matter in which language the test is. These studies have shown mixed results, with some indicating rater bias, and others not finding significant bias. Perhaps the inconsistency in results is an indication that rater bias is
a complex concept, with many intervening variables. Even so, there have been few if any studies that examine the rater’s first language relationship with the test language. The current study addresses this relationship and possible bias by dividing the non-native speaker raters into groups by their first language. This analysis reveals whether or not there are any patterns to rating behaviors by first language.

2.5.2 Rater proficiency level

When native speakers are preferred as raters over non-native speakers, it is often because they are assumed to have high, ideal language proficiency. Raters are typically native speakers, language teachers, or experts in the language, though these rater qualifications do not guarantee uniformity in language ability. Consequently, few studies have actually attempted to measure and document a correlation, or lack thereof, between the language proficiency of raters and the raters’ evaluations.

Native speakers of a language cannot be assumed to all have a high level of language proficiency. In their study, Hamilton, Lopes, McNamara, and Sheridan (1993) administered English exams to several native speaker groups, including vocational students, university students, graduate students, professors, and attorneys, with widely variable results within and across the groups. Even though some language teachers have had to pass a language exam to qualify for their positions, considerable variation in their proficiencies exists, and the exam that they took may have little construct relevance to the test that they were being asked to rate. Other factors such as time in a native-speaking country or other measures of language expertise do not offer a dependable correlation between the amount or type of language exposure and proficiency.
With this variation in rater language ability, it is surprising that more attention has not been brought to the issue of rater language proficiency.

A limited amount of research has been conducted on language teacher proficiency. Elder (1993) investigates the use of an International English Language Testing Service (IELTS) score to determine English teacher qualification. She examines the IELTS as a performance predictor, compared it to university screening procedures, and tries to determine an appropriate cut score for teacher qualification. Elder discovers that in the teacher context, the IELTS score is relevant when dealing with low levels of teacher proficiency, but not for highly proficient teachers. The results indicate that the listening IELTS score was most useful as a predictor.

Elder (2001) digs deeper into these issues, examining the implications and appropriateness of such testing. There is a question as to whether or not it is appropriate to use a proficiency test for a specialty field, such as language teaching. Test developers would need to specify the range of language that is needed for that position. If a test were designed to qualify potential language teachers to be hired, it would also be difficult to create tasks that are relevant to that position and can be easily administered.

It has been the tradition in SLA and Language Testing research that native speakers are employed as comparative models and language evaluators with the assumption that they are model language users (Davies, 2011). Few researchers have taken the time to actually factor in the native speaker’s actual language proficiency. Since studies have shown that native speakers do not possess homogenous language ability, their proficiencies should be assessed.
2.6 Rater evaluation of linguistic criteria

Even though there is no general consensus as to whether or not native and non-native speaker raters assign comparable scores, most studies that have examined raters’ evaluation of different linguistic features in the examinees’ performances found differences between the native and non-native raters. Some studies, such as Barnwell (1989b), reported an overall tendency for native speaker raters to be stricter than non-natives in overall score assignment. Broad assessments like this have a limited utility when investigating differences between native and non-native speaker rater ability. Other research studies have found evaluation differences only in certain linguistic features.

In his meta analysis, Ludwig (1982) reviews twelve studies on the subject, investigating differences between native and non-native groups in terms of comprehensibility, irritation, and acceptability. Ludwig finds that although learner proficiency scores did not significantly differ between native and non-native high-proficiency raters, non-natives focus more on grammar, are more intolerant of errors, and are more cautious in rating high-level examinees.

In the late 1980s to early 1990s, a strand of English as a Second Language research appeared comparing native and non-native teachers’ evaluations of written essays (Connor-Linton, 1995a, 1995b; Hinkel, 1994; Kobayashi, 1992; Sheorey, 1986; Takashima, 1987). Some of the findings suggest that native speakers are affected more by the comprehensibility/irritation continuum, focusing on meaning over form. Raters also have a tendency to base the acceptability of a structure on the likelihood that such an error would be committed within their group. Researchers conclude that non-native teachers evaluate grammatical errors more harshly than native teachers, though they have found some interesting differences in the evaluation processes of the two groups. Sheorey (1986) concludes that both native and non-native teachers find verb
errors to be serious, but that non-native teachers emphasize lexical errors more. Takashima (1987) finds that non-native evaluators have difficulty identifying article, word choice and sentence formation errors. In Takashima’s study, the evaluator and the non-native students whose essays are rated were both Japanese, perhaps adding to the non-native’s difficulty in error identification.

Kobayashi (1992) also discovers differences in the two groups, with native speakers being stricter about grammar, but more positive about organization and clarity of meaning. The difference that Hinkel (1994) discovers between native and non-native speakers does not have to do with errors, but rather pragmatics. In Hinkel’s study, non-native speaker evaluators do not have access to advanced notions and conceptualizations needed for high-level writing, making them less effective raters.

Differences between native and non-native raters’ analyses of the linguistic characteristics of a sample have continued to be identified. Pronunciation and politeness are more strictly assessed by natives (Brown, 1995), while grammar has more weight in scores for non-native raters (Fayer & Krasinski, 1987; van Maele, 1994). Moreover, Reed and Cohen (2001) find that natives tend to focus more on intonation and register.

Zhang and Elder (2011) is another example of a Language Testing study that compared both the holistic scores and linguistic feature analysis of native and non-native speaker raters on a speaking test. Like many others, they do not find a significant difference between the holistic final ratings of the groups, but they do find differences in how the native and non-native raters analyze the samples. These are differences in the amount of comments made by native and non-native raters about the varying aspects of the performances. Zhang and Elder notice that the non-native speaker raters (NNES) make significantly more comments in the area of linguistic
resources (vocabulary and general linguistic resources). Native English speakers (NES) make significantly more comments in the areas of interaction, demeanor, compensation strategy, and other general comments.

Zhang and Elder conclude that the native speaker raters focus on a broader range of aspects of the examinee’s performances. They interpret these results to say that non-native speakers are more focused on form, whereas native speakers are more focused on communication. Zhang and Elder find this result to be interesting because it does not accord with statements made by English as a Lingua Franca (ELF) supporters who argue that non-native speakers are concerned with communication and not form (Jenkins, 2000). A lack of focus on form may make ELF speakers (non-native speakers) disadvantaged in taking tests, as tests tend to emphasize correctness and standard language norms.

When the analysis that raters use to arrive at their final scores is analyzed, differences between native and non-native speakers’ approaches to evaluating the languages emerge. It is therefore not sufficient to limit research between rater groups to one statistical point, the final or holistic rating; it is necessary to delve deeper into the raters’ rationale for the score. Investigating what is behind the score will elucidate whether or not differences between native and non-native raters exist, an important question in SLA and Language Testing.

2.7 Limitations to previous studies

The relation between being a native speaker and speaking test rating reliability should consider all levels of speaking proficiency, not just those typically attained by language learners. In Kim’s (2009b) study comparing native speaker and non-native speaker ratings of English speaking ability, the author endeavors to include a variety of examinee productions in her
sample, including a variety of examinee levels and test task types. Participant examinees are selected from a variety of levels of English language classes and eight different speaking items are developed, including three different tasks. However, Kim includes only language learners in her study, to the exclusion of high-level learners who may be equivalent to native speakers. These administrative constraints diminish the relevance of the samples in relation to the test construct. Additionally, although her examinees come from different levels of English classes, membership in each class is determined by a combined ability across speaking, reading, writing, and listening abilities, and not speaking alone, the skill measured in the study. This led to class levels that are not necessarily representative of the full range of speaker proficiency, limiting the validity and generalizability of Kim’s results. Finally, there is no consideration for the level of the student or the task in the analysis of the data. In other studies, such as Zhang and Elder (2011), the level of examinee responses is described as a “wide range,” but not explicitly defined. Level of task is also not taken into account into the analysis.

Barnwell’s work (1989b) on rater variables examines closely the construct of the ACTFL oral proficiency interview. The author selects tests from four examinees, two at the lower range of the ACTFL scale, and two at the upper range of the ACTFL scale, in an effort to have a variety of examinees. Barnwell has native speaker raters give holistic scores to the samples, because the ACTFL scale references how effective an examinee’s speech would be when interacting with native speakers. Including only four examinees in the sample makes the sample susceptible to the characteristics of those particular individuals, and limits the generalizability of the results. Furthermore, the ACTFL raters are trained to have varying expectations of the examinee’s performance on tasks of different levels, and to consider task level in assigning the final rating. The untrained native speaker raters make no consideration for task level in their
evaluations, but only give their overall impression of the examinee. Therefore, the two rater
groups being compared have differing ideas of the construct being examined, and could not be
expected to perform similarly.

Important to the construct of a proficiency test is the opportunity for the examinee to be
able to display her holistic ability across a range of tasks. Therefore, speaking tests typically give
examinees multiple attempts to display their skill across tasks and topics. For example,
Barnwell’s (1989b) use of the ACTFL speaking test gave raters a representative sample of
language to evaluate, including a variety of language tasks, totaling up to about 30 minutes of
examinee language. Conversely, other studies comparing rater groups have had raters evaluate
very limited sample sizes. Although Eckes (2008) uses an established test, the TestDaF, with a
range of tasks, the analysis focuses largely on the individual task ratings rather than the final
score. Consequently, the results are based on ratings of speech samples that are very short,
probably less than one minute in length, rather than a holistic consideration of individual
examinees’ performances across a range of tasks. Kim (2009b) considers each of the eight task
samples collected from an examinee as separate data points, so she is not analyzing an
examinee’s sample holistically as is typically meant by speaking proficiency. The only valid data
point in such a test should be the final, overall rating which results from a performance across the
tasks. Kang (2008) defends using only one type of task, with responses of less than one minute,
in the rating sample by increasing the rating pool to more than 120 raters. Although there is
strength in the number of raters, Kim’s conclusions are limited to one task, which has the
potential to be inherently unreliable as a representative of the examinee’s proficiency.

Research focused on rater variables typically analyzes criterion-referenced evaluations
rather than norm-referenced evaluations, due to its prevalence in contexts where language is
being evaluated. The rating scale that raters use in their evaluations, therefore, becomes an important influencing variable in this strand of research. Eckes (2008) examines an established test, using its official raters, so he naturally has the raters use the criteria on which the test is constructed for their evaluations of the responses. The criteria explicitly describe the linguistic features required for each level, and the raters are experienced in interpreting the criteria in a reliably uniform manner. Linguistically specific criteria are used in Barnwell’s (1989b) study, where trained raters gave ACTFL ratings as the baseline for native speaker comparison. However the non-native speaker group has no training in the ACTFL rating scale, therefore are highly subject to their individual biases when evaluating the examinees.

Kang’s (2008) study asks raters to evaluate various language features on invalidated, seven-point Likert scales. The raters are asked to give a ‘holistic’ score ranging from “low proficiency” (1) to “moderate proficiency” (4) to “high proficiency” (7). No additional explanation is given for these points, and no linguistic features are attributed to them. Similarly, Kim (2009b) creates a four-point Likert scale for the overall evaluations of the samples in her study. Instead of evaluating linguistic features, raters are asked to evaluate task completion: from almost always successful to almost always unsuccessful. The authors in both of these studies asserted that their experience as language teachers give the participants the requisite guidance needed to reliably evaluate the samples. However, raters who depend on broad, non-linguistic rating criteria are subject to individual biases attributable to their experience and characteristics, such as age, gender, occupation, international experience, personality, cultural background, and opinion (Barnwell, 1989b; Eckes, 2008; Galloway, 1980; Kang, 2008; Ludwig, 1982; Reed & Cohen, 2001). Even though Zhang and Elder (2011) use an established test, the researchers decide not to use its established rating criteria, believing that using such a scale would constrain
the raters’ natural evaluations. They used a rating scale of 1 (very poor) to 5 (excellent), with no additional information provided. This approach limits the interpretation of results because actual test rating would employ the pre-established rating scale.

Most rater research acknowledges the importance of rater training either by using previously trained raters or by giving novice raters some amount of training. However, training is often minimal and certainly not uniform across studies. In some studies of native speaker versus non-native speaker raters, rater training means a brief meeting with raters to explain the rating procedures and answer questions (Barnwell, 1989b; Kim, 2009b). Studies where raters experience only brief training sessions referred to the teaching or language experience of the raters as part of their training (Kang, 2008; Kim, 2009b; Zhang & Elder, 2011), though such experience rarely includes education in Language Testing theory. In other instances, training was previously provided by established, large-scale testing programs (Barnwell, 1989b; Eckes, 2008) where raters undergo intense training programs and meet certification requirements. In these cases, training typically lasts days or weeks, involving detailed explanation of evaluation criteria, reviews of sample productions, and final qualifying exams for raters. These studies value the role of the training process in maintaining high inter-rater reliability statistics, so that the ratings produced were consistent (Eckes, 2008). Efforts were made to minimize individual variation between raters through training, so that raters could be interchangeable (Eckes, 2008). Because the approach toward rater training is so different among the studies on rater behavior, it is difficult to compare the results and conclusions of the studies directly. In all practicality, the more high-stakes the exam, the more training, quality control checks, and retraining the raters should undergo. Large-scale testing programs always use trained raters; therefore, research on similar testing circumstances should include trained raters.
In summary, many of the Language Testing research studies that are most often cited for showing no difference between the abilities of native and non-native speaker raters to evaluate language have notable limitations. The examinees that are being rated are typically learners, so they have low to mid range abilities. Raters are asked to give language proficiency ratings (which are generally holistic) on short samples that sometimes do not have a variety of language tasks. In many cases, the raters had a very limited training or no training at all. Also, they often did not use an established set of rating criteria, but rather gave impressionistic evaluations on a numeric scale that had no explanation. A lack of training, explicit scale, and normalizing for raters is incongruent with the practices of most large-scale Language Testing organizations, so it is difficult to determine whether these results are applicable. Despite these weaknesses, as weaknesses are inherent in all research, the number of studies on the subject of native and non-native raters continues to increase. This additional research will help support the emergence of patterns among the results, informing both Language Testing and SLA practices.

2.8 Contributions of current study

The operational testing context available in the US Federal Government offers a context that can address many of the weaknesses in the previous studies, adding to the overall body of research addressing native and non-native speaker differences, the effect of rater characteristics, and the construct of speaking proficiency exams. US Federal Government testing programs conduct thousands of speaking tests annually, with examinees ranging from low-level language learners to highly articulate and precise speakers (both native and non-native). Raters generally evaluate extended samples of speech to reliably establish a rating based on consistent ability. In order to perform this task, the raters must successfully complete extensive training and
certification requirements, and regularly undergo quality control checks. This detailed training includes an in-depth understanding of the ILR Skill Level Descriptions and the protocols for conducting such examinations. The current study intends to add to the present body of research on native speakers in the general linguistics, SLA, and Language Testing fields by offering evidence from a unique context, the US Federal Government. The use of highly proficient non-native speaker raters will provide new data that will help define what a native speaker is (or is not), reveal whether or not there are significant competence differences between highly proficiency native and non-native speakers (at least in terms of evaluating speaking ability), and determine if it is appropriate to use non-native speakers as raters.
Chapter 3: Speaking Proficiency Testing at the FBI

3.1 Overview

The United States Federal Government has been conducting language proficiency testing since the mid 1950s, longer than any other language testing organization that is currently in existence. In the process, government testers have collaborated extensively with each other about the language testing issues that underlie their work. Nonetheless, because of the government’s internal operational focus and its corresponding lack of resources to support publishing, very little of the information and knowledge that the government has generated has been made known to academia and industry. As a result, academia and industry have lacked a clear understanding of the rationale behind government testing practices and, largely as a result, have criticized government testing organizations’ practices and their interpretations of the proficiency concept. A better understanding of the context in which the government operates and how government testing organizations apply testing concepts will help illuminate the rationale behind government testing practices, as well as contribute to academia and industry’s collaboration over testing theory.

3.2 The origins of proficiency testing

In the 1970s, the concept of language proficiency was already well established in the United States Federal Government. Outside the government, the idea of language proficiency in academia emerged with the concept of communicative competence. Hymes (1972) stated that a complete understanding of a native speaker's language use must extend beyond Chomsky's idea of linguistic competence to include communicative competence. Communicative competence
included not only a native speaker's intuitive knowledge of all the linguistic features and rules that govern them within a language, but also incorporated competence in the social rules of the language, including communicative competence, register, and other conventions. Savignon (1972) claimed that communicative competence involves both linguistic ability, which is the ability to function in a communicative setting, and paralinguistic ability, which was the ability to adapt to input. These authors along with Schulz and Bart (1975) emphasized that communication was of utmost importance, with form secondary. The three areas that were of focus in communicative competence were function, content, and accuracy (Higgs & Clifford, 1982). The definition of “proficiency” continued to be discussed and improved on in academic contexts.

Bachman and Palmer (1982) introduced the notion of communicative proficiency and suggested that it should be used to refer to language ability. Communicative proficiency was described as an integrated skill, involving language competencies such as grammar, discourse, elocution, and sociolinguistics as displayed though the different skill modalities (receptive, productive, oral, and visual).

United States Federal Government proficiency testing fleshed out these models of proficiency in the Oral Proficiency Interview (OPI). The ILR concept of proficiency borrowed from Hymes’ 1972 communicative competence model, but added Higgs and Clifford’s concept of utilizing native speaker intuitive knowledge of both the linguistic and social rules appropriate to a particular linguistic environment (Higgs & Clifford, 1982). As a result, the native speaker became central to the concept of proficiency in the government. The US Federal Government needed to evaluate not only effective communication, but also whether or not personnel could be fully integrated into a foreign language context. Answering that question required evaluation of a person’s content, function, and accuracy in the target language. For this reason, the US Federal
Government adopted proficiency testing at all levels, including the sophisticated and eloquent levels of language proficiency that many native speakers do not attain.

Lowe (1983) defined US Federal Government language proficiency as “the global rating of general language ability over a wide range of functions and topics at any given level.” The ILR Skill Level Descriptions for Proficiency are meant to address a person’s functional linguistic ability regardless of how the language was acquired. The ILR refers to “native speakers” in order to state that an individual who achieves a Level 5 has the language abilities functionally equivalent to that which a native speaker is thought to possess. Specifically, the ILR Level 5 is a “highly articulate well-educated native speaker,” which is defined by demonstrated ability rather than whether the individual is actually a native born speaker. The Descriptions describe an examinee’s ability to master increasingly difficult linguistic functions, in increasingly broad contexts, with increasing expectations for accuracy. The ILR Levels are not meant to be all-inclusive descriptions, but rather a description of the salient features of the level, detailing what examinees are able to accomplish as well as their linguistic limitations (Lowe, 1998).

The discussions of communicative competence set the stage for academia to explore what the United States Federal Government had been doing in the realm of language proficiency. At the same time, US Federal Government agencies reached out to academia for additional foreign language resources. In 1978, the Foreign Service Institute held familiarization workshops on the oral interview with members of the academic community. These workshops sparked interest in the US Federal Government testing system, and later that year the Educational Testing Service received a grant from Congress to explore the Interagency Language Roundtable system in greater depth and determine whether or not it would be applicable outside the government context. This project was known as “the common metric project” and concluded that the ILR
SLDs were not applicable outside the US Federal Government in their form at that time, because, the project found, the ILR does not discriminate well enough among the lower levels of beginning language abilities.

3.3 Proficiency in the US Federal Government

Lowe, Child and Mackey defined proficiency in US Federal Government contexts as “the functional ability to accomplish communication tasks through unrehearsed language” (Lowe, Child, & Mackey, 2000, 9). Proficiency is concerned with a broad functional ability over a wide range of content, meaning a person's ability to perform linguistic functions or oral communication tasks in a manner appropriate for the country or culture where the language is spoken. Proficiency concerns itself with both what an examinee can do with the language as well as how well the examinee can do it: proficiency measures both the quantity and quality of tasks performed (Lowe et al., 2000).

US Federal Government agencies that have language training schools use proficiency testing the end of most courses. In the United States Federal Government, proficiency testing is also used for making decisions on hiring and promotion as well as incentive pay and assignments abroad. Although proficiency testing may not seem the most appropriate approach in each of these contests, US Federal Government agencies need linguists who are flexible, who can be placed in a variety of situations and circumstances and function appropriately in the language. US Federal Government agencies also need a rating that can be useful across all examinee types and for a variety of purposes (FBI, 2009).

The Speaking Proficiency Test (SPT) and the Oral Proficiency Interview (OPI), another interview-style test administered in the US Federal Government, are not an achievement tests—
i.e., they are not measures of a specific set of knowledge. The questions posed in these oral assessments are not designed to cover course content, but rather a general ability to function at a particular level. Nor are the SPT and OPI performance tests, which are used to indicate a person's ability to accomplish tasks related to specific professional contexts. Although US Federal Government linguists are expected to use their language proficiency in the course of performing specific linguistic tasks, such as audio monitoring, translation, and interpretation, the SPT and OPI are designed to measure in overall, broad terms the examinee’s ability to speak. Proficiency and performance blend together in instances where the variety of tasks used to determine proficiency are geared towards particular job-specific tasks. Performance defined in this way parallels academia’s concept of language for specific purposes (FBI, 2009).

3.4 Rating Criteria: The Interagency Language Roundtable Skill Level Descriptions

The original purpose of the ILR scale was to test adult Americans born in the United States who were learning a second language, most of whom came from the Foreign Service community. Eventually, the ILR was used to test other federal employees and also came to be used by the academic and for-profit sectors. In today’s US Federal Government, the ILR scale is used to assess all types of language speakers – no matter where their language was acquired here or abroad. Agencies across the US Federal Government use the ILR Skill Level Descriptions to describe the language proficiency of native speakers as well as language learners. Because of the functional nature of the ILR scale, this practice has worked well for the government (FBI, 2009).

The Interagency Language Roundtable Skill Level Descriptions provide a scale that is functional in nature. The primary purpose of the ILR Skill Level Descriptions is to describe how language professionals use their language proficiency, ranging from Level 0, No Proficiency,
through Level 5, Functionally Native Proficiency. Although Level 5’s Description uses the word “native,” Functionally Native Proficiency is not necessarily the same thing as a native-born speaker (ILR, 1985). As established previously, not all native speakers are homogenous (Hamilton et al., 1993); being a native-born speaker in itself is insufficient to earn a Level 5 (Brooks & Brau, 2007).

The US Federal Government employs professionals with highly developed language proficiencies in order to fulfill specialized mission demands. These personnel must be proficient at levels beyond what would be expected by the average speaker of the language. For example, diplomats working for the Department of State need to be able to negotiate policy effectively using the appropriate cultural and sociolinguistic strategies. For another example, covert operatives working for an organization such as the CIA need to be linguistically imperceptible when working undercover – operatives need to be able to blend into their environments.

Effectively tailoring one’s language to a range of audiences, maintaining the appropriate tone during a conversation, and presenting an unconventional, complex idea in a manner where nuance or meaning could adversely affect its acceptance, are all tasks that are carried out by US Federal Government positions. As a consequence, the ILR scale needs to carefully distinguish general professional speaking proficiency from the highest levels of speaking proficiency. For most language learning and professional purposes, reaching ILR Level 3 (General Professional Proficiency) is more than sufficient. There are positions, however, that require greater depth and breadth, and the ILR scale therefore distinguishes between the basic, professional level that begins at Level 3 through the advanced professional Level 4 to Level 5, where the range of language associated with the “highly articulate, well educated native speaker” is expected (FBI, 2009)
<table>
<thead>
<tr>
<th>Base Level Summary Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 No Proficiency</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2 Limited Working Proficiency</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4 Advanced Professional Proficiency</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Figure 1. ILR Skill Level Descriptions Summary Statements (FBI, 2009)

---

2 Although FBI does not ordinarily conduct face-to-face tests, as discussed previously, various government agencies have conducted research that validates the comparability between face-to-face and over-the-phone tests (Fischer, 2006).
Each of the ILR Skill Level Descriptions is comprised of holistic statements reflecting a speaker's overall ability. They contain information about both the language tasks that an examinee is able to perform, the quality of that performance and the topic areas that speakers at that level can typically handle. This scale is intended to be an ordinal scale because its ranges do not have a constant interval, neither are they in a mathematical ratio (Lowe, 2012). The summary statements from each of the ILR Levels are shown in Figure 1, while the entire Skill Level Descriptions for Proficiency – Speaking are in Appendix A.

Each level represents a range of abilities, delimited by the Descriptions themselves. This ability must be demonstrated consistently throughout the test, not just on one or two occasions. In order for an examinee to receive a particular rating, the examinee must demonstrate every single part of the Skill Level Description that correlates to the rating. The SLDs describe a non-compensatory core for language proficiency, meaning that if an examinee does not display consistent ability to demonstrate minimum features of that level of language proficiency, the examinee cannot earn that level. The non-compensatory nature of this scale serves two purposes: it reduces the subjectivity in raters’ decisions to assign one level or another and delivers a rating to a government agency that represents an examinee's ability to consistently perform all language functions at that level with the necessary accuracy (FBI, 2009). The aggregate nature of language learning means that many language learners find moving from Level 0 to Level 1 of the ILR scale less difficult than reaching each successive higher level. Some learners, and even native speakers, seem to fossilize; they are not able to move past a certain level (FBI, 2009).

ILR Level 2 and 3 are most critical levels in the US Federal Government context. Level 2, Limited Working Proficiency, is characterized by the ability to satisfy routine social demands and limited work requirements. Level 2 speakers can execute a range of functions such as
conducting routine work-related interactions, and they are able to participate in personal
interactions with facility. Level 3 speakers, on the other hand, show greater depth and breadth of
ability. They are able to speak the language with sufficient structural accuracy and vocabulary to
participate effectively in most formal and informal conversations in practical, social and
professional topics (FBI, 2009).

Level 2 and Level 3 speakers differ in their control of organization. Where Level 2
speakers use minimal coherence, Level 3 speakers use cohesive discourse, which means they are
able to clarify, support opinions, justify decisions and deliver extended monologues. Level 2
speakers control basic structures but often show weakness, whereas Level 3 speakers effectively
use language structures to convey meaning accurately. There are also notable differences in
vocabulary between Levels 2 and 3. Level 2 speakers use vocabulary that is appropriate in high-
frequency settings, while Level 3 speakers use the language clearly and relatively naturally to
elaborate concepts freely and make ideas easily understandable. Level 2 speakers speak with
confidence but not facility; a speaker at Level 3 speaks readily and fills in pauses suitably.
Differences are also apparent between these levels in terms of pronunciation: Level 2 speakers
often miscommunicate due to pronunciation, but Level 3 speakers’ errors virtually never
interfere with understanding. Differences in levels of socio-cultural awareness can also be noted
between Level 2 and Level 3 speakers, with Level 2 speakers comfortable in social situations,
and Level 3 speakers effective across social, practical and professional topics (FBI, 2009).

At Level 4, language is no longer merely just a vehicle for conveying information
between two people, but is also a tool to achieve certain, higher-level purposes with an
interlocutor. The speaker is able to tailor, persuade, counsel, negotiate, and construct his
responses with strategies appropriate to achieve his goal in communication. At Level 4 there are
few flaws in the language and only minor mistakes in his elocution, while the examinee is able to employ an arsenal of linguistic tools to enhance his message, including rhetorical speech devices and collocations. Level 4 speakers are rare. Advanced cognitive ability and knowledge are prerequisites for Level 4; however, many speakers who are highly intelligent cannot produce Level 4 speech consistently across a wide range of subjects. Additionally, there is no specific piece of knowledge that a Level 4 speaker must have or is asked to produce during a test. Level 5 speakers do everything Level 4 speakers do, but with such control that they do not make any noticeable mistakes or slips. Additionally, Level 5 speakers have more creativity with the language and are able to cover a wider range of topics (FBI, 2009).

The ILR Skill Level Descriptions were designed for government stakeholders, including employees, supervisors, and language students. The ILRs are comparative, in that they describe both what the speaker can and cannot do at a particular level, and thereby provide both the ceiling and the floor for the level. They are also comparative in that the Descriptions carry the same themes from one level to the next, so that an examinee at one level can be compared to the level above and below. The Descriptions themselves operate by the principles of gestalt and outlining. They give the essence of what the examinee can do at a particular level, but are not meant to be exhaustive. The center of the description informs the periphery. When the gestalt revealed by the examinee is incomplete, the rating is lowered. Each ILR Description begins with an overall characterization, followed by qualifying statements, the concluding statement, and then further exemplification. The descriptions contain thresholds of the level (Lowe, 1998).

However, each description shows worst-case, average case, and best-case perspectives for each level. These perspectives are based on findings from working with thousands and thousands of language learners within the government context, not upon constructs of language
learning. They were developed organically from examples, not mechanically from theory. The goal of the constructors of the Descriptions was to provide bracketing and interpolation, providing evidence of a floor and ceiling for each level. In the Skill Level Descriptions for Speaking Proficiency, the majority of the ranges are well bounded, with the exception of the Level 3/Level 3+ boundary (Lowe, 2001).

Lowe (2001) describes why perhaps the Speaking Skill Level Description is easier to use than other similar Skill Level Descriptions, such as Listening, Reading, and Writing. The Speaking Skill Level Descriptions have clearer boundaries between most levels than the other skills do. Historically, the Speaking Skill Level Descriptions have been the most needed of all of the four major proficiency skills. Speaking testing has been both administratively accessible and widely used among US Federal Government agencies. Because the Speaking Skill Level Description has been measured more often than the other Skill Level Descriptions, it has been revised more frequently than the other Skill Level Descriptions. The Speaking Skill Level Description is the most readily usable also because it is the most experiential of the Skill Level Descriptions. It is empirical in that it shows worst-case, average-case, and best-case boundaries between almost every level. In contrast, the Skill Level Descriptions for Listening, Reading, and Writing have anywhere between two and six unbounded levels.

3.5 Proficiency testing instruments

The FBI’s SPT measures a functional ability in language, regardless of how, where, and under what conditions that language was learned or acquired. Raters measure an examinee’s ability to appropriately accomplish real-world communication tasks in a variety of target-culture contexts, including general conversation, work-related tasks, and complex, sophisticated tasks
both within and outside of the individual’s expertise (FBI, 2009). Testers pose questions that require the examinee to interweave language features to complete the task posed. As a criterion-referenced test, the SPT rating reflects an examinee’s ability to function based on pre-determined criteria, or objectives, and not reflective of relative ability of examinees. The criteria used for the SPT are the ILR Skill Level Descriptions. As a holistically rated test, the SPT measures what the examinee does well, and whether or not it is done consistently, rather than only what the examinee lacks in speaking ability. SPT scoring involves an evaluation of a wide array of linguistic functions elicited during the course of the SPT. The rater must bear in mind all aspects of an examinee’s performance during the test and during the rating process. The result is a global rating of language ability based on the gestalt of the performance (Lowe, 1983).

The SPT and the OPI are generally administered in three phases. The first phase is the warm-up. In the warm-up, the testers work to make the examinee feel at ease and get into the flow of using the language. Additionally, the testers glean information about the examinee's background and familiar topics from the examinee. Lastly, the testers sense a hypothetical level of the examinee's ability. This hypothetical level becomes the foundation for the second phase of the test: the core phase (FBI, 2009).

During the core phase, the testers work to test their hypothetical level by eliciting responses aimed at that level of ability; these elicitations are called level checks. Level checks help to confirm the examinee's ability to perform at their level of comfort. Interspersed among the level checks are probes, or elicitations aimed at raising the examinee's level of ability to the next ILR level. Testers alternate between level checks and probes to find a level at which the examinee can perform consistently (also referred to as the floor), adjusting the overall level of the test when necessary. The probes help to establish the level the examinee cannot yet fully
attain (also referred to as the ceiling). Elicitations may include a variety of question types, statements made by the tester, situations, and other activities. Once the testers are sure that they have gathered enough information to be able to assign an accurate rating to the speech sample, they moved to the wind-down portion of the test. In the wind-down, the testers lower the level of the test to a comfortable level so the examinee is able to leave the test with a feeling of accomplishment (FBI, 2009).

The SPT as used by the FBI is a revised version of the US Federal Government’s original Oral Proficiency Interview. The FBI, CIA, and DLI adopted the SPT. Use of the SPT methodology brought noticeable improvements to government oral proficiency testing by returning to its core philosophy of eliciting language in a natural, conversational style and gave flexibility to testers so that the test appeared more natural because it tapped more authentic language. For example, instead of depending on scripted questions written on cards, testers were trained to develop questions through the course of the test by combining tasks appropriate for the level with relevant topics as a test progressed. This allowed for a more natural flow of conversation throughout the course of the test, with one topic developing into another topic, instead of abrupt shifts in the test making it seem less like conversation and more like an examination. Nonetheless, testers still retained sets of prepared questions that they could weave into the test as appropriate. Another change in the SPT’s methodology was that testers no longer had to painstakingly work from Level 0 up to the appropriate level based on the hypothetical level gleaned from the warm-up. Another advance that was made in the test is that some of the tasks that were knowledge-based, such as the completion of proverbs, were eliminated.

One drawback of the SPT methodology is that it requires testers to be highly trained and adept at administering the test to ensure that a ratable sample of language is acquired. Although
the test is designed to elicit a more authentic sample of speech, there is more danger that the testers do not sample the speech appropriately because they may not be able form the appropriate tasks and administer the elicitations adequately. Due to these drawbacks, any testing programs using the SPT methodology must ensure that its testers are well trained and constantly supervised to ensure rater reliability and test validity (Herzog, 2003).

At the turn of the 21st Century, testing at some US Federal Government agencies was growing exponentially. Agencies that have language schools typically depended on their instructors to be oral proficiency testers as an ancillary duty. This sometimes meant that there were large numbers of personnel working as testers and yet little time or priority given to tester training and reliability. This lack of control over the testers endangered the administration of the SPT methodology. In 2000, the Defense Language Institute (DLI) was in such a position, with approximately 300 instructors serving as oral proficiency testers. In order to better control the methodology and ensure better uniformity in test administration, DLI reverted back to the OPI methodology. This more regimented methodology helped DLI to have reliable test ratings despite a large number of testers. The OPI 2000 retained some of the advances made in the SPT’s methodology, such as the concept of using the hypothetical level from the warm-up to begin asking questions during the core of the test. The FBI and the CIA had smaller testing programs, with 100 to 200 testers per program, and opted to retain the SPT methodology (Herzog, 2003).
Chapter 4: Methodology and Research Design

4.1 Overview

This study compares scores assigned by groups of raters of the FBI Speaking Proficiency Test (SPT). The participants were thirty trained and certified testers/raters in eight different languages: English, Arabic, Farsi, French, German, Mandarin Chinese, Spanish, and Vietnamese. All of the raters had good English proficiency, receiving English speaking proficiency ratings of ILR Level 2+ through ILR Level 5. The rater participants are asked to listen and then assign a rating for twenty-five complete English SPTs. The SPTs were selected to represent a variety of speaker abilities, ranging from ILR Level 2 to 5, and to also include examinees who are both native speakers and non-native speakers of English. To prevent rater bias for or against one examinee type, about half of the examinees were native speakers and half were non-native speakers. Participants listened to the tests, rated them, and then completed individual reports for each exam, noting their overall rating, linguistic category ratings (ratings of functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness) and justifications for the ratings given.

For analysis, the raters were initially divided into two groups, English native speaker raters (n=6) and non-native speaker English raters (n=24). Since the definition of a native speaker can be based on the following criteria a) speaking the language very well (language proficiency) and b) the first language acquired (native language), the non-native speakers were further subdivided in to groups based on those principles. Non-native speakers were divided into groups according to their a) English speaking proficiency level (creating three additional groups)
and b) their first/native language (creating seven additional groups). These new groupings were then compared to the native speaker raters.

Means, medians, and standard deviations of final ratings of all exams were compared between the native and non-native speaker groups to address the first research question. Analysis of Variance (ANOVA) was used to determine if the mean ratings of the two groups were significantly different. Then NS and NNS ratings of exams were at each rated exam level (Level 2/2+, Level 3/3+, and Level 4/4+/5 exams). Next, the raters were grouped according to their own rated English proficiency, for the second research question, and according to their first languages, to address the third research question. Finally, the fourth research question (how the participant raters evaluated the linguistic categories) was addressed by comparing those more specific rating categories across the same groups as above, using multivariate analysis of variance (MANOVAs).

4.2 Participants

The FBI maintains a cadre of almost 200 speaking testers who administer oral proficiency evaluations in approximately 70 languages, conducting a total of more than 3,000 speaking tests annually. The Speaking Proficiency Tests (SPTs) are conducted via telephone, and typically two testers work as a team to administer the test to the examinee. Both testers administer and rate the test, giving both holistic final ratings and linguistic category ratings for each examinee. Depending on operational need, raters may be asked to administer and rate up to 30 tests a month. Some raters are US Federal Government employees and some work for the government as contractors. Most of those who are government employees are in the position of Language Analyst. Language Analysts complete language transfer tasks, such as verbatim
translation, summary translation, and interpretation. For these people, testing duties are ancillary, and they are not compensated in addition to their regular salary, as testing tasks are performed during regular work hours. Being a rater, however, makes them more competitive for promotion. More than half of the testers are Contract Testers, who generally do not perform other services for the FBI other than administering and evaluating SPTs. They are paid a flat fee for each test they rate (FBI, 2009).

4.2.1 Rater qualifications and training

Becoming a speaking proficiency tester is a rigorous process. Initially, candidates for the position must meet qualification requirements, which include completing an application package and interview, as well as meeting speaking proficiency, native speaker, availability, and background requirements. Candidates for the English tester position must be native speakers of English and take an English speaking proficiency test and receive an ILR rating of Level 5 in order to qualify. Candidates for all other languages must be rated ILR Level 5 in their native language and a Level 2+ in English to qualify. Qualified candidates are invited to attend a two-week training workshop in Washington, DC (FBI, 2009).

The first week of the workshop entails eight hours a day of theoretical background in proficiency testing, speech elicitation techniques and rating theory pertaining to testing in their languages. The workshop curriculum focuses on explaining the ILR Skill Level Descriptions both as the construct for the exam and as the protocol for rating. Much time is spent discussing the differences between the levels and explicating what each level means, including disambiguating phrases and understanding the philosophy of the ILR.
The ILR Skill Level Descriptions are a single, uniform set of criteria used by raters of all languages; they are stated in English, and are not translated into other languages. Therefore, the raters of all languages are trained to interpret the same set of criteria in order to have a common conceptualization of the criteria, regardless of whether or not they are a native speaker, their English speaking proficiency level, or their first language. Workshop activities include lectures, practical activities, extensive readings, review and evaluation of over 30 SPTs in English and their test language, and conducting practice SPTs with volunteers. Each night candidates are required to complete two to three hours of homework related to the material they learned that day. During the first week, trainees review approximately ten complete tests, both in English and in their own native language. In the second week, candidates practice their new skills, administering and rating approximately twenty additional speaking proficiency tests in their language (FBI, 2009).

The workshop is completed with a three-part final exam: a theoretical test, a task in which the candidate must produce a set of questions appropriate to eliciting language at various levels, and a rating test, where the candidate must rate and complete reports for two exams. Successful completion of these three parts of the final exam in addition to a positive recommendation from a master tester, who observed the practical part of the workshop, grants a trainee provisional certification as a tester and rater. Tester/rater candidates are then paired with master testers to complete ten additional tests. After successful completion of ten tests, raters are then considered certified. Ten to twenty percent of SPT workshop participants do not meet the requirements set forth in the workshop and training period and are discontinued as testers. Only after successfully completing all parts of SPT Tester training and evaluation can they begin to evaluate real FBI examinees.
The common understanding of the ILR Skill Level Descriptions among and across test languages acts as a control variable for the research design, limiting individual biases in the raters. The extensive training that the raters undergo supports increased rater reliability and reduces the subjectivity of the ratings. The training puts all of the raters on an equal playing field in their approach to the rating sample, even though English is the first or strongest language of the non-native speakers. It shows whether or not rater groups can be comparable with the appropriate training, or whether there are differences in the evaluation of the raters that cannot be overcome despite training.

4.2.2 Participants’ background

To address the first research question which focuses on the difference between native and non-native speaker evaluations, the participant raters were divided into two rater groups: native speakers of English (NS) and non-native speakers (NNS) of English. The FBI’s operational definition of a native speaker is someone who is raised, including living and being educated) in a society where the language is spoken as well as educated in that language though completion of secondary school. This definition is not meant to refute an individual’s belief in being a native speaker. The native speaker definition in the FBI context is used to select applicants who are more likely to receive ILR Level 5 in Speaking, saving the agency resources in the hiring process.

---

3 The purpose of this definition is only for research purposes and selection of FBI SPT Testers. An examinee’s linguistic history is not considered in the language evaluation processes or hiring of any other FBI positions.
NNSs are considered to be all other speakers of the language, including heritage speakers, language learners, and many people who do not neatly fit into one of the aforementioned groups. The FBI recognizes that a person’s whether or not a person is a native speaker is not a reliable predictor of speaking proficiency, but it has found that NNSs are unlikely to receive ILR Level 5.

The NS raters in this study are qualified FBI English raters, and are native English speakers rated ILR Level 5 in Speaking English. The NNS raters are current SPT raters of Arabic, Farsi, French, German, Spanish, Mandarin Chinese, and Vietnamese and therefore are native speakers, rated ILR Level 5, in those languages, respectively. SPT raters of foreign languages have all been rated at least ILR 2+ in English speaking proficiency; the English speaking proficiency levels of the NNS raters participating in this study range from ILR Level 2+ to 4+. In fact, the majority of English native speakers who take the test (including those who are college educated) are rated in the 2+ to 3+ range (Brooks & Brau, 2007). So, the NNS are highly-proficient English speakers, with proficiencies equaling or in some cases, even exceeding that of the typical English native speaker. All of the participants in the study have been raters for at least one year, and successfully completed quality reviews of their rating ability in the previous year, indicating that they are reliable raters in their respective native languages. Raters were solicited from the pool of current FBI raters, and were certified to give tests in their native languages. Raters received financial compensation equal to their normal evaluation rate, $80 per test.

4 Depending on agency needs, SPT Tester applicants who do not meet these requirements may be permitted to attend tester training if they are ILR 5s in Speaking.
4.3 Sample: Speaking Proficiency Tests (SPTs)

The exams that the raters evaluated were Speaking Proficiency Tests (SPTs). SPTs, as the name indicates, are intended to assess an examinee’s speaking proficiency, or their holistic, functional ability to use the language, regardless of how the test takers acquired their language (FBI, 2009). The US Federal Government’s approach to language proficiency evolved through the latter part of the 20th Century, in part influencing and in part being influenced by the concept of proficiency in academia (Lowe & Stansfield, 1988). This approach to language testing within the US Federal Government has not been largely documented outside the internal government manuals and procedural documents that are not publicly available. The theory behind the Government’s approach is central to understanding its particular construct of proficiency and the interpretation of government proficiency ratings.

The speaking samples that the raters in this study were authentic English SPTs administered at the FBI (n=25). The English SPTs selected to be rated in the study represent the range of typical FBI examinees. The distribution of ratings of the SPT exams selected for this study is listed in Table 1. The tests included thirteen native English speaker examinees and twelve non-native English speaker examinees. The examinees in each group have varying proficiency levels, ranging from ILR Level 2+ to ILR Level 5. The native speaker examinees in the sample range from ILR Level 3 to ILR Level 5, with two or three examinees at each level. Similarly, two or three non-native speaker SPTs are included in the sample at each level, representing ILR Level 2+ to ILR Level 4. No native speaker Level 2+ examinees were included in the sample because they are rare among FBI’s corpus of tests and unrepresentative of typical FBI examinee results. Likewise, non-native speaker ILR Level 4+ or ILR Level 5 examinees were excluded because they rarely occur and could not be located in the FBI SPT test corpus.
Table 1. English SPT Exam Distribution

<table>
<thead>
<tr>
<th>ILR Level</th>
<th>NS</th>
<th>NNS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4+</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3+</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>2+</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>12</td>
<td>25</td>
</tr>
</tbody>
</table>

Each of the recorded tests selected for the sample was initially selected based on the rating given by the original two raters. Each test was then reviewed to verify the original rating, determine if the examinee is a native or non-native speaker of English and check for issues with the recordings. The samples ranged in length from 35 minutes to 50 minutes.

During data collection, the rater participants received a set of recordings of the English SPTs to rate. The order of the exams was randomized five times, to create sets with five different orderings. Raters were randomly assigned to one of the five groups to control for an ordering effect. The raters were then asked to rate the tests as they normally would, listening to the exam, reviewing the rating criteria and completing a form including final ratings, linguistic subcategory ratings, and justifications for their ratings. Raters were given a total of two months in which they had to rate all 25 SPTs. Raters were instructed not to rewind, stop or review the test.
which are not allowed in regular operational test rating or test reviews). Repetition was prohibited because repeated listening causes raters to lose the holistic nature of the final rating by overanalyzing individual errors.

Although the research methodology tried to replicate the normal operational testing process as much as possible, there were a few deviations that were unavoidable in order to properly conduct the research. The main deviation is that the raters typically act as both testers and raters simultaneously, both giving and rating the test at the same time in tandem with a co-tester. Typically, there are two tester/raters on the phone in a three-way conversation with the examinee. As a result, the raters in this study had a much lower cognitive load, since they only had to rate the test and not administer the test, which would include developing and posing appropriate elicitations and interacting with the examinee and co-tester. As a result, the raters in the study might have been able to pay more attention to the sample they are evaluating, notice more errors and potentially assign lower ratings. All of the raters in the study have reviewed and rated previously-administered tests in the same manner as this study during the annual Quality Assurance process, but most raters only rate tests in this manner two to five times per year. Since this study deals with rating only, and not test administration, the insights gained can help the selection of raters that can conduct reviews, and not new test administrators.

It was normal for participants to not see the examinees in the tests because, at the FBI, all SPTs are administered via telephone conference call. Since testers administer all SPTs telephonically, they generally give tests from their homes or offices, in the same locations as where they rated the exams for the study. In typical test administration, both testers participate in administering the test and rating the test.
4.4 The ratings

The raters evaluated the sample SPTs according to the FBI rating protocol. Rating activity began during the test itself, as the raters observed the questions given by the testers and took careful notes about the examinee’s performance in response to the testers’ directed elicitations. In SPT elicitation practice, the testers try to determine the highest ILR level at which the examinee can function consistently. The testers crafted questions both at the level at which the examinee functioned appropriately (level-checks) and also questions at the next higher level (probes) to challenge the examinee to function at a more complex level. When the test began, testers formulated an initial hypothesis of the examinee’s working level proficiency; they then continually evaluated and modified this hypothesis, based on the results of the level-checks and probes. The raters examined both the testers’ choice of questions as well as the examinee’s ability to respond appropriately to those questions to determine a working level for the test, and eventually a final rating.

After the raters finished listening to each SPT, they began the post-test evaluation process. The raters reviewed their notes to find strengths and weaknesses in the examinee’s performance. The raters assigned a holistic final rating, as well as linguistic category ratings in functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness. The linguistic category ratings were reported as main levels (i.e., 1, 2, 3, 4 or 5), whereas the final rating could be a main level ILR rating or a plus level rating.

The ILR rating system is holistic and non-compensatory. The ILR Skill Level Descriptions (Appendix A) details the functional ability required to receive a given rating. As a result, an examinee could not receive any particular ILR rating unless each criterion described in the rating was met. For example, although an examinee may have met almost all the criteria
listed in ILR Level 3, if even one of these criteria was not met, the resulting rating can be no higher than an ILR Level 2+. FBI speaking test raters first determined the linguistic category ratings through the use of the Speech Evaluation Guide, shown in Appendix B, a rating tool that organizes the ILR Skill Level Description criteria by linguistic category so that they were more easily comparable across levels. Once each of the linguistic category ratings was assigned, the lowest of these ratings became the base level for the final rating. Typically, ratings across the linguistic categories would range across two or three levels. In an uneven performance, a native speaker examinee could be rated an ILR Level 2 in vocabulary, but a Level 4 in pronunciation (or vice versa for a non-native examinee). Nevertheless, the base rating would be ILR 2 if it were the lowest ILR Level rating assigned. In other words, linguistic category ratings were never averaged to arrive at a final rating. The raters then reviewed the full ILR Skill Level Description for the base level (for example, ILR Level 2). If the examinee fully met the Level 2 criteria, the rater would review the ILR Level 2+ Description. If the examinee met all the criteria of the Level 2+, than that became the final rating. If not, the final rating was ILR Level 2.

Ratings were justified through comments included on the Individual Tester Report (ITR), which was completed by each tester after the final rating was determined. (An example ITR is included as Appendix C.) These reports are written in English, no matter the language in which the test was administered. In addition to basic administrative information, the report includes all of the assigned ratings, justifications for the ratings, and an overview of the content of the test. Raters were required to give an overall, general justification for the final rating, which included the features of the examinee’s performance that were the most salient for determining the rating. Then the raters gave specific justifications to support each of their linguistic category ratings. Raters also listed the topics that were covered during the course of the test to assure that a range
of topics were introduced and that no controversial, personal, or sensitive topics (e.g. abortion, religion, homosexuality) that would upset the examinee or give the appearance of tester bias occurred. The two role-play situations that the testers selected for the examinee were listed, including the task that the situation required and comments about the examinee’s performance on the situations. Raters also documented how and where linguistic breakdowns occurred during the test. The raters detailed two examples of probes that evidenced the examinee’s lack of ability to meet the linguistic demands of the next higher ILR level. In this section, the rater paraphrased the question, determined the task and ILR level of the question, and detailed how the examinee did not meet the expected requirements according to the ILR criteria. Finally, the rater reports included an area for any noteworthy comments not directly pertaining to the examinee’s language ability, such as that the examinee had an illness that may have impacted the test performance.

4.5 Data

Once reports were completed, they were submitted by email. The information was then compiled into a spreadsheet, including information about the raters, such as their native speaker status, first language, and English proficiency level. The ratings that were used for quantitative analyses included the final holistic ILR rating and linguistic category ratings for functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness.
Subsequent to the ratings, the participants provided a detailed justification for each assessment; however, the justifications were not analyzed for this study\(^5\).

### 4.6 Analysis of the data

In order to measure the consistency of ratings and homogeneity of rater groups, inter-rater reliability of each of the rater groups was measured separately using Krippendorff’s alpha. Raters were divided by whether or not they were native speakers (NS raters and NNS raters), English speaking proficiency (Level 2/2+ raters, Level 3/3+ raters, Level 4/4+ raters, and Level 5 raters) and first language (English, Arabic, Farsi, French, German, Mandarin Chinese, Spanish and Vietnamese raters). Krippendorff’s alpha was used because the rating scale was designed as an ordinal scale, and this analysis considered the proximity of differing ratings in its calculations. Krippendorff’s alpha was used to determine if the groups were able to meet the acceptable standard for rater reliability. An inter-rater reliability coefficient of 0.8 or higher is generally considered acceptable in most situations; a minimum of 0.7 is considered acceptable for many studies (Krippendorff, 2004; Lombard, Snyder-Duch, Bracken, 2003). The results show that rater groups did not always meet these thresholds.

\(^5\) Comments data collected in the reports were not included in the analysis for this particular study. Raters provided general comments on the overall performance, listing the features of the examinee’s ability that were most influential on the final rating. Raters then provided comments on each of the linguistic categories, detailing strengths and/or weaknesses in those particular areas. In these comments, the raters were advised not to simply copy the text from the ILR Skill Level Descriptions onto the report verbatim, but rather explain in detail how the examinee’s performance did or did not meet ILR criteria. Raters were also asked to indicate two elicitations used in the test where the examinee’s response did not meet the functional or linguistic expectations required for the intended level, known as breakdowns. Also, information about the topics included and the role-play situations was not used.
After reliability was examined, the ratings from the different rater groups were compared. As all three research questions examine group differences, the final test ratings were compared using Analysis of Variance (ANOVA) to determine if there was a significant difference between the ratings given by different groups of raters, as well as the effect sizes of any significant differences. Before any parametric tests such as ANOVAs were conducted, the prerequisite assumptions were assessed; the data were analyzed using descriptive statistics and histograms to check the data’s normality, independence, homogeneity of variance, and whether or not the data was interval in nature. When the assumptions were not met, statistics that made adjustments for the discrepancy in the data were reported, such as the Greenhouse-Geissler statistic, which corrected for violations of the assumption of sphericity relevant to MANOVAs.

Power analysis revealed that 30 subjects divided in two groups (NS and NNS raters) with 25 repetitions, an assumed effect size of \( r = 0.2 \), and an alpha of 0.05 had an estimated power of 1.00. An effect size of \( r = 0.2 \) was chosen because it was between a small (\( r = 0.1 \)) to medium (\( r = 0.3 \)) effect size, and it accounted for 4% of the variance. As there were so many factors contributing to variation in test ratings (examinee performance factors, task-related factors, etc.), an effect size at this level would indicate an influence of whether or not the rater is a native speaker on the examinee’s ratings. Results from the ANOVAs (test of significance, effect sizes, and power) are reported.

To test for potential differences between raters due to English proficiency, the raters were assigned to groups according to their English speaking proficiency. This rearrangement of raters resulted in four groups: ILR Level 5 raters (\( n = 6 \)), ILR Level 4/4+ raters (\( n = 4 \)), ILR Level 3/3+ raters (\( n = 14 \)), and ILR 2/2+ raters (\( n = 6 \)). ANOVAs were again used to determine differences between these four groups’ final exam ratings, after the appropriate assumptions were assessed.
Power analysis revealed that 30 subjects divided in four groups with 25 repetitions, an assumed effect size of $r = 0.2$, and an alpha of 0.05 had an estimated power of 0.98.

A third set of analyses on final exam ratings tested for differences due to raters’ first language. This division resulted in eight different rater groups: English ($n=6$), Arabic ($n=4$), Farsi ($n=3$), French ($n=3$), German ($n=3$), Mandarin ($n=4$), Spanish ($n=4$), and Vietnamese ($n=3$). ANOVAs were again employed to look at the differences between the rater groups after assumptions for parametric tests were checked a third time. Power analysis revealed that 30 subjects divided into eight groups with 25 repetitions, an assumed effect size of $f = 0.2$, and an alpha of 0.05 had an estimated power of 0.80.

The previous analyses addressed the first two research questions, but did not address how the raters arrive at their final ratings. Therefore, an analysis of the raters’ linguistic category ratings was needed to explore how the raters arrived at the final rating, giving insight into how they defined the construct of the exam. Raters’ linguistic category ratings (functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness) were analyzed through Multivariate Analyses of Variance (MANOVAs), where the compounding effect of differences between rater groups were analyzed as a whole and also differences between linguistic category ratings were also addressed.
Chapter 5: Results

5.1 Overview

This study examined the differences between how native speaker (NS) raters and non-native speaker (NNS) raters evaluated speaking proficiency, particularly at high levels of acquisition. The analysis of data included both the ratings assigned by the rater groups as well as the justifications given for the ratings. Since the native speaker was defined both as an ideal speaker of the language and as a native acquirer of the language, the data were analyzed divided the raters into three groupings according to these different native speaker constructs: native and non-native speakers, English proficiency levels of the raters, and first language of the raters. Both the final and linguistic category ratings were compared. These analyses were derived from the following research questions.

1. Do native and non-native speaker raters assign comparable ratings on speaking tests?
2. Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?
3. Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?
4. Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

Data were analyzed using both descriptive statistics and parametric tests. Parametric tests included both analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA).
Both ANOVAs and MANOVAs required that certain assumptions be met for the normality of data. Prior to implementing these tests, the data were examined for normality of distribution.

5.2 Normality of the data

ANOVA are parametric tests, and therefore require parametric data, which are data with a normal distribution. This study’s data met the four assumptions required for parametric tests (Field, 2005):

1. **Independence of ratings: Ratings should be given independently of each other.** In this data set, the ratings were given independently. The ILR SLDs were criterion-referenced and the raters were trained not to compare samples to each other. Additionally, the order in which the samples were rated was mixed for the different raters. This way, there would be no ordering effect (e.g., where one examinee was always rated after another examinee and the rater potentially influenced by the previous examinee’s performance).

2. **Interval data: Data points should be measured on an interval scale.** Although the ILR SLDs were not originally an interval scale, they have been converted to an interval scale for the purpose of this research. The ILR Levels ranged from 0 to 5, so the base level rating was converted to an interval number and a plus level rating was converted to 0.6. For example, a rating of 2+ was converted to 2.6. The conversion of ILR ratings to interval numbers has been practiced numerous times in prior Language Testing research (Clark, 1988; Stansfield et al., 1990). The equivalence of the plus level rating to 0.6 was the standard in the US Federal Government, and it was routinely coded as a 0.6 in lieu of a plus.
3. **Homogeneity of variance**: Variances throughout the data should be the same. For the ANOVAs used in the analyses, this assumption was checked through Levene’s test, and data was interpreted accordingly.

4. **Normally distributed data**: Data should come from a normally distributed population. To check this assumption, statistics of skewness and kurtosis were examined and histograms were reviewed for normality. These statistics were reviewed for each rater group prior to conducting parametric tests.

![Figure 2. Selected Exams by Level and Examinee Native Speaker Group](image)

Since much of the previous research on NS versus NNS speaker raters focused on the academic environment, the subjects that were rated were typically classroom language learners. On the ILR scale, these examinees would probably have been rated anywhere between the ILR 0 and 2 levels. One aspect of this study that distinguished it from the previous research was that the
focus was on higher proficiency examinees than were used in previous comparable research studies. To achieve this goal, twenty-five exams were selected ranging from ILR Level 2+ to ILR Level 5 (Figure 2).

Both native and non-native examinees’ speaking exams were selected for rating in the sample according to how they would naturally occur in the population, but still showed a more or less even distribution across levels. Within the FBI applicant population, there were very few NS examinees that received Level 2+ ratings and very few NNS examinees received Level 4+ or 5 ratings, so they were not represented. The remaining ILR levels had representation from both NS and NNS examinees.

Initially, the normality of the data in the overall sample was reviewed. Because the sample was quite large (n=750; 30 raters times 25 interviews), assumptions of normality could not reliably be determined through z-scores of skewness and kurtosis alone (Field, 2005). The skewness score was (0.22) and the kurtosis score was (-0.80) for all of the exams combined. As both were near 0, there was no significant skewness. The large sample number made the standard errors for skewness and kurtosis 0.09 and 0.18 respectively.

A review of the histograms of the rated exams’ original ratings (Figure 3) and the raters’ actual ratings (Figure 4) revealed similar distributions. Even though the actual ratings that were assigned were lower than the original ratings given to the exams, the distributions of the two groups are similarly shaped. A normal distribution of ratings was evidence that the exams were representative of the sample of FBI English language proficiency examinees.
5.3 Exam level redistribution

The raters in the study largely assigned ratings to the sample tests that were lower than the original ratings assigned. There were several explanations for this phenomenon. Typically, the original raters performed duties as both test administrators and raters simultaneously. Their cognitive load was much more intense than the raters in the study, who had only to rate the test from a recording. The original raters had the additional tasks of formulating appropriate questions and situations, presenting the elicitations (questions/situations) in a smooth manner, interacting with the examinee, interacting with the co-tester and rating the exam.

The cognitive load on the group of raters in the research, who only had to rate the exam, was much lighter. They could focus more on errors that occurred and take more scrupulous notes on what they observed, which, if anything tended to result in a lower rating, as indicated in Table 2.

Raters typically only rate from recordings a couple of times a year. While ratings from recordings were considered valid, the methodology that they followed for the study was not the
raters’ typical practice. Using recordings was unavoidable for the present research. Also, the fact that the original raters were part of the interview may have led them to evaluate the examinee’s performance differently. Moreover, the tests that were rated were selected from a population of exams that were administered between 2004 and 2006 in an effort to make sure that the original testers/raters on the test were not currently in the English rater population. The testing and rating principles and procedures have since been further codified, resulting in more consistent results (FBI, 2009).

Analyses of the data set included an evaluation of the descriptive statistics as well as parametric tests for all the exams grouped together, as well as analyses of the exams by ILR Level: Level 2/2+ exams, Level 3/3+ exams, and Level 4/4+/5 exams. It was considered inappropriate to categorize exams by their original ratings because the current raters (according to current practices) had a consensus that the exam was at a different level. Since one of the research questions deals with why raters assign the ratings they do, it was deemed appropriate to redistribute the exams into new ILR Level groupings. For example, if the original raters thought that an exam belonged at the ILR Level 3 (placing it in the Level 3/3+ exams group), while the majority of raters in this study gave it a Level 2+ (placing it in the Level 2/2+ exams group), raters’ sub ratings and comments justifying the rating would align with the Level 2+ final rating, therefore it was important to keep this tests’ ratings and sub ratings with the 2+ group.
Table 2. Redistribution of Ratings from Original ILR Levels to New Exam Groupings

<table>
<thead>
<tr>
<th>Original ILR Level</th>
<th>Examinee Native Speaker Group</th>
<th>NS Median</th>
<th>NNS Median</th>
<th>Average of NS and NNS Median</th>
<th>New Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+</td>
<td>NNS</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2/2+</td>
</tr>
<tr>
<td>2+</td>
<td>NNS</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2/2+</td>
</tr>
<tr>
<td>3+</td>
<td>NNS</td>
<td>2.6</td>
<td>2.6</td>
<td>2.6</td>
<td>2/2+</td>
</tr>
<tr>
<td>2+</td>
<td>NNS</td>
<td>2.6</td>
<td>2.8</td>
<td>2.7</td>
<td>2/2+</td>
</tr>
<tr>
<td>3</td>
<td>NNS</td>
<td>2.8</td>
<td>2.6</td>
<td>2.7</td>
<td>2/2+</td>
</tr>
<tr>
<td>3</td>
<td>NS</td>
<td>3</td>
<td>2.6</td>
<td>2.8</td>
<td>2/2+</td>
</tr>
<tr>
<td>3</td>
<td>NNS</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3/3+</td>
</tr>
<tr>
<td>3</td>
<td>NS</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3/3+</td>
</tr>
<tr>
<td>3</td>
<td>NS</td>
<td>3.3</td>
<td>3</td>
<td>3.15</td>
<td>3/3+</td>
</tr>
<tr>
<td>4</td>
<td>NNS</td>
<td>3.3</td>
<td>3</td>
<td>3.15</td>
<td>3/3+</td>
</tr>
<tr>
<td>3+</td>
<td>NNS</td>
<td>3.6</td>
<td>3</td>
<td>3.3</td>
<td>3/3+</td>
</tr>
<tr>
<td>3+</td>
<td>NS</td>
<td>3.6</td>
<td>3</td>
<td>3.3</td>
<td>3/3+</td>
</tr>
<tr>
<td>3</td>
<td>NNS</td>
<td>3.3</td>
<td>3.6</td>
<td>3.45</td>
<td>3/3+</td>
</tr>
<tr>
<td>3+</td>
<td>NS</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3/3+</td>
</tr>
<tr>
<td>3+</td>
<td>NS</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3/3+</td>
</tr>
<tr>
<td>4</td>
<td>NNS</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3/3+</td>
</tr>
<tr>
<td>4</td>
<td>NS</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3/3+</td>
</tr>
<tr>
<td>4</td>
<td>NNS</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3/3+</td>
</tr>
<tr>
<td>4</td>
<td>NS</td>
<td>4</td>
<td>3.6</td>
<td>3.8</td>
<td>3/3+</td>
</tr>
<tr>
<td>3+</td>
<td>NNS</td>
<td>3.8</td>
<td>4</td>
<td>3.9</td>
<td>3/3+</td>
</tr>
<tr>
<td>4+</td>
<td>NS</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4/4+/5</td>
</tr>
<tr>
<td>5</td>
<td>NS</td>
<td>4.3</td>
<td>4</td>
<td>4.15</td>
<td>4/4+/5</td>
</tr>
<tr>
<td>5</td>
<td>NS</td>
<td>4.8</td>
<td>4.6</td>
<td>4.7</td>
<td>4/4+/5</td>
</tr>
<tr>
<td>5</td>
<td>NS</td>
<td>4.8</td>
<td>4.6</td>
<td>4.7</td>
<td>4/4+/5</td>
</tr>
<tr>
<td>4+</td>
<td>NS</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4/4+/5</td>
</tr>
</tbody>
</table>
For those reasons, the original ratings were altered to match the adjusted ratings given by the participant raters. Since the NS and NNS groups did not have the same numbers of raters, the median rating of each group was calculated, as the median was less susceptible than the mean to the influence of outlier ratings. The medians of the NS and NNS raters were averaged, so the NS or NNS ratings would have equal influence on the new rating. Then new exam level groups were compiled for the analyses below. Table 2 showed the original ratings, the average of the NS and NNS rater group medians, and the new grouping for the exam.

5.4 Native and non-native speaker rater comparison: Final ratings

5.4.1 Comparison of NS and NNS raters’ final ratings: Introduction.

To answer the first research question (Do native and non-native speaker raters assign comparable ratings on speaking tests?), the raters were grouped into NS and NNS groups, and the final ratings on the 25 FBI English SPTs were examined. Therefore, the first analysis measured the degree of similarity/difference between the final ratings given by NS and NNS raters of speaking proficiency tests at ILR Levels 2 through 5. This rating was an overall, holistic score that reflected an examinee’s sustained ability to complete functional tasks with the necessary quality. The overall final ratings from raters from NS and NNS backgrounds were compared using descriptive statistics and ANOVAs.

5.4.2 Comparison of NS and NNS raters’ final ratings: Inter-rater reliability.

The inter-rater reliability statistics of the NS and NNS rater groups were examined. Inter-rater reliability describes the consistency of ratings among the raters within each group (Field, 2005). A higher reliability statistic means that there the members of the group had similar ratings
to other group members. Each group’s inter-rater reliability statistic was measured against the standard of acceptability for the purpose. For this study, the statistic would ideally be greater than 0.8, although at least 0.7 (Krippendorff, 2004; Lombard et al., 2003). For this research, Krippendorff’s alpha was used because it accommodates interval ratings, like the transformed interval ILR ratings in this study. Additionally, Krippendorff’s alpha allows for the comparison of two or more rater groups, which were used in later analyses. Krippendorff’s alpha also takes into account the number of raters in the group, adjusting for large or small groups.

Table 3. Inter-rater Reliability Statistics: Native Speaker Group

<table>
<thead>
<tr>
<th>Rater native speaker group</th>
<th>N coders</th>
<th>N cases</th>
<th>N decisions</th>
<th>Krippendorff’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>6</td>
<td>25</td>
<td>150</td>
<td>0.77</td>
</tr>
<tr>
<td>NNS</td>
<td>24</td>
<td>25</td>
<td>600</td>
<td>0.59</td>
</tr>
<tr>
<td>All raters</td>
<td>30</td>
<td>25</td>
<td>750</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3 shows that the NS raters had a higher Krippendorff’s alpha value than the NNS raters. An alpha value of 0.77 for the NS raters is considered acceptable for holistic ratings of proficiency, although not ideal (Krippendorff, 2004; Lombard et al., 2003). However, the NNS raters’ alpha of 0.59 was below the 0.70 minimum acceptability threshold.
5.4.3 Comparison of NS and NNS raters’ final ratings: All exams.

As seen in Table 4, there was virtually no descriptive difference between the mean ratings given by the native speaker raters and the non-native speaker raters. The native speaker raters assigned a mean rating of 3.45 across all exams, with a standard error of 0.06; the non-native speaker raters assigned a mean rating of 3.44, with a standard error of 0.04. The range of scores assigned by the two groups was identical, with both groups giving scores ranging from ILR 1+ to ILR 5. Both groups had a median score of 3.6.

Table 4. Final Ratings for All Exams: Native vs. Non-native Raters

<table>
<thead>
<tr>
<th>Rater native speaker group</th>
<th>Sample size</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Standard error</th>
<th>Skewness (SE)</th>
<th>Kurtosis (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native speakers (n=6)</td>
<td>150</td>
<td>3.45</td>
<td>0.79</td>
<td>0.06</td>
<td>0.30 (0.20)</td>
<td>-0.36 (0.39)</td>
</tr>
<tr>
<td>Non-native speakers (n=24)</td>
<td>600</td>
<td>3.44</td>
<td>0.90</td>
<td>0.04</td>
<td>0.21 (0.10)</td>
<td>-0.88 (0.20)</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>750</td>
<td>3.44</td>
<td>0.88</td>
<td>0.03</td>
<td>0.22 (0.09)</td>
<td>-0.80 (0.18)</td>
</tr>
</tbody>
</table>

Further review of the descriptive statistics revealed some differences in the distribution of the scores of the two rater groups. The non-native speaker raters had a slightly larger standard deviation from the mean (0.90) than the native speaker raters (0.79).

Neither the native nor non-native rater groups were particularly skewed. Overall the sample had a z-score of skewness (determined by dividing the skewness statistic by the skewness standard error) of 2.46. The non-native speaker raters had a z-score of skewness of 2.06 with the native speaker raters having a z-score of 1.53. The NS raters had a z-score of kurtosis of -0.92,
not significantly different from a normal distribution. In contrast, the NNS raters’ distribution was platykurtic, having as z-score of -4.42. Due to the large size of the sample, it was important to also inspect the distributions visually rather than depending solely on analysis from z-score (Field, 2005). Figures 5 and 6 showed the histograms of the NS raters and the NNS raters respectively. These box plots revealed very little significant difference between the two groups’ distributions, considering both skewness and kurtosis.

Figure 5. NS Raters’ Final Ratings

Figure 6. NNS Raters’ Final Ratings

Prior to conducting the parametric tests on the data, the four assumptions for parametric data needed to be reviewed. As established previously, the data were independent and interval. The review of skewness, kurtosis and the histograms of the rater groups revealed that the groups’ distribution of ratings did not differ from normal. Due to the large number of data points in the sample, a review of the histograms in

Figure 5 and Figure 6 confirmed their normal distribution and obviated any significant z-score for skewness and kurtosis. To check for the homogeneity of variance, Levene’s test was
Examined for the two groups. Levene’s test revealed a significant result at \( p = 0.01 \), meaning that the error variance between the two groups was not equal. As a result, the Welch’s \( F \) statistic was reported for the ANOVA.

Considering that the main measures of central tendency, the mean and the medians, of the two rater groups were nearly or exactly identical, it was no surprise that an ANOVA of the NS and the NNS rater groups revealed no significant difference. There was not a significant effect of being a native speaker on final exam ratings, \( F(1, 749) = 0.02, p = 0.88 \). A post hoc power analysis revealed that the power coefficient was 1.00, which meant if a significant result had existed, than it would have been detected (Type \( \beta \) error). The final ratings given by the NS and NNS rater groups to all exams were not different.

5.4.4 Comparison of NS and NNS raters’ final ratings: Analysis of exams by level of proficiency

Analyzing all exams together, as above, implied that it takes the same skill for a rater to make a determination about the proficiency of an ILR Level 2+ speaker as about an ILR Level 5 speaker. Each of the ILR base levels has a different set of functions on which the examinee must demonstrate his or her ability (FBI, 2009). Additionally, there are different expectations for the examinee’s quality of linguistic performance at each level. Since many of the NNS raters have not achieved the level of proficiency that they are rating, their capability to rate may vary across levels. As a result, it was important to look at the NS and NNS raters’ evaluations on each of the ILR levels of exams: Level 2/2+ exams, Level 3/3+ exams, and Level 4/4+/5 exams. Results from the descriptive statistics for these three sub-groupings were included in Table 5.
Table 5. Final Ratings for Each Proficiency Level: NS vs. NNS Raters

<table>
<thead>
<tr>
<th>Rater native speaker group</th>
<th>Sample size</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>(SE)</th>
<th>Kurtosis (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Level 2/2+ Exams</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native speakers (n=6)</td>
<td>36</td>
<td>2.62</td>
<td>0.45</td>
<td>0.08</td>
<td>0.20 (0.39)</td>
<td>0.72 (0.77)</td>
</tr>
<tr>
<td>Non-native speakers (n=24)</td>
<td>144</td>
<td>2.54</td>
<td>0.53</td>
<td>0.04</td>
<td>0.69 (0.20)</td>
<td>1.00 (0.40)</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>2.55</td>
<td>0.51</td>
<td>0.04</td>
<td>0.60 (0.18)</td>
<td>0.91 (0.36)</td>
</tr>
<tr>
<td><strong>Level 3/3+ Exams</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native speakers (n=6)</td>
<td>84</td>
<td>3.41</td>
<td>0.46</td>
<td>0.05</td>
<td>-0.06 (0.26)</td>
<td>0.80 (0.52)</td>
</tr>
<tr>
<td>Non-native speakers (n=24)</td>
<td>336</td>
<td>3.47</td>
<td>0.69</td>
<td>0.04</td>
<td>0.38 (0.13)</td>
<td>-0.42 (0.27)</td>
</tr>
<tr>
<td>Total</td>
<td>420</td>
<td>3.46</td>
<td>0.65</td>
<td>0.03</td>
<td>0.38 (0.12)</td>
<td>-0.18 (0.24)</td>
</tr>
<tr>
<td><strong>Level 4/4+/5 Exams</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native speakers (n=6)</td>
<td>30</td>
<td>4.56</td>
<td>0.51</td>
<td>0.09</td>
<td>-0.89 (0.43)</td>
<td>-0.60 (0.83)</td>
</tr>
<tr>
<td>Non-native speakers (n=24)</td>
<td>120</td>
<td>4.44</td>
<td>0.62</td>
<td>0.06</td>
<td>-0.89 (0.22)</td>
<td>-0.28 (0.44)</td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>4.46</td>
<td>0.60</td>
<td>0.05</td>
<td>-0.92 (0.20)</td>
<td>-0.20 (0.39)</td>
</tr>
</tbody>
</table>

On the ILR Level 2/2+ exams, the medians of both the native speaker and the non-native speaker rater groups were the same: 2+. Likewise, there was very little difference between their mean ratings. The NS raters gave a mean score of 2.62, whereas the NNS raters assigned a mean score of 2.54 to the ILR Level 2/2+ exams. The standard deviation was slightly larger for the NNS raters, (0.53) than for the NS raters (0.45). For the NS raters, the majority of the scores
were in the 2+ to 3 range, with just a few raters assigning 1+, 2, or 3+ scores. For the NNS raters, the scores extended from 1+ to 4, with one rating of 4+, two whole levels higher than the median rating of 2+ for exams in this group.

Additional information about the two rater groups’ distributions was gleaned from an examination of their histograms (Figure 7 and Figure 8) and the skewness and kurtosis statistics (Table 5). The NS raters’ distribution of scores did not indicate any significant skewness or kurtosis. The skewness and kurtosis statistics revealed that the NNS raters had significant ($p < 0.05$) positive $z$-scores of skewness (3.44) and kurtosis (2.40). The positive value of skewness indicated a greater number of low scores in the NNS raters’ final rating distributions.

![Figure 7. NS Raters’ Final Ratings of L 2/2+ Exams](image)

![Figure 8. NNS Raters’ Final Ratings of L 2/2+ Exams](image)

In the ILR Level 3/3+ exams (Table 5), again there were few differences between the measures of central tendency of the NS and NNS rater groups. There was a median rating of 3+ (3.6) from both groups of raters. The mean of the NS raters was 3.41; the mean of the NNS raters
was only 0.06 higher at 3.47. The non-native raters’ ratings were distributed between ILR Levels 2 and 5, and the native speakers’ ratings were distributed between ILR Levels 2 and 4+. The NS raters had a narrower standard deviation (0.46) than the NNS raters (0.69). Since plus level scores were converted to 0.6 (with 0.4 being the difference between a plus level and the next higher base level), this 0.23 difference reflects approximately one half of a plus level.

Much of the information about the samples’ distributions was reflected also in the histograms in Figure 9 and Figure 10. The z-score of skewness for the non-native speakers’ ratings of Level 3 performances was 2.82, indicating more lower scores from the non-native speaker raters; however, the large sample size tended to exaggerate skewness z-scores (Field, 2005) and the distribution of non-native raters’ scores of Level 3 performances was not considered so skewed as to invalidate means-based statistical analyses. The native speaker group distribution showed no significant skewness or kurtosis.

![Figure 9. NS Raters’ Final Ratings of L 3/3+ Exams](image)

![Figure 10. NNS Raters’ Final Ratings of L 3/3+ Exams](image)
There was a similar pattern in the ILR Level 4/4+/5 exams as the ILR Level 2/2+ and 3/3+ exams (Table 5). The native and non-native speaker rater groups shared a median of 4.6. The mean native speaker rating was 4.56 and non-native speaker rating is 4.44, as indicated in Table 3. The native speaker raters had a smaller standard deviation (0.51) than the non-native raters (0.62), indicating more uniform ratings within the native rater group. The range of ratings for the NNS raters was from ILR Level 2+ to 5, whereas the NS raters assigned ratings from the 3+ to 5 range.

Figure 11. NS Raters’ Final Ratings of L 4/5 Exams

Figure 12. NNS Raters’ Final Ratings of L 4/5 Exams

Again, viewing the histograms of the rater groups (Figure 11 and Figure 12) added information about the samples’ distributions. The histograms showed that there was a tendency of both groups to give high ratings for the Level 4/4+/5 examinees. This was likely due the fact
that the ILR Level 5 rating is the highest rating attainable, so the distribution was truncated at that point. An analysis of skewness and kurtosis showed that the non-native speaker rater group was negatively skewed, with a z-score statistic of -4.01. The native speaker rater’s z-score statistic of -2.07 showed the same tendency. Both were significant at $p < .05$. On the contrary, neither group had a significant z-score for kurtosis.

Prior to conducting the parametric tests on the data, the four assumptions for parametric data were reviewed, as was done previously. Again, the data were independent and interval. Our review of skewness, kurtosis and the histograms of the rater groups revealed that the groups’ distribution of ratings did not differ from normal. To check for the homogeneity of variance, Levene’s test was examined for the two groups, across the three exam level groupings. Levene’s test did not reveal a significant result for the NS/NNS group comparison for the Level 2/2+ exams, $p = 0.08$, meaning that the error variance between the two groups is equal. However, Levene’s test did produce a significant result for the Level 3/3+ exams and the Level 4/4+/5 exams, with $p = 0.00$ and $p = 0.05$ respectively. As a result, the Welch’s F statistic from the ANOVA reflected adjustments that were made for the significant homogeneity of variance for the Levels 3/3+ and 4/4+/5 exam groupings.

Although the measures of central tendency looked similar between the native and non-native speaker raters and among the three groups of varying levels of exams, an analysis of variance (ANOVA) determined if any differences were significant. Results from the one-way ANOVAs comparing the native and non-native speaker group ratings in each of the three exam level groups were reported in Table 6.
Table 6. Analyses of Variance for Native Speaker Group by Exam Level

<table>
<thead>
<tr>
<th>Source</th>
<th>df, error</th>
<th>$F$</th>
<th>$p$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2/2+ exams</td>
<td>1, 178</td>
<td>0.81</td>
<td>0.37</td>
<td>0.01</td>
</tr>
<tr>
<td>Level 3/3+ exams</td>
<td>1, 418</td>
<td>0.53</td>
<td>0.47</td>
<td>0.00</td>
</tr>
<tr>
<td>Level 4/4+/5 exams</td>
<td>1, 148</td>
<td>0.98</td>
<td>0.32</td>
<td>0.01</td>
</tr>
</tbody>
</table>

As Table 6 indicated, the value of $p$ was greater than 0.05 for each exam level, indicating no exam levels where the rater’s nativity made a significant difference in the final rating of the exams: Level 2/2+ exams $F(1, 178) = 0.81, p = 0.37$; Level 3/3+ exams $F(1, 418) = 0.53, p = 0.47$; Level 4/4+/5 exams $F(1, 148) = 0.98, p = 0.32$. A post hoc power analysis revealed that the power coefficient was 0.97 for the Level 2/2+ exams, 1.00 for the Level 3/3+ exams, and 0.94 for the Level 4/4+/5 exams, which means there was little to no opportunity for a not detecting a significant result that existed (Type $\beta$ error).

Overall, there was no evidence that NS and NNS raters assigned significantly different final ratings on speaking proficiency exams at ILR Level 2 through 5. These results provided the answer to the first research question:

Do native and non-native speaker raters assign comparable ratings on speaking tests?

The data in the current study showed no statistically significant differences between the ratings of the native speaker and non-native speaker rater groups.

Even though there was not a significant difference between the means, there was some
evidence of a potential trend in the distribution of the data in the two rater groups. In each analysis, a larger standard deviation appeared in the NNS rater group than in the NS rater group. In the end, however, any differences in ratings averaged out overall, with the result of no significant differences between the ratings of NS and NNS raters.

5.5 Rater English proficiency comparison: Final ratings

5.5.1 Comparison of raters’ final ratings by English proficiency: Introduction

As discussed previously, the issue of what it is to be a native speaker was debated, as being a native speaker can mean how a person acquired the language or how well a person speaks the language. Much of Language Testing and Second Language Acquisition research used native speakers as a baseline for comparison, assuming that their language performances reflect ideal speakers, yet rarely was the language proficiency of the native speaker actually quantified. In this section, the study explored differences of the final ratings assigned by raters at different ILR Base Level ranges: Level 2 (2+ raters); Level 3 (3/3+ raters); Level 4 (4/4+ raters); and Level 5 (5 raters). This rater grouping addressed the second research question: Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency? The Level 5 raters were the NS raters, as there were not any NNS raters in the pool of qualified and available participants, although Level 5 NNS exist. For the same reason, the Levels 2+ through 4+ raters were all NNS raters. The final ratings of raters with varying English speaking proficiencies were compared using descriptive statistics and ANOVAs.
5.5.2 Comparison of raters’ final ratings by English proficiency: Inter-rater reliability

The results of inter-rater reliability among the new grouping of raters according to English speaking proficiency were displayed in Table 7. Inter-rater reliability was measured using Krippendorff’s alpha. As before, the highest alpha value was found in the Level 5 (NS/English) raters at 0.77, which was an acceptable inter-rater reliability statistic for tests similar to the SPT (Krippendorff, 2004; Lombard et al., 2003). The other three English proficiency groups’ inter-rater reliability statistics fell below the 0.70 acceptability threshold for proficiency testing. Interestingly, the Level 2 and Level 3 speaker group had the second-highest Krippendorff’s alpha value (0.62), not the Level 4 group (0.58), whose English proficiency most closely matched that of the Level 5 raters. Finally, the alpha for the entire set of raters of 0.62 remained below the 0.70 minimum acceptability threshold.

Table 7. Inter-rater Reliability Statistics: Rater English Proficiency Level

<table>
<thead>
<tr>
<th>Rater English proficiency level</th>
<th>N coders</th>
<th>N cases</th>
<th>N decisions</th>
<th>Krippendorff's alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2/2+</td>
<td>6</td>
<td>25</td>
<td>150</td>
<td>0.62</td>
</tr>
<tr>
<td>Level 3/3+</td>
<td>14</td>
<td>25</td>
<td>350</td>
<td>0.62</td>
</tr>
<tr>
<td>Level 4/4+</td>
<td>4</td>
<td>25</td>
<td>100</td>
<td>0.58</td>
</tr>
<tr>
<td>Level 5</td>
<td>6</td>
<td>25</td>
<td>150</td>
<td>0.77</td>
</tr>
<tr>
<td>All raters</td>
<td>30</td>
<td>25</td>
<td>750</td>
<td>0.62</td>
</tr>
</tbody>
</table>
5.5.3 Comparison of raters’ final ratings by English proficiency: All exams.

A review of the means of the four rater groups split by English speaking proficiency (shown in Table 8) revealed some differences among the final ratings. The ILR Level 5 speakers (previously known as the native speaker raters) assigned a mean rating of 3.45 across all exams, with a standard error of 0.06. The Level 4 range raters showed a similar final rating, with a mean of 3.46 and a standard error of 0.09. The Level 3 raters gave higher ratings, with a mean of 3.62 and a standard error of 0.05. The lowest mean rating was from the Level 2 rater group, who assigned a mean final rating of 3.24 and a standard error of 0.07. The lowest mean rating of the Level 2 group was also reflected in the medians of the groups. The Level 2 raters had a median final rating of 3, while the other groups had a median rating of 3.6 (ILR 3+).

<table>
<thead>
<tr>
<th>Rater English speaking proficiency level</th>
<th>Sample size</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Standard error</th>
<th>Skewness (SE)</th>
<th>Kurtosis (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2 raters (n=6)</td>
<td>150</td>
<td>3.24</td>
<td>0.89</td>
<td>0.07</td>
<td>0.42 (0.20)</td>
<td>-0.44 (0.39)</td>
</tr>
<tr>
<td>Level 3 raters (n=14)</td>
<td>350</td>
<td>3.52</td>
<td>0.91</td>
<td>0.05</td>
<td>0.14 (0.13)</td>
<td>-1.07 (0.26)</td>
</tr>
<tr>
<td>Level 4 raters (n=4)</td>
<td>100</td>
<td>3.46</td>
<td>0.85</td>
<td>0.09</td>
<td>0.16 (0.24)</td>
<td>-0.53 (0.48)</td>
</tr>
<tr>
<td>Level 5 raters (n=6)</td>
<td>150</td>
<td>3.45</td>
<td>0.79</td>
<td>0.06</td>
<td>0.30 (0.20)</td>
<td>-0.36 (0.39)</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>750</td>
<td>3.44</td>
<td>0.88</td>
<td>0.03</td>
<td>0.22 (0.09)</td>
<td>-0.80 (0.18)</td>
</tr>
</tbody>
</table>

As in the NS/NNS comparison, the Level 5 raters had a smaller standard deviation than the other groups in the comparison. Here, the Level 5 raters had a standard deviation of 0.79,
with the other groups having larger standard deviations. The Level 2 raters had a standard
deviation of 0.89, the Level 3 raters had the largest standard deviation of 0.91, and the Level 4
raters’ standard deviation was 0.85.

None of the rater groups were particularly skewed. As mentioned in the previous
comparison, the sample had a \( z \)-score of skewness of 2.46. The only rater group that had a \( z \)-
score of skewness above the 1.96 mark (indicating it is significant at \( p = 0.05 \)) was the Level 2
group, with a \( z \)-score of 2.13. The Level 5 raters had a \( z \)-score of kurtosis of -0.92, not
significantly different from a normal distribution. In contrast, the Level 3 raters’ distribution was
platykurtic, having as \( z \)-score of -4.12. Due to the large sample size, it was important to also
inspect the distributions visually rather than depending solely on analysis from \( z \)-score (Field,
2005).

![Figure 13. Histogram of L2 Raters’ Final Ratings](image1)

![Figure 14. Histogram of L3 Raters’ Final Ratings](image2)
Figure 13, Figure 14, Figure 15, and Figure 16 show the histograms of the raters with different English speaking proficiency levels. These histograms revealed very little difference between the Level 2, 3, 4, and 5 groups’ distributions, considering both skewness and kurtosis.

Prior to conducting the ANOVAs, the four assumptions for parametric data were reviewed and met. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis showed that the data met the normality requirement. Levene’s test revealed a significant result at $p = 0.03$, meaning that the error variance among the four language groups was not equal. As a result, the Welch’s $F$ statistic from the ANOVA was used, which adjusts for the lack of homogeneity of variance.

The native speaker/non-native speaker rater group comparison revealed no statistical difference between the raters groups, but when the NNS raters were categorized by English
speaking proficiency, significant differences between the groups were revealed. In this case, a one-way ANOVA revealed a significant effect of rater English speaking proficiency level on final ILR rating for the exams, $F(3, 746) = 3.71, p = 0.01$, partial $\eta^2 = 0.02$. A post hoc power analysis revealed that the power coefficient was 1.00, which meant if there had been a significant result, it would have been detected (Type $\beta$ error).

The effect size (partial $\eta^2$) of 0.02 was very small\textsuperscript{6}, meaning that overall, the raters’ English proficiency level accounted for very little of the differences in ratings among the rater proficiency level groups. Bonferroni post hoc tests showed significant differences between Level 2+ raters and Level 3/3+ raters ($p = 0.01$), with the Level 3 raters assigning ratings at an average of 0.28 higher than the Level 2 raters. No significant differences were found between the Level 5 raters and any of the other English proficiency level groups.

5.5.4 Comparison of raters’ final ratings by English proficiency: Analysis of exams by level of proficiency

As before, the exams were divided into three categories (Level 2/2+, Level 3/3+, and Level 4/4+/5) to determine if the level of the exam affected whether or not there were any differences among the rater group divided by English proficiency. Since some of raters were evaluating tests at levels above their own proficiency ratings, the accuracy of their ratings suffer. Results from the descriptive statistics for these three sub-groupings were listed in Table 9.

\textsuperscript{6} The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
Table 9. Final Ratings for Each Exam Proficiency Level: Rater English Proficiency Levels

<table>
<thead>
<tr>
<th>Rater English speaking proficiency level</th>
<th>Sample size</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Standard error</th>
<th>Skewness (SE)</th>
<th>Kurtosis (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 2/2+ Exams (n=6)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 raters (n=6)</td>
<td>36</td>
<td>2.39</td>
<td>0.50</td>
<td>0.08</td>
<td>0.38 (0.39)</td>
<td>0.13 (0.77)</td>
</tr>
<tr>
<td>Level 3 raters (n=14)</td>
<td>84</td>
<td>2.58</td>
<td>0.52</td>
<td>0.06</td>
<td>1.07 (0.26)</td>
<td>2.07 (0.52)</td>
</tr>
<tr>
<td>Level 4 raters (n=4)</td>
<td>24</td>
<td>2.59</td>
<td>0.56</td>
<td>0.12</td>
<td>0.01 (0.47)</td>
<td>-1.06 (0.92)</td>
</tr>
<tr>
<td>Level 5 raters (n=6)</td>
<td>36</td>
<td>2.62</td>
<td>0.45</td>
<td>0.08</td>
<td>0.20 (0.39)</td>
<td>0.72 (0.77)</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>180</td>
<td>2.55</td>
<td>0.51</td>
<td>0.04</td>
<td>0.60 (0.18)</td>
<td>0.91 (0.36)</td>
</tr>
<tr>
<td><strong>Level 3/3+ Exams (n=14)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 raters (n=6)</td>
<td>84</td>
<td>3.22</td>
<td>0.65</td>
<td>0.07</td>
<td>0.59 (0.26)</td>
<td>0.64 (0.52)</td>
</tr>
<tr>
<td>Level 3 raters (n=14)</td>
<td>196</td>
<td>3.57</td>
<td>0.71</td>
<td>0.05</td>
<td>0.31 (0.17)</td>
<td>-0.69 (0.35)</td>
</tr>
<tr>
<td>Level 4 raters (n=4)</td>
<td>56</td>
<td>3.51</td>
<td>0.62</td>
<td>0.08</td>
<td>0.35 (0.32)</td>
<td>-0.46 (0.63)</td>
</tr>
<tr>
<td>Level 5 raters (n=6)</td>
<td>84</td>
<td>3.41</td>
<td>0.46</td>
<td>0.05</td>
<td>-0.06 (0.26)</td>
<td>0.80 (0.52)</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>420</td>
<td>3.46</td>
<td>0.65</td>
<td>0.03</td>
<td>0.38 (0.12)</td>
<td>-0.18 (0.24)</td>
</tr>
<tr>
<td><strong>Level 4/4+/5 Exams (n=5)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 raters (n=6)</td>
<td>30</td>
<td>4.30</td>
<td>0.67</td>
<td>0.12</td>
<td>-0.44 (0.43)</td>
<td>-1.24 (0.83)</td>
</tr>
<tr>
<td>Level 3 raters (n=14)</td>
<td>70</td>
<td>4.52</td>
<td>0.58</td>
<td>0.07</td>
<td>-1.32 (0.29)</td>
<td>1.29 (0.57)</td>
</tr>
<tr>
<td>Level 4 raters (n=4)</td>
<td>20</td>
<td>4.36</td>
<td>0.69</td>
<td>0.15</td>
<td>-0.45 (0.51)</td>
<td>-1.36 (0.99)</td>
</tr>
<tr>
<td>Level 5 raters (n=6)</td>
<td>30</td>
<td>4.56</td>
<td>0.51</td>
<td>0.09</td>
<td>-0.89 (0.43)</td>
<td>-0.60 (0.83)</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>150</td>
<td>4.46</td>
<td>0.60</td>
<td>0.05</td>
<td>-0.92 (0.20)</td>
<td>-0.20 (0.39)</td>
</tr>
</tbody>
</table>
In the ILR Level 2/2+ exams, the Level 5 raters had the highest mean score (2.62) and the Level 2 raters had the lowest mean score (2.39) for a difference of 0.23. There was not much difference between the mean ratings of the four groups, but the mean rating increased slightly as the English proficiency of the rater group increased. The Level 3 raters gave a mean assigned a mean rating of 2.58. The Level 4 raters gave a 2.59 overall. The median of all four rater groups was the same: 2+.

The standard deviation varied by a small amount between the four rater proficiency groups. The Level 5 raters had the smallest standard deviation of 0.45, and the Level 4 speakers’ standard deviation is 0.56. The Levels 2, 4, and 5 raters assigned ratings from the Level 1+ to Level 3+ on ILR Level 2/2+ exams. The Level 3 raters had ratings that extend all the way to the Level 4+, giving them the broadest range of ratings.

Additional information about the rater groups’ distributions on the Level 2/2+ exams was gleaned from an examination of the skewness and kurtosis statistics (Table 9). The Level 2, Level 4 and Level 5 raters’ distribution of scores did not indicate any significant skewness or kurtosis. The skewness and kurtosis statistics revealed that the Level 3 raters had significant \( p < 0.05 \) positive z-scores of skewness (4.08) and kurtosis (3.98). However, the Level 3 group had the largest number of raters, so the normality of its distribution was determined by examining the histograms instead of significant z-score statistics.

In the ILR Level 3/3+ exams, the Level 2 group raters had a median rating of Level 3, lower that than the other groups’ medians of Level 3+. The Level 2 group also had the lowest mean (3.22). The mean of the Level 5 raters was 3.41, with the remaining two groups’ means being higher than the Level 5 raters, with a mean rating of 3.51 from the Level 4 raters and 3.57 from the Level 3 raters. The ratings were distributed between ILR Levels 2 and 4+ ranges most
rater groups, with the Level 3 group extending to the Level 3 range. The Level 5 raters had the lowest standard deviation, 0.46, which was 0.25 lower than the group with the highest standard deviation, the Level 3 raters, 0.71. This difference reflected approximately one half of a plus level.

The z-score of skewness for the Level 2 rater group (derived from Table 9) was 2.25, significant at $p < 0.05$. The statistics from Table 9 did not reveal that any of the rater groups had a significant level of kurtosis except for Level 3 raters, who had a kurtosis $z$-score of -1.99. Both groups’ skewness and kurtosis values did not affect the normality of the data because of the large sample size.

There was a somewhat similar pattern in the ILR Level 4/4+/5 exams as the ILR Level 2/2+ and 3/3+ exams. The medians of all four rater groups were 4+. The Level 5 and Level 3 raters had similar mean ratings (4.56 and 4.52 respectively); whereas, the Level 2 and Level 4 raters had lower ratings (with mean ratings of 4.30 and 4.36, respectively). Table 9 also revealed the standard distribution of 0.51 for the Level 5 raters, still the smallest among the four rater groups. The Level 2, 3 and 4 groups had standard deviations of 0.67, 0.58, and 0.69 respectively. The Level 5 raters assigned ratings from the ILR Level 3+ to 5 ranges; whereas, the Levels 2 and 4 raters assigned ratings from the broader 3 to 5 range, and the Level 3 raters assigned ratings from the even broader 2+ to 5 range. Although the Level 3 group had a more similar mean and standard deviation to the Level 5 group than the other two groups, its $z$-scores for both skewness (-4.61) and kurtosis (2.27) were significant at $p < 0.05$; however, again the measure of skewness was mitigated by the group’s large $n$.

Differences between the means did appear to be more substantive among the English proficiency level groups comparison than the native/non-native speaker group comparison. An
analysis of variance (ANOVA) determined if any group differences were significant at each of the three levels of exams. Results from the three one-way ANOVAs for each exam level were reported in Table 10.

Table 10. Analyses of Variance for Rater English Proficiency by Exam Level

<table>
<thead>
<tr>
<th>Source</th>
<th>df, error</th>
<th>F</th>
<th>p</th>
<th>Partial η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2/2+ exams</td>
<td>3, 176</td>
<td>1.51</td>
<td>0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>Level 3/3+ exams</td>
<td>3, 416</td>
<td>6.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Level 4/4+/5 exams</td>
<td>3, 146</td>
<td>1.41</td>
<td>0.24</td>
<td>0.03</td>
</tr>
</tbody>
</table>

As Table 10 indicated, for the Level 2/2+ exams, the raters’ proficiency level was not significantly related to their ratings: $F(3, 176) = 1.51, p = 0.21$, partial $\eta^2 = 0.03$. Similarly, for the Level 4/4+/5 exams, raters’ proficiency level was not significantly related to their rating: $F(3, 146) = 1.41, p = 0.24$, partial $\eta^2 = 0.03$.

The Level 3/3+ exams told a different story. Here, the English proficiency of the raters did affect the rating assigned: $F(3, 416) = 6.00, p = 0.00$, partial $\eta^2 = 0.04$. A post hoc power

---

7 All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that the data met the normality requirement. Levene’s test revealed whether or not homogeneity of variance was met, and if not, the Welch’s F statistic adjusted for the lack of homogeneity of variance.

8 The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
analysis revealed that the power coefficient was 0.90, which means if a significant result had existed, than it would have been detected (Type β error). The relatively low partial $\eta^2$ indicated that rater proficiency had very little effect on ratings. Bonferroni post hoc tests showed significant differences between Level 2 and Level 3 raters ($p = 0.00$), with the Level 2 raters giving final ratings of 0.35 lower than the Level 3 raters. Similarly, there was a nearly significant difference between the Level 2 and Level 4 raters ($p = 0.059$). Here the Level 2 raters were 0.29 lower than the Level 4 raters. Again, there were no significant differences between the Level 5 groups and any of the other 3 groups.

Overall, there was evidence that the English speaking proficiency of raters was related to the final ratings assigned for exams overall and in the Level 3/3+ range exams. These differences did not appear in the Level 2/2+ exams and the Level 4/4+/5 exams. These results addressed the second research question:

Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?

The data showed some significant differences among the raters of varying English speaking proficiency English speaking proficiency levels. No significant differences emerged between the Level 5 raters (who were the native speakers) and any of the other English proficiency level groups, the non-native speakers.

These results should be interpreted with caution. There was a significant difference between the English proficiency level groups’ means overall. When the ratings were considered by exam level, the significant difference occurred in the Level 3/3+ exams. However, the partial
\( \eta^2 \) was small in both instances, meaning that it only contributed to a small part of rating variation. Still, there appeared to be some trends in the distribution of the data in the rater groups, as appeared in the native speaker/non-native speaker comparison in the first research question. Consistently, the Level 5 speakers had a narrower standard deviation than any of the non-native speaker groups, who had English proficiency levels of ILR Level 2+-4+.

5.6 Rater native language comparison: Final ratings

5.6.1 Comparison of raters’ final ratings by native language: Introduction

Grouping native speakers of various language backgrounds into a NNS raters group assumed that NNS are a homogeneous group, or at least they shared a unified trait (or lack of trait) of being not native. The results from the comparison of NS and NNS raters showed that there was no significant difference between the final ratings of the two groups, but do those results hold regardless of the native language of the rater? To investigate further the question of whether NS and NNS raters assigned equivalent ratings, NS raters’ (English raters’) performances were compared to each of the seven languages that were the NNS raters’ native languages separately (Arabic, Farsi, French, German, Mandarin, Spanish, and Vietnamese). This analysis addressed the third research question: Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?

5.6.2 Comparison of raters’ final ratings by native language: Inter-rater reliability

An examination of inter-rater reliability, measured using Krippendorf’s alpha, was conducted and results were displayed in Table 11. The English rater had the highest alpha statistic, 0.77. The other native language rater group that had an inter-rater reliability high
enough to be consistent in their ratings was the German rater group, with an alpha of 0.74. All of the other six rater groups fell below the 0.7 minimum standard (Krippendorff, 2004; Lombard et al., 2003). The Mandarin, Farsi and Spanish raters had the lowest alpha statistic at 0.52. Again, the alpha for the entire set of raters of 0.62 was below the 0.70 minimum acceptability threshold.

Table 11. Inter-rater Reliability Statistics: Rater Native Language

<table>
<thead>
<tr>
<th>Rater native language</th>
<th>N coders</th>
<th>N cases</th>
<th>N decisions</th>
<th>Krippendorff’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>6</td>
<td>25</td>
<td>150</td>
<td>0.77</td>
</tr>
<tr>
<td>Arabic</td>
<td>4</td>
<td>25</td>
<td>100</td>
<td>0.60</td>
</tr>
<tr>
<td>Farsi</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>0.53</td>
</tr>
<tr>
<td>French</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>0.67</td>
</tr>
<tr>
<td>German</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>0.74</td>
</tr>
<tr>
<td>Mandarin</td>
<td>4</td>
<td>25</td>
<td>100</td>
<td>0.52</td>
</tr>
<tr>
<td>Spanish</td>
<td>4</td>
<td>25</td>
<td>100</td>
<td>0.53</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>0.67</td>
</tr>
<tr>
<td>All raters</td>
<td>30</td>
<td>25</td>
<td>750</td>
<td>0.62</td>
</tr>
</tbody>
</table>

5.6.3 Comparison of raters’ final ratings by native language: All exams

The analysis of native and non-native speaker rater groups revealed that there was no real difference between the mean ratings of the two groups. Results from Table 12 reveal that the English raters’ (formerly the NS raters’ in previous analyses) final rating was situated in the
middle of the distribution of final ratings of the native language groups. The Arabic speaker raters (3.67) and the Vietnamese speaker raters (3.64) assigned the highest ratings. The Mandarin speaker raters (3.26) assigned the lowest ratings, with an average 0.41 below the Arabic raters or approximately a plus level.

Table 12. Final Ratings for All Exams: Rater Native Language

<table>
<thead>
<tr>
<th>Rater native language</th>
<th>Sample size</th>
<th>Mean</th>
<th>Standard error</th>
<th>Standard deviation</th>
<th>Skewness (SE)</th>
<th>Kurtosis (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (n=6)</td>
<td>150</td>
<td>3.45</td>
<td>0.06</td>
<td>0.79</td>
<td>0.30 (0.20)</td>
<td>-0.36 (0.39)</td>
</tr>
<tr>
<td>Arabic (n=4)</td>
<td>100</td>
<td>3.67</td>
<td>0.08</td>
<td>0.82</td>
<td>0.18 (0.24)</td>
<td>-1.07 (0.48)</td>
</tr>
<tr>
<td>Farsi (n=3)</td>
<td>75</td>
<td>3.34</td>
<td>0.12</td>
<td>1.02</td>
<td>0.25 (0.28)</td>
<td>-1.14 (0.55)</td>
</tr>
<tr>
<td>French (n=3)</td>
<td>75</td>
<td>3.34</td>
<td>0.11</td>
<td>0.97</td>
<td>0.27 (0.28)</td>
<td>-0.76 (0.55)</td>
</tr>
<tr>
<td>German (n=3)</td>
<td>75</td>
<td>3.48</td>
<td>0.12</td>
<td>1.02</td>
<td>0.11 (0.28)</td>
<td>-1.28 (0.55)</td>
</tr>
<tr>
<td>Mandarin (n=4)</td>
<td>100</td>
<td>3.26</td>
<td>0.09</td>
<td>0.86</td>
<td>0.48 (0.24)</td>
<td>-0.63 (0.48)</td>
</tr>
<tr>
<td>Spanish (n=4)</td>
<td>100</td>
<td>3.36</td>
<td>0.07</td>
<td>0.73</td>
<td>0.37 (0.24)</td>
<td>-0.18 (0.48)</td>
</tr>
<tr>
<td>Vietnamese (n=3)</td>
<td>75</td>
<td>3.64</td>
<td>0.10</td>
<td>0.86</td>
<td>-0.04 (0.28)</td>
<td>-0.82 (0.55)</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>750</td>
<td>3.44</td>
<td>0.03</td>
<td>0.88</td>
<td>0.22 (0.09)</td>
<td>-0.80 (0.18)</td>
</tr>
</tbody>
</table>

The standard deviations of the language groups varied from 0.73 (Spanish) and 0.79 (English) to 1.02 (Farsi and German). In an analysis of the distributions of the ratings, $z$-scores were derived for skewness and kurtosis. The only language rater group that was significantly skewed was the Mandarin raters ($z$-score = 1.99). Three language rater groups had significant $z$-
scores for kurtosis ($p < 0.05$): Farsi (-2.09), Arabic (-2.23), and German (-2.24). Although each of these statistics was significant, they were not significant by much, and therefore did not reflect an abnormal distribution.

The box plots for the rater groups (Figure 17) gave additional information about the distribution of the ratings. Here, it was revealed that the medians for the rater groups differ. The English, Arabic, German, and Vietnamese rater groups had a median of 3.6 (ILR Level 3+). The Spanish group had a median of 3.3.

![Box plots showing rating distribution for different rater groups](image)

Figure 17. Rater Native Language Groups’ Final Ratings for All Exams
Finally, the remaining groups, to include Farsi, French, and Mandarin, had a median of 3.0 (ILR Level 3). Here, the differences between the distributions of the rater groups were also depicted, with the majority of English and Spanish groups’ ratings clustered between ILR Level 2+ and 4. Other groups had broader distributions from ILR Level 2 to 5, with Farsi and French raters giving several ratings at the ILR Level 1+.

Prior to conducting the ANOVAs, the four assumptions for parametric data needed to be reviewed and met. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis showed that the data meet the normality requirements. The Levene’s test revealed a significant result at $p = 0.00$, meaning that the error variance among the eight language groups was not equal. As a result, the Welch’s $F$ statistic was referenced, which adjusted for the lack of homogeneity of variance.

In order to determine whether or not there was a statistical difference between the means of the different rater language groups, an analysis of variance was conducted. Overall, the differences between the means of the rater language groups were significant: $F(7, 749) = 2.52, p = 0.01$, partial $\eta^2 = 0.02$. A post hoc power analysis revealed that the power coefficient was 0.85, which meant if there had been a significant difference, as there was, it would have been detected (Type β error). The effect size (partial $\eta^2$) was small at 0.02\(^9\). The small effect size meant that the raters’ native language did have an impact on their ratings of English, but it accounted for very little of the difference among the native language groups. Post hoc t-tests revealed that 0.41

\(^9\) The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
difference between the Mandarin and Arabic rater groups was significant \( (p = 0.03) \), with the Arabic raters giving higher ratings than the Mandarin raters.

5.6.4 Comparison of raters’ final ratings by native language: Analysis of exams by level

An analysis of the final ratings by exam level revealed whether or not differences between English and other language raters existed at Level 2/2+, Level 3/3+, and Level 4/4+/5 exams (Table 13).

Table 13. Final Ratings for Each Proficiency Level: Rater Native Language Groups

<table>
<thead>
<tr>
<th>Native Language</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Level 2/2+ Exams</td>
<td></td>
<td>Level 3/3+ Exams</td>
<td></td>
<td>Level 4/4+/5 Exams</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n=6)</td>
<td></td>
<td>(n=14)</td>
<td></td>
<td>(n=5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English (n=6)</td>
<td>36</td>
<td>2.62</td>
<td>0.45</td>
<td>84</td>
<td>3.41</td>
<td>0.46</td>
<td>30</td>
<td>4.56</td>
<td>0.51</td>
</tr>
<tr>
<td>Arabic (n=4)</td>
<td>24</td>
<td>2.83</td>
<td>0.38</td>
<td>56</td>
<td>3.69</td>
<td>0.65</td>
<td>20</td>
<td>4.59</td>
<td>0.55</td>
</tr>
<tr>
<td>Farsi (n=3)</td>
<td>18</td>
<td>2.51</td>
<td>0.77</td>
<td>42</td>
<td>3.25</td>
<td>0.82</td>
<td>15</td>
<td>4.59</td>
<td>0.49</td>
</tr>
<tr>
<td>French (n=3)</td>
<td>18</td>
<td>2.26</td>
<td>0.51</td>
<td>42</td>
<td>3.42</td>
<td>0.68</td>
<td>15</td>
<td>4.43</td>
<td>0.71</td>
</tr>
<tr>
<td>German (n=3)</td>
<td>18</td>
<td>2.29</td>
<td>0.34</td>
<td>42</td>
<td>3.54</td>
<td>0.76</td>
<td>15</td>
<td>4.73</td>
<td>0.43</td>
</tr>
<tr>
<td>Mandarin (n=4)</td>
<td>24</td>
<td>2.44</td>
<td>0.38</td>
<td>56</td>
<td>3.31</td>
<td>0.73</td>
<td>20</td>
<td>4.11</td>
<td>0.71</td>
</tr>
<tr>
<td>Spanish (n=4)</td>
<td>24</td>
<td>2.63</td>
<td>0.47</td>
<td>56</td>
<td>3.38</td>
<td>0.51</td>
<td>20</td>
<td>4.18</td>
<td>0.65</td>
</tr>
<tr>
<td>Vietnamese (n=3)</td>
<td>18</td>
<td>2.69</td>
<td>0.60</td>
<td>42</td>
<td>3.71</td>
<td>0.61</td>
<td>15</td>
<td>4.59</td>
<td>0.54</td>
</tr>
<tr>
<td>Total (n=30)</td>
<td>180</td>
<td>2.55</td>
<td>0.51</td>
<td>420</td>
<td>3.46</td>
<td>0.65</td>
<td>150</td>
<td>4.46</td>
<td>0.60</td>
</tr>
</tbody>
</table>

144
In the Level 2/2+ exams, the mean ratings ranged between 2.26 (French) and 2.83 (Arabic), for a spread of 0.57 approximately a plus level. The English raters’ mean rating was 2.62, which was in the middle of the range, and only 0.07 higher than the mean of the all raters together. The standard deviations ranged between 0.34 (German) and 0.77 (Farsi).

In the Level 3/3+ exams, the mean ratings ranged between 3.25 (Farsi) and 3.71 (Vietnamese), for a spread of 0.46, also approximating a plus level. The English raters’ mean rating was 3.41, which was again in the middle of the range, and this time 0.05 lower than the mean of the group. The standard deviations ranged between 0.46 (English) and 0.82 (Farsi). Although the English raters’ standard deviation was practically the same in the Level 2 and Level 3 range exams, they went from having the fourth lowest (out of eight) standard deviation to the lowest standard deviation of the rater language groups.

In the Level 4/4+/5 exams, the mean ratings ranged between 4.11 (Mandarin) and 4.73 (German), for a spread of 0.62, the largest range of means found among the three exam groups. This difference represented a plus level rating on the ILR scale. The English raters’ mean rating was 4.56, showing that the English scores were consistently in the middle of the range, on this occasion being 0.10 higher than the mean of the group. The standard deviations ranged between 0.43 (German) and 0.71 (French and Mandarin). As in the Level 3 range exams, the English raters had a low standard deviation (0.51), second lowest of the group.

Particular rater language groups were either mostly consistently higher than the English group (Arabic, Mandarin, Spanish, Vietnamese) or lower than the English group (Farsi, French). Interestingly, the German raters vary from being stricter than the English group on the lower-rated exams, to more lenient on the higher level exams. The smallest standard deviations overall
occurred in the lower level exams, indicating more consistency among language groups at that level.

Previously, the results of the ANOVA for all exams together indicated that the rater’s native language made a significant difference in the rating that was produced, although the effect of that contribution was very small overall. The application of the same parametric test to the exams at different levels of exams revealed significant differences again at each level\(^\text{10}\). In these cases, there was not only a strong power statistic, but also a larger effect size than for the ANOVA of all tests combined. Results for the ANOVAs for rater native language by exam level were listed in Table 14.

Table 14. Analyses of Variance for Rater Native Language by Exam Level

<table>
<thead>
<tr>
<th>Source</th>
<th>df, error</th>
<th>F</th>
<th>p</th>
<th>Partial (\eta^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2/2+ exams</td>
<td>7, 172</td>
<td>3.40</td>
<td>0.00</td>
<td>0.12</td>
</tr>
<tr>
<td>Level 3/3+ exams</td>
<td>7, 412</td>
<td>3.34</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>Level 4/4+/5 exams</td>
<td>7, 142</td>
<td>2.67</td>
<td>0.01</td>
<td>0.12</td>
</tr>
</tbody>
</table>

\(^{10}\) All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that the data met the normality requirement. Levene’s test revealed whether or not homogeneity of variance was met, and if not, the Welch’s F statistic adjusted for the lack of homogeneity of variance.
The ANOVA of the ILR Level 2/2+ exams showed a significant difference among ratings of the native language groups: $F(7, 172) = 3.40, p = 0.00$, partial $\eta^2 = 0.12^{11}$. This difference meant that in the ILR Level 2/2+ exams, the rater’s native language made a significant difference in the final rating produced. Moreover, the partial $\eta^2$ of 0.12 showed that the raters’ native language did account for a notable part of the variance. Bonferroni post hoc tests revealed that significant differences were found between Arabic raters and French raters, with the Arabic raters’ having a mean final rating of 0.58 over the French raters ($p = 0.01$). Additionally, the Arabic raters had a significant ($p = 0.01$) final rating 0.54 higher than the German raters. It was important to note that none of the NNS first language groups’ mean ratings differed significantly from the English group’s mean rating.

Similar to the Level 2/2+ exams, the ANOVA of the ILR Level 3/3+ exams showed a significant difference among ratings of the native language groups: $F(7, 412) = 3.34, p = 0.00$, partial $\eta^2 = 0.05^{12}$. This means that in the mid-range exams as well, the rater’s native language made a significant difference in the final rating produced. However, the partial $\eta^2$ of 0.05 did not reveal a strong effect of rater language on ratings of Level 3/3+ exams. The Bonferroni post hoc tests revealed that the Arabic raters had a mean final rating significantly higher than that of the Farsi raters ($p = 0.02$). Additionally, the Vietnamese raters had a final rating significantly higher ($p = 0.03$) than the Farsi raters. The respective differences of 0.45 for the Arabic raters and 0.46 for the Vietnamese raters over the Farsi raters represented a difference on average of a plus level.

---

11 A post hoc power analysis revealed that the power coefficient is 0.85, which means there was little opportunity for a not detecting a significant result that exists (Type β error).
12 A post hoc power analysis revealed that the power coefficient is 0.85, which means there was little opportunity for a not detecting a significant result that exists (Type β error).
Continuing the trend, the ANOVA of the ILR Level 4/4+/5 exams showed a clear significant difference among ratings of the native language groups, ANOVA: $F(7, 142) = 2.67, p = 0.01$, partial $\eta^2 = 0.12$. The rater’s native language made a significant difference in the final ratings given to high-level exams. The partial $\eta^2$ of 0.12\(^{13}\) indicated a notable effect of rater language on ratings of high-level exams. Bonferroni post hoc tests indicated that there was a difference approaching significance ($p = 0.056$) between German raters and Mandarin raters, with the German raters’ having a mean final rating of 0.62 over the Mandarin raters.

Overall, an analysis of the non-native speaker (NNS) rater groups revealed that the NNS raters do not act homogenously. In fact, there were significant differences between the different first language groups over all the exams as well as when the exams were separated by level. These results addressed the third research question:

Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?

In the distribution of ratings, the English (NS) raters did fall in the middle of the spectrum of language raters, and never showed any significant differences with any of the other first language groups. Whether or not the rater’s first language was closely related to English, such as German, or very different from English, such as Mandarin, there was not a significant difference between that language group and the English language group. There were, however, ______________

\(^{13}\) A post hoc power analysis revealed that the power coefficient is 0.85, which means there was little opportunity for a not detecting a significant result that exists (Type $\beta$ error).
significant differences among the language groups themselves. The Arabic and Vietnamese raters had a tendency to give higher ratings at times, and the Farsi and Mandarin gave lower ratings at times. The German raters gave low ratings to the lower level tests and higher ratings to the higher level tests. Trends did emerge among the language groups, but no trend involved the English raters.

5.7 Native and non-native speaker rater comparison: Linguistic category ratings

5.7.1 Comparison of NS and NNS raters’ category ratings: Introduction

In the previous three sections, the final ratings of the 25 SPT interviews were explored, first looking at a native speaker rater versus non-native speaker rater comparison. Later, the rater groups were divided by their English language proficiency: the native, English speakers being the Level 5 raters, and the other NNS raters being divided into Level 4, 3, and 2 raters. Finally the ratings of the English raters (NS) were compared to the ratings of the seven other native language groups represented (NNSs). All of these analyses looked at the final ratings issued, which were the scores of record for the exams. However, those were not the only ratings assigned. As a regular part of the rating process, SPT raters also assigned seven linguistic category ratings in the areas of functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness. These ratings could only be ILR base levels (0 - 5, no pluses assigned) and informed the final rating. An examination of the linguistic category ratings provided insight into how raters arrived at their final rating and what linguistic characteristics were critical to raters from different groups. This view one step further into the mind of the rater helped further define the construct being used in the final rating process, thereby further defining the construct of the exam. In this analysis, the linguistic category ratings
native speaker and non-native speaker raters were compared first, using descriptive statistics and MANOVAs. The results address the fourth research question: Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

5.7.2 Comparison of NS and NNS raters’ category ratings: All exams

Table 15 presented the means and standard deviations for the linguistic category ratings for all 25 exams. Overall, the mean ratings were lowest for the functions category and highest for the pronunciation category.

Table 15. Linguistic Category Ratings for All Exams: Rater Native Speaker Group

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS</td>
<td>NNS</td>
</tr>
<tr>
<td>Functions</td>
<td>3.33</td>
<td>3.26</td>
</tr>
<tr>
<td>Organization</td>
<td>3.55</td>
<td>3.39</td>
</tr>
<tr>
<td>Structures</td>
<td>3.45</td>
<td>3.42</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3.51</td>
<td>3.40</td>
</tr>
<tr>
<td>Fluency</td>
<td>3.45</td>
<td>3.54</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3.81</td>
<td>3.83</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3.40</td>
<td>3.42</td>
</tr>
</tbody>
</table>

NS (n = 150: 6 participants x 25 exams); NNS (n = 600: 24 participants x 25 exams)
In most of the linguistic categories, the ratings of the NS and NNS raters were very similar. There was only one category where the differences between the two groups were at least 0.10 of a rating different. In the organization category, the NS raters assigned a mean rating of 3.55, and the NNS raters gave a rating of 0.16 lower, at 3.39. Both the NS and NNS rater groups had median ratings of 3 across the categories, except for pronunciation, where there was a median rating of 4.

In all of the linguistic categories, the standard deviation of the NS raters was lower than the NNS raters, except for structures where the standard deviation was the same. The two largest differences in standard deviations between the two groups occurred in the fluency and social/cultural appropriateness categories, where the NS ratings’ standard deviations were 0.22 and 0.20 higher respectively. A difference of this amount was approximately half a plus level.

As with the ANOVAs conducted previously, the four assumptions of parametric data were reviewed prior to conducting MANOVAs of the linguistic category ratings. These ratings were independent and interval. Histograms showed that the distribution of each rater groups’ ratings did not differ from normal. Any correction needed in the homogeneity of variance was made by the use of Wilks’ lambda in the interpretation of the MANOVA results, and was indicated in the appropriate results table.

Considering that there were no differences between the medians of the linguistic category ratings from the NS raters and the NNS raters and only some small differences appeared when the means of the linguistic categories were examined, little difference in measures of central tendency in parametric tests was expected. In order to determine if any difference that occurred was significant, a multivariate analysis of variance (MANOVA) was conducted, which included
a measure of overall difference among all the linguistic categories combined as well as comparisons category by category.

Table 16. MANOVA for Linguistic Category Rating by Native Speaker Group

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>df, error</th>
<th>F</th>
<th>p</th>
<th>Partial $\eta^2$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall$^{14}$</td>
<td>7, 742</td>
<td>3.80</td>
<td>0.00</td>
<td>0.04</td>
<td>0.98</td>
</tr>
<tr>
<td>Functions</td>
<td>1, 748</td>
<td>0.60</td>
<td>0.44</td>
<td>0.00</td>
<td>0.12</td>
</tr>
<tr>
<td>Organization</td>
<td>1, 748</td>
<td>3.18</td>
<td>0.08</td>
<td>0.00</td>
<td>0.43</td>
</tr>
<tr>
<td>Structures</td>
<td>1, 748</td>
<td>0.09</td>
<td>0.76</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1, 748</td>
<td>1.68</td>
<td>0.20</td>
<td>0.00</td>
<td>0.25</td>
</tr>
<tr>
<td>Fluency</td>
<td>1, 748</td>
<td>1.14</td>
<td>0.29</td>
<td>0.00</td>
<td>0.19</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>1, 748</td>
<td>0.03</td>
<td>0.88</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>1, 748</td>
<td>0.08</td>
<td>0.77</td>
<td>0.00</td>
<td>0.06</td>
</tr>
</tbody>
</table>

As a part of the MANOVA, the result of the overall multivariate test revealed a significant difference between the NS and NNS raters among the linguistic category ratings, $F(7, 742) = 3.80, p = 0.00$, partial $\eta^2 = 0.04^{15}$. This statistic was shown in Table 16. Even though the

$^{14}$ Wilks’ lambda
$^{15}$ A post hoc power analysis revealed that the power coefficient is 0.85, which means there was little opportunity for a not detecting a significant result that exists (Type $\beta$ error).
difference is significant, the partial $\eta^2$ of 0.04 was very low, meaning that overall, the raters’
nativity did not account for much of the difference that occurred.

Reviewing the results of the individual linguistic category analyses (Table 16) showed
that none of the individual categories had a significant result at $p < 0.05$. The organization ratings
comparison between the NS and NNS raters approaches that mark, but at $p = 0.08$, it was not
significant: $F(1, 748) = 3.18, p = 0.08$, partial $\eta^2 = 0.00^{16}$. Overall, while there was a significant
difference between NS and NNS raters’ ratings on the linguistic category levels, it was not
concentrated in one linguistic category and it did not lead to a significant difference in the two
groups’ final ratings.

5.7.3 Comparison of NS and NNS raters’ category ratings: Analysis of exams by
level of proficiency

The exams were again divided into three categories (Level 2/2+, Level 3/3+, and Level
4/4+/5) to determine if the rater’s nativity had a different effect on the linguistic category ratings
at different exam levels. In the ILR Level 2/2+ exams, the NS raters assigned higher ratings than
the NNS raters in every category except for structures, as seen in Table 17. The largest difference
between the NS and NNS raters occurred in the organization category, with the NS mean being
0.32 higher than the NNS raters, which is approaching a plus level difference. The NS raters’
mean rating for structures was lower than the NNS raters, but only by 0.06. The majority of the
medians for the native speakers were Level 3, except for functions and structures. The NNS

\[ ^{16} \text{A post hoc power analysis revealed that the power coefficient is 0.85, which means there was little}
\text{opportunity for a not detecting a significant result that exists (Type } \beta \text{ error).} \]
raters had mostly Level 2 median scores for the linguistic categories, except for fluency and pronunciation. This meant that the NNS raters had lower median scores (Level 2) for the categories of organization, vocabulary, and social/cultural appropriateness than the NS raters (Level 3). Similar to what had been seen previously, Table 17 also tells us that the NS raters had narrower standard deviations than the NNS raters in most of the categories, except for functions and structures.

Table 17. Linguistic Category Ratings for Level 2/2+ Exams: Rater Native Speaker Group

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS</td>
<td>NNS</td>
</tr>
<tr>
<td>Functions</td>
<td>2.47</td>
<td>2.37</td>
</tr>
<tr>
<td>Organization</td>
<td>2.81</td>
<td>2.49</td>
</tr>
<tr>
<td>Structures</td>
<td>2.39</td>
<td>2.45</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>2.61</td>
<td>2.44</td>
</tr>
<tr>
<td>Fluency</td>
<td>2.81</td>
<td>2.64</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>2.92</td>
<td>2.77</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NS (n = 36: 6 participants x 6 exams); NNS (n = 144: 24 participants x 6 exams)
In order to determine whether or not the differences seen in the descriptive statistics were significant, MANOVAs were conducted for each of the exam level groupings. Table 18 showed the results for the Level 2/2+ exams, which indicated that when all the linguistic category ratings were considered together, there was a significant effect of rater nativity on the

---

17 Wilks’ lambda

18 All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that the data met the normality requirement. Wilks’ lambda was used to correct for any issues with homogeneity of variance.
linguistic category ratings: $F(7, 172) = 2.99, p = 0.01$, partial $\eta^2 = 0.11$. However, the partial $\eta^2$ indicated that being a native speaker or not accounts only for 0.11 of the variance, which was nevertheless a notable piece of the variation.

Table 18 also showed the ANOVAs for each linguistic category. As might have been predicted by the descriptive statistics, the only one that was significant at $p < 0.05$ is organization, $F(1, 178) = 8.40, p = 0.00$, partial $\eta^2 = 0.05$. Again though, this difference between NS and NNS raters only accounted for a small portion of the variance, 0.05.

Table 19 showed the means and standard deviations for the NS and NNS raters groups in each of the linguistic subcategories. The NS raters assigned higher ratings in functions, organization, and vocabulary, but lower in their ratings of fluency, pronunciation, and social/cultural appropriateness. Just as in the ratings from all exams combined, the mean structures rating was the same between the NS and NNS raters. In this instance, the largest difference between the NS and NNS raters occurred in the fluency category, with the NS mean being 0.19 lower than the NNS raters. A review of the medians given by the two groups revealed that both groups gave median ratings of ILR Level 3 in all categories except fluency and pronunciation. For pronunciation, both groups assigned a median rating of Level 4. In fluency, the NNS gave a rating of Level 4, and the NS had a median rating of 3.

---

19 The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).

20 The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
Table 19. Linguistic Category Ratings for Level 3/3+ Exams: Rater Native Speaker Group

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS</td>
<td>NNS</td>
</tr>
<tr>
<td>Functions</td>
<td>3.30</td>
<td>3.27</td>
</tr>
<tr>
<td>Organization</td>
<td>3.50</td>
<td>3.41</td>
</tr>
<tr>
<td>Structures</td>
<td>3.45</td>
<td>3.45</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3.50</td>
<td>3.45</td>
</tr>
<tr>
<td>Fluency</td>
<td>3.38</td>
<td>3.57</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3.80</td>
<td>3.92</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3.32</td>
<td>3.46</td>
</tr>
</tbody>
</table>

NS (n = 84: 6 participants x 14 exams); NNS (n = 336: 24 participants x 14 exams)

Again there was a pattern that the NS raters have lower standard deviations than the NNS raters. In this set of exams, the NS raters had lower standard deviations in every linguistic category, ranging from 0.11 points lower in structures to as much as 0.31 points lower in social cultural appropriateness.

Table 20 showed the results for the Level 3/3+ exams MANOVA\(^{21}\). Previously it was established that the trends for the NS/NNS differences in the mid-range exams followed the

---

\(^{21}\) All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that
trends of the overall exams that the largest differences occurred in the fluency category and the social/cultural appropriateness category, with the NNS raters assigning higher ratings in both.

Table 20. MANOVA for Linguistic Category Rating by Native Speaker Group: Level 3/3+

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>df, error</th>
<th>$F$</th>
<th>$p$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>7, 412</td>
<td>3.09</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>Functions</td>
<td>1, 418</td>
<td>0.12</td>
<td>0.73</td>
<td>0.00</td>
</tr>
<tr>
<td>Organization</td>
<td>1, 418</td>
<td>0.96</td>
<td>0.33</td>
<td>0.00</td>
</tr>
<tr>
<td>Structures</td>
<td>1, 418</td>
<td>0.00</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1, 418</td>
<td>0.23</td>
<td>0.63</td>
<td>0.00</td>
</tr>
<tr>
<td>Fluency</td>
<td>1, 418</td>
<td>3.76</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>1, 418</td>
<td>1.31</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>Social/Cultural Appropriateness</td>
<td>1, 418</td>
<td>2.17</td>
<td>0.14</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Like in MANOVA for all tests combined, the multivariate tests results indicated that when all the linguistic category ratings were considered together, there was a significant affect of rater nativity on the linguistic category ratings: $F(7, 412) = 3.09, p = 0.00$, partial $\eta^2 = 0.05$. The data met the normality requirement. Wilks’ lambda was used to correct for any issues with homogeneity of variance.

22 Wilks’ lambda
effect size (0.05) indicated that whether or not the rater was a native speaker accounted for a small portion of the variance\textsuperscript{23}.

The ANOVAs for each linguistic category were displayed in Table 20. As might be predicted by the descriptive statistics, the only one that was significant at $p < 0.05$ was fluency, $F(1, 418) = 3.76$, $p = 0.05$, partial $\eta^2 = 0.01$, power = 0.49. With a partial $\eta^2$ of 0.01, the significant difference between NS and NNS raters’ fluency ratings in Level 3/3+ exams accounted for very little of the variation in the ratings.

In the ILR Level 4/4+/5 exams, the NS raters assigned mean ratings higher than the NNS raters in every linguistic category except fluency, as seen in Table 21. The only category where the ratings were comparable between the two rater groups was social/cultural appropriateness. The largest difference between two rater groups was the difference of 0.25 in vocabulary. Both groups gave median ratings of ILR Level 4 in all categories except organization and fluency. For organization, the NS raters had a median of 4.5, whereas the NNS raters had a median of Level 5. In fluency, the NNS gave a rating of Level 5 and the NS had a median rating of 4. Overall, the NS and NNS raters groups were quite comparable at this level.

\textsuperscript{23} The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
Table 21. Linguistic Category Ratings for Level 4/4+/5 Exams: Rater Native Speaker Group

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th></th>
<th></th>
<th>Standard Deviation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS</td>
<td>NNS</td>
<td>Total</td>
<td>NS</td>
<td>NNS</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Functions</td>
<td>4.43</td>
<td>4.31</td>
<td>4.33</td>
<td>0.63</td>
<td>0.78</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>4.57</td>
<td>4.43</td>
<td>4.45</td>
<td>0.57</td>
<td>0.73</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Structures</td>
<td>4.73</td>
<td>4.53</td>
<td>4.57</td>
<td>0.45</td>
<td>0.67</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>4.63</td>
<td>4.38</td>
<td>4.43</td>
<td>0.62</td>
<td>0.75</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>4.40</td>
<td>4.54</td>
<td>4.51</td>
<td>0.56</td>
<td>0.62</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Pronunciation</td>
<td>4.93</td>
<td>4.83</td>
<td>4.85</td>
<td>0.25</td>
<td>0.44</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>4.43</td>
<td>4.39</td>
<td>4.40</td>
<td>0.68</td>
<td>0.75</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>

NS (n = 30: 6 participants x 5 exams); NNS (n = 120: 24 participants x 5 exams)

The NS raters consistently again had lower standard deviations than the NNS raters. However, these standard deviation differences were lower than the previous sets of exams, possibly because at this level, raters were prevented from overrating too much, as the scale ended at ILR 5. The descriptive statistics for the Level 4/4+/5 exams, revealed that the NS raters gave higher ratings in all the linguistic categories except fluency.
Table 22. MANOVA for Linguistic Category Rating by Native Speaker Group: Level 4/4+/5 Exams

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>df, error</th>
<th>$F$</th>
<th>$p$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>7, 142</td>
<td>2.22</td>
<td>0.04</td>
<td>0.10</td>
</tr>
<tr>
<td>Functions</td>
<td>1, 148</td>
<td>0.67</td>
<td>0.42</td>
<td>0.00</td>
</tr>
<tr>
<td>Organization</td>
<td>1, 148</td>
<td>0.98</td>
<td>0.32</td>
<td>0.01</td>
</tr>
<tr>
<td>Structures</td>
<td>1, 148</td>
<td>2.58</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1, 148</td>
<td>2.87</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>Fluency</td>
<td>1, 148</td>
<td>1.30</td>
<td>0.26</td>
<td>0.01</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>1, 148</td>
<td>1.45</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>Social/Cultural Appropriateness</td>
<td>1, 148</td>
<td>0.08</td>
<td>0.78</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The multivariate tests results, shown in Table 22, indicated that when all the linguistic category ratings were considered together, rater nativity had a significant effect on the linguistic category ratings: $F(7, 142) = 2.22, p = 0.04$, partial $\eta^2 = 0.10$. This result meant that the MANOVAs found significant differences between the NS and NNS raters at each of the exam.

---

21 Wilks’ lambda
25 All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that the data met the normality requirement. Wilks’ lambda was used to correct for any issues with homogeneity of variance.
levels, as well as when all the exams were combined. The partial $\eta^2$ indicates that whether or not a rater was a native speaker accounted for a notable portion of the variance, 0.10.$^{26}$

The descriptive statistics revealed that the largest difference between the NS raters and the NNS raters in exams at the Level 4/4+/5 occurred in the vocabulary category. The ANOVAs for each linguistic category displayed in Table 22 indicated that of all the linguistic categories, the $p$ value for vocabulary was closest to significant, although it did not pass the $p < 0.05$ criterion.

Overall, there was evidence that whether or not the rater was a native speaker affected the linguistic category ratings assigned for exams overall and at each of the exam level ranges. It should be noted that many of the significant results that were discovered had very small effect sizes, indicating that they were not the only contributors toward the overall variance between the scores. These results spoke to the answer of the fourth research question:

Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

In one interpretation of these results, significant differences appeared between NS raters and NNS raters’ ratings of the linguistic category scores, both overall on exams of all levels and at each particular exam level range. The linguistic category ratings indicated that the NS raters

$^{26}$ The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
and NNS raters were giving different weight or preference to different aspects of speech, even though these appeared to wash out prior to the assignment of the final rating.

A different interpretation of the results was that most of the ANOVAs of the individual linguistic categories did not find a significant difference between the NS and NNS raters. Exceptions to this generalization were that NS raters tended to rate organization higher overall and in Level 2/2+ exams and NNS raters tended to rate fluency higher in Level 3/3+ exams. When differences were noted, the small partial $\eta^2$ means that they were contributors, but only minor contributors.

Some of the trends in the previous descriptive statistics recur in these analyses. Almost uniformly, the NNS raters had a higher standard deviation than the NS raters. Additionally, in most instances, the NS raters assigned higher ratings than the NNS raters in the linguistic categories.

### 5.8 Rater English proficiency comparison: Linguistic category ratings

#### 5.8.1 Comparison of raters’ category ratings by English proficiency: Introduction

Finally, the linguistic category ratings were examined once again, this time with the NNS raters divided by their English speaking ability, as was done with the final ratings to answer research question 2. Again, this lead to four rating groups, the Level 2, 3, and 4 English speakers (NNS) and the Level 5 English speakers (NS). The results from the previous section indicated that rater English proficiency had more of an impact on final exam ratings than whether or not the rater was a native speaker. In line with that conclusion, it was expected that rater English speaking proficiency would have had an impact on the linguistic category ratings both over all exams and at each exam level. Again, this analysis provided further insight into how the raters in
different groupings arrive at their final rating and what linguistic characteristics were critical to
the raters from different groups, also contributing to research question four. Both descriptive
statistics and MANOVAs were analyzed for the results.

5.8.2 Comparison of raters’ category ratings by English proficiency: All exams

Table 23 showed that the raters with the lowest English proficiency (the Level 2 raters)
assigned the lowest ratings on linguistic categories, except for the category of pronunciation. No
one proficiency level group consistently assigned the highest ratings across the linguistic
categories. In many instances, the ratings for the Level 3, 4, and 5 groups were very similar, with
the Level 2 raters being at times anywhere from 0.20 to 0.40 (nearly half and ILR base level)
lower than another group on average. The largest ranges between the group means occurred in
the functions and organization categories, where the Level 2 raters were 0.40 lower than the
Level 3 and Level 5 raters respectively. In the case of functions, the Level 2 raters actually had a
mean rating of less than 3 across all exams.

Table 23 showed that of the English proficiency rater groups, the Level 5 raters almost
always had the lowest standard deviation across all linguistic categories, except in structures. In
many cases, the Level 3 raters had the highest standard deviation, though not in structures and
pronunciation. The largest spread between two groups’ standard deviations occurred in functions
and pronunciation, where the Level 3 raters had a standard deviation that was 0.20 higher than
the Level 5 raters. The medians were similar across the groups, with the median for the category
ratings being mostly Level 3.
Table 23. Linguistic Category Ratings for All Exams: Rater English Speaking Proficiency

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L2</td>
<td>L3</td>
</tr>
<tr>
<td>Functions</td>
<td>2.97</td>
<td>3.37</td>
</tr>
<tr>
<td>Organization</td>
<td>3.15</td>
<td>3.48</td>
</tr>
<tr>
<td>Structures</td>
<td>3.29</td>
<td>3.47</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3.17</td>
<td>3.46</td>
</tr>
<tr>
<td>Fluency</td>
<td>3.39</td>
<td>3.57</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3.86</td>
<td>3.75</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3.19</td>
<td>3.51</td>
</tr>
</tbody>
</table>

L2 (n = 150: 6 participants x 25 exams); L3 (n = 350: 14 participants x 25 exams); L4 (n = 100: 4 participants x 25 exams); L5 (n = 150: 6 participants x 25 exams)

Considering that there were no differences between the medians of the linguistic category ratings among the English proficiency groups, a significant difference among the rater groups was not expected. There were some notable differences in the means of the groups. Therefore, a multivariate analysis of variance (MANOVA) was conducted, including a measure of overall difference among all the linguistic categories combined as well as comparisons category by
category\textsuperscript{27}. The result of the overall multivariate test revealed a significant difference among the English speaking proficiency level groupings of raters in their linguistic category ratings, $F(21, 2125.42) = 4.77, p = 0.00$, partial $\eta^2 = 0.04$\textsuperscript{28}. The partial $\eta^2$ of 0.04 was very low\textsuperscript{29}, meaning that the raters’ English proficiency did not account for much of the rating differences observed.

Reviewing the results of the individual linguistic category analyses (Table 24) showed that several of the individual categories had a significant result at $p < 0.05$, but each with relatively low effect sizes (partial $\eta^2$). The low effect sized meant that the significant difference did not account for much of the variance\textsuperscript{30}. The functions ratings comparison had the highest partial $\eta^2$ at 0.03: $F(3, 746) = 6.56, p = 0.00$, partial $\eta^2 = 0.03$. Other than that, significant differences occurred in organization ($F(3, 746) = 6.56, p = 0.00$, partial $\eta^2 = 0.02$), vocabulary ($F(3, 746) = 4.09, p = 0.01$, partial $\eta^2 = 0.02$), pronunciation ($F(3, 746) = 2.56, p = 0.05$, partial $\eta^2 = 0.01$), and social/cultural appropriateness ($F(3, 746) = 4.06, p = 0.01$, partial $\eta^2 = 0.02$). Overall, raters of differing proficiency levels differed significantly in rating all of the linguistic categories except structures and fluency.

\textsuperscript{27} All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that the data met the normality requirement. Wilks’ lambda was used to correct for any issues with homogeneity of variance.
\textsuperscript{28} The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
\textsuperscript{29} The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
\textsuperscript{30} The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
Table 24. MANOVA for Linguistic Category Rating by Rater English Proficiency

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>df, error</th>
<th>$F$</th>
<th>$p$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall$^{31}$</td>
<td>21, 2125.43</td>
<td>4.77</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Functions</td>
<td>3, 746</td>
<td>6.56</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Organization</td>
<td>3, 746</td>
<td>5.33</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Structures</td>
<td>3, 746</td>
<td>1.17</td>
<td>0.32</td>
<td>0.01</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3, 746</td>
<td>4.09</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Fluency</td>
<td>3, 746</td>
<td>2.24</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3, 746</td>
<td>2.56</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3, 746</td>
<td>4.06</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Bonferroni post hoc tests gave additional information about instances where significant differences existed. In the functions category, the Level 2 raters’ mean rating was 0.40 lower than that of the Level 3 raters ($p = 0.00$), 0.34 lower than that of the Level 4 raters ($p = 0.03$) and 0.35 lower than that of the Level 5 raters ($p = 0.01$). In the organization category, the Level 2 raters’ mean rating was 0.32 lower than that of the Level 3 raters ($p = 0.00$) and 0.39 lower than that of the Level 5 raters ($p = 0.00$), but not significantly lower than that of the Level 4 raters. In the vocabulary category, the Level 2 raters’ mean rating was 0.29 lower than that of the Level 3 raters ($p = 0.02$), 0.34 lower than that of the Level 4 raters ($p = 0.05$) and 0.34 lower than that of

$^{31}$ Wilks’ lambda
the Level 5 raters ($p = 0.02$). In the social/cultural appropriateness category, the Level 2 raters’ mean rating was 0.32 lower than that of the Level 3 raters ($p = 0.00$), but not significantly different than any of the other groups. Finally, in the pronunciation category, one significant difference was discovered between two groups that did not involve the Level 2 raters. The Level 3 raters’ mean rating was 0.32 lower than that of the Level 4 raters ($p = 0.04$). Overall, the Level 2 raters contributed to many of the significant differences, and were typically significantly lower than other, higher proficiency rater groups.

5.8.3 Comparison of raters’ category ratings by English proficiency: Analysis of exams by level of proficiency

The exams were again divided into three categories (Level 2/2+, Level 3/3+, and Level 4/4+/5) to determine if rater English speaking proficiency had a different effect on the linguistic category ratings of different exam levels.

In the ILR Level 2/2+ exams, the Level 2 raters gave lower ratings than raters from the higher English ability groups in every category except for pronunciation, as seen in Table 25. The group with the highest mean ratings was consistently either the Level 4 or 5 raters. The difference between the groups with the highest and lowest means ranges from 0.23 to 0.48, which was approximately half an ILR base level. The largest difference between any rater groups occurred in the organization category, with the Level 4 raters’ mean being 0.48 higher than that of the Level 2 raters. The majority (five out of seven) of the medians for the Level 3, 4, and 5 were at the Level 3, except for structures and social/cultural appropriateness for the Level 4 raters and functions and structures for the Level 5 raters. The Level 2 raters had mostly Level 2 median scores (five out of seven) for the categories, except for fluency and pronunciation.
Table 25. Linguistic Category Ratings for Level 2/2+ Exams: Rater English Proficiency

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L2</td>
<td>L3</td>
</tr>
<tr>
<td>Functions</td>
<td>2.11</td>
<td>2.43</td>
</tr>
<tr>
<td>Organization</td>
<td>2.33</td>
<td>2.52</td>
</tr>
<tr>
<td>Structures</td>
<td>2.31</td>
<td>2.49</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>2.31</td>
<td>2.43</td>
</tr>
<tr>
<td>Fluency</td>
<td>2.58</td>
<td>2.62</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>2.72</td>
<td>2.71</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>2.47</td>
<td>2.56</td>
</tr>
</tbody>
</table>

L2 (n = 36: 6 participants x 6 exams); L3 (n = 84: 14 participants x 6 exams); L4 (n = 24: 4 participants x 6 exams); L5 (n = 36: 6 participants x 6 exams)

In order to determine whether or not the differences seen in the descriptive statistics were significant, MANOVAs were conducted for each of the exam level groupings. Table 26 showed the results for the Level 2/2+ exams. The multivariate tests results indicated that when all the linguistic category ratings were considered together, there was a significant difference.

---

32 All assumptions for parametric tests were met prior to conducting the ANOVAs. The data were independent and interval. Reviews of the histograms in combination of skewness and kurtosis show that the data met the normality requirement. Wilks’ lambda was used to correct for any issues with homogeneity of variance.
between English proficiency level groups on the linguistic category ratings: $F(21, 488.70) = 2.09, p = 0.00$, partial $\eta^2 = 0.08$. However, the partial $\eta^2$ indicated that rater English proficiency level accounted only for 0.08 of the variance, which was nevertheless a notable (Cohen, 1973).

Table 26: MANOVA for Linguistic Category Rating by Rater English Proficiency: Level 2/2+

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>$df$, error</th>
<th>$F$</th>
<th>$p$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall(^{33})</td>
<td>21, 488.70</td>
<td>2.09</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>Functions</td>
<td>3, 176</td>
<td>3.70</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>Organization</td>
<td>3, 176</td>
<td>4.26</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>Structures</td>
<td>3, 176</td>
<td>1.01</td>
<td>0.39</td>
<td>0.02</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3, 176</td>
<td>3.24</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>Fluency</td>
<td>3, 176</td>
<td>1.19</td>
<td>0.32</td>
<td>0.02</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3, 176</td>
<td>1.45</td>
<td>0.23</td>
<td>0.02</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3, 176</td>
<td>1.00</td>
<td>0.40</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 26 also showed the ANOVAs for each linguistic category. Three out of the seven linguistic categories appeared as significantly affected by the raters’ English proficiency ($p < \ldots$

\(^{33}\) Wilks’ lambda
0.05). Significant differences occurred in functions ($F(3, 176) = 3.70, p = 0.01$, partial $\eta^2 = 0.06$), organization ($F(3, 176) = 4.26, p = 0.01$, partial $\eta^2 = 0.07$), and vocabulary ($F(3, 176) = 3.24, p = 0.02$, partial $\eta^2 = 0.05$). However, these differences among the raters groups of difference English speaking proficiencies only accounted for a relatively small portion of the variance, ranging from 0.05 to 0.07 (Cohen, 1973).

Bonferroni post hoc tests revealed that for functions, organization, and vocabulary, the Level 2 raters had scores that were significantly lower than some of the other, higher English proficiency raters (Levels 3, 4, and 5). In the functions category, the Level 2 raters’ mean rating was 0.32 lower than that of the Level 3 raters ($p = 0.04$) and 0.43 lower than that of the Level 4 raters ($p = 0.03$). In the organization category, the Level 2 raters’ mean rating was 0.47 lower than that of the Level 5 raters ($p = 0.00$). In the vocabulary category, the Level 2 raters’ mean rating was 0.40 lower than that of the Level 4 raters ($p = 0.05$).

In the ILR Level 3/3+ exams, the Level 2 raters continued to give lower ratings in every category except for pronunciation, where the Level 5 raters had the lowest mean rating (Table 27). In many of the categories, the Level 4 raters had the highest means, for instance in organization, vocabulary, fluency, and pronunciation. When the Level 4 raters did not have the highest ratings, the Level 3 raters had high ratings. In the Level 3/3+ exams, there seemed to often be a considerable spread between the mean ratings the highest and lowest groups. In five out of seven categories, the difference was 0.40 or greater, which was approximately half of an ILR base level.
Table 27. Linguistic Category Ratings for Level 3/3+ Exams: Rater English Proficiency

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L2</td>
<td>L3</td>
</tr>
<tr>
<td>Functions</td>
<td>2.94</td>
<td>3.39</td>
</tr>
<tr>
<td>Organization</td>
<td>3.10</td>
<td>3.51</td>
</tr>
<tr>
<td>Structures</td>
<td>3.27</td>
<td>3.52</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3.14</td>
<td>3.55</td>
</tr>
<tr>
<td>Fluency</td>
<td>3.36</td>
<td>3.63</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3.95</td>
<td>3.83</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td>3.15</td>
<td>3.58</td>
</tr>
</tbody>
</table>

L2 (n = 84: 6 participants x 14 exams); L3 (n = 196: 14 participants x 14 exams); L4 (n = 56: 4 participants x 14 exams); L5 (n = 84: 6 participants x 14 exams)

In the remaining two categories (structures and fluency), the differences were also large, at 0.25 and 0.28 respectively. In most of the categories for most of the groups, the median rating was ILR Level 3. All four groups have a median rating of Level 4 for pronunciation.
Table 28. MANOVA for Linguistic Category Rating by Rater English Proficiency: Level 3/3+

Exams

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>df, error</th>
<th>$F$</th>
<th>$p$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall$^{34}$</td>
<td>21, 1170.85</td>
<td>3.78</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Functions</td>
<td>3, 416</td>
<td>8.38</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Organization</td>
<td>3, 416</td>
<td>6.67</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>Structures</td>
<td>3, 416</td>
<td>1.87</td>
<td>0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3, 416</td>
<td>5.77</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Fluency</td>
<td>3, 416</td>
<td>3.98</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3, 416</td>
<td>2.92</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3, 416</td>
<td>6.84</td>
<td>0.00</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 28 showed the results for the Level 3/3+ exams. Some of the trends seen in the previous analyses recurred here. For example, the ratings on the Level 2 raters were lower than those of the other raters and the Level 5 speakers had a smaller standard deviation than the other rater groups. The multivariate tests results indicated that when all the linguistic category ratings were considered together, there was a significant effect of rater English proficiency on the linguistic category ratings: $F(21, 1170.85) = 3.78$, $p = 0.00$, partial $\eta^2 = 0.06$. Again, the

$^{34}$ Wilks’ lambda
medium effect size (partial $\eta^2$) indicated that English proficiency level accounts for a moderate portion of the variance, 0.06 (Cohen, 1973).

The ANOVAs for each linguistic category were displayed in Table 28. At this level, there were several linguistic category ANOVAs that returned significant differences. In fact, every linguistic feature except structures was significant. Significant differences occurred in functions ($F(3, 416) = 8.38, p = 0.00, \text{partial } \eta^2 = 0.06$), organization ($F(3, 416) = 6.67, p = 0.00, \text{partial } \eta^2 = 0.05$), vocabulary ($F(3, 416) = 5.77, p = 0.00, \text{partial } \eta^2 = 0.04$), fluency ($F(3, 416) = 3.98, p = 0.01, \text{partial } \eta^2 = 0.03$), pronunciation ($F(3, 416) = 2.92, p = 0.03, \text{partial } \eta^2 = 0.02$), social/cultural appropriateness ($F(3, 416) = 6.84, p = 0.00, \text{partial } \eta^2 = 0.05$). However, again, the small partial $\eta^2$ statistics (0.02–0.06) indicated that the differences do not account for much of the variation.\(^{35}\)

Bonferroni post hoc tests revealed that there were significant differences among the levels in six of the seven linguistic categories. For functions, organization, and vocabulary, the Level 2 raters had scores that were significantly lower than each of the other, higher English proficiency raters (Levels 3, 4, and 5). For these three linguistic categories, there were no differences between any two other rater groups. In the functions category, the Level 2 raters’ mean rating was 0.45 lower than that of the Level 3 raters ($p = 0.00$), 0.38 lower than that of the Level 4 raters ($p = 0.01$), and 0.36 lower than that of the Level 5 raters ($p = 0.00$). In the organization category, the Level 2 raters’ mean rating was 0.41 lower than that of the Level 3 

\(^{35}\)The effect size (partial $\eta^2$) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).
The Level 2 raters also contributed to significant differences in social/cultural appropriateness and fluency, though there was not a significant difference with each of the other three levels. In the social/cultural appropriateness category, the Level 2 raters’ mean rating was 0.42 lower than that of the Level 3 raters \((p = 0.00)\) and 0.36 lower than that of the Level 4 raters \((p = 0.04)\). In the fluency category, the Level 2 raters’ mean rating was 0.28 lower than that of the Level 3 raters \((p = 0.04)\) and there was no significant difference with any other English proficiency group. In the pronunciation category, the significant difference occurred between Level 3 and 4 English proficiency raters. The Level 4 raters’ mean rating was 0.36 higher than that of the Level 3 raters \((p = 0.04)\).

In a departure from the pattern previously established, the Level 2 raters did not typically have the lowest mean ratings among the rater groups in the ILR Level 4/4+/5 exams (Table 29). In the case of pronunciation, the mean rating was the highest out of the rater groups. At a mean of 4.97, it meant that the Level 2 raters almost uniformly gave Level 5 ratings to the examinees, since no rating higher than Level 5 is possible. The Level 5 raters had the highest rating in the categories of organization, structures, and vocabulary; whereas, the Level 4 raters had the lowest mean ratings for these categories. Otherwise, there was not much of a pattern among the other ratings. All four English proficiency groups assigned a median score of 5 for structures and pronunciation. The Level 2 raters had the lowest set of medians, with most of the medians being at Level 4. The other level groups had mostly a median of 5 for the linguistic categories on the
high level exams.

Table 29. Linguistic Category Ratings for Level 4/4+/5 Exams: Rater English Proficiency

<table>
<thead>
<tr>
<th>Linguistic Category</th>
<th>Mean Rating</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L2</td>
<td>L3</td>
</tr>
<tr>
<td>Functions</td>
<td>4.10</td>
<td>4.43</td>
</tr>
<tr>
<td>Organization</td>
<td>4.30</td>
<td>4.53</td>
</tr>
<tr>
<td>Structures</td>
<td>4.53</td>
<td>4.53</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>4.30</td>
<td>4.46</td>
</tr>
<tr>
<td>Fluency</td>
<td>4.43</td>
<td>4.56</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>4.97</td>
<td>4.74</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td>4.17</td>
<td>4.47</td>
</tr>
</tbody>
</table>

L2 (n = 30: 6 participants x 5 exams); L3 (n = 70: 14 participants x 5 exams); L4 (n = 20: 4 participants x 5 exams); L5 (n = 30: 6 participants x 5 exams)

Table 30 showed the MANOVA and individual linguistic category results for the Level 4/4+5 exams. In this set, the descriptive statistics revealed a mixture of high and low means among the four English proficiency rater groups. Nevertheless, the multivariate tests results indicated that when all the linguistic category ratings were considered together, there was a significant difference of rater English proficiency on the linguistic category ratings: $F(21,$
\[402.56 = 2.55, p = 0.00, \text{ partial } \eta^2 = 0.11\]. Again, the partial \( \eta^2 \) indicated that English proficiency level accounted for a considerable portion of the variance, 0.11\(^{36}\).

Table 30. MANOVA for Linguistic Category Rating by Rater English Proficiency: Level 4/4+/5 Exams

<table>
<thead>
<tr>
<th>Linguistic category</th>
<th>df, error</th>
<th>( F )</th>
<th>( p )</th>
<th>Partial ( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall(^{37})</td>
<td>21, 402.56</td>
<td>2.55</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td>Functions</td>
<td>3, 146</td>
<td>1.77</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>Organization</td>
<td>3, 146</td>
<td>1.59</td>
<td>0.20</td>
<td>0.03</td>
</tr>
<tr>
<td>Structures</td>
<td>3, 146</td>
<td>0.86</td>
<td>0.46</td>
<td>0.02</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3, 146</td>
<td>1.56</td>
<td>0.20</td>
<td>0.03</td>
</tr>
<tr>
<td>Fluency</td>
<td>3, 146</td>
<td>0.97</td>
<td>0.41</td>
<td>0.02</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>3, 146</td>
<td>3.40</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Social/Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td>3, 146</td>
<td>1.30</td>
<td>0.28</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Interestingly, the largest difference between means from the descriptive statistics occurred in vocabulary, with the Level 5 raters being 0.38 higher than the Level 4 raters, yet this was not the linguistic category where a significant difference occurs. Instead, the ANOVA for

\[^{36}\text{The effect size (partial } \eta^2 \text{) for this data was interpreted as follows: 0.01 - 0.03 small; 0.04 - 0.09 medium; 0.10 and above large (Cohen, 1973).}\]

\[^{37}\text{Wilks' lambda}\]

177
pronunciation revealed a significant difference $F(3, 146) = 3.40, p = 0.02$, partial $\eta^2 = 0.07$, power $= 0.76$. The Bonferroni post hoc tests revealed that there was a difference that approached significance between the Level 2 and Level 3 raters’ mean ratings ($p = 0.066$), with the Level 3 raters’ mean being 0.33 lower than that of the Level 2 raters.

Overall, there was evidence that raters’ English proficiency level affected the linguistic category ratings assigned for exams overall and at each of the exam level ranges. It should be noted that many of the significant results that were discovered had relatively small effect sizes, indicating that they were not the only contributors toward the overall variance between the scores. These results speak to the answer of the fourth research question:

Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

Even though this data was not directly addressing the native speaker/non-native speaker dichotomy, it provided further evidence that what was more relevant than being a native speaker was how well the rater displays the skill that was being evaluated. Here, more significant results were discovered than when looking at the NS/NNS categories. Still, when differences were noted, the relatively small effect sizes meant that they were contributors, but only minor contributors. Some of the trends in the previous descriptive statistics recurred in these analyses as well. Almost uniformly, the English proficiency Levels 2 through 4 raters (NNS) raters had higher standard deviations than the Level 5 (NS) raters.
Chapter 6: Discussion

6.1 Overview

The fundamental issue of this study is whether or not native and non-native speaker raters rate English speaking tests differently from each other. This study follows a long legacy of studies looking at differences between native and non-native speakers, both from the fields of Language Testing research and in Second Language Acquisition research (Barnwell, 1989b; Birdsong, 1992; 2005; Coppetiers, 1987; Fayer & Krasinski, 1987; Galloway, 1980; Kim, 2009b; Ludwig, 1982; Montrul & Slabakova, 2003; Rossiter, 2009; Shi, 2001; White & Genesee, 1996; Zhang & Elder, 2011). Language Testing research contrasts native and non-native raters to answer a practical question: Can non-native speakers be validly used as raters? Second Language Acquisition research in this area focuses on the broader, theoretical issue of ultimate attainment: Can non-native speakers functionally acquire the language to the same extent that native speakers acquire the language? The implications of the results on both the rater and the examinees are discussed: What is the impact of sustaining or abandoning the native speaker concept in Language Testing? As mentioned in the literature review, the conclusions made by previous studies have had mixed results, with some finding significant differences between native and non-native speaker raters (Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979), and others finding no such significant differences (Barnwell, 1989b; Eckes, 2008; Kim, 2009a; Kim, 2009b; Ludwig, 1982).

This dissertation intends to examine the native/non-native speaker issue in a new context: language testing within the US Federal Government. Language testing in the US Federal Government context offers better control over some rater/participant variables and addresses a
greater range of language proficiency levels by moving beyond the lower levels of language proficiency typically encountered in literature and into professional, advanced levels of language use. Moreover, this study’s rater participants were actual raters used by the US Federal Government, and the exams rated by them in this study were exams actually administered by the US Federal Government. Additionally, this study explores the construct of a native speaker at a deeper level, by challenging the assumptions underlying that label and considering the implications for testers who have not earned or been given the native speaker label. In the end of this study, the results and conclusions of this study provide practical recommendations for selecting speaking test raters and contribute to the discussion on the native speaker construct.

The current study compares native and non-native speakers’ ratings of English oral proficiency exams given at the FBI. This study’s data include both the final ratings assigned to the twenty-five exams as well as their linguistic category sub-ratings. Raters are grouped by three criteria: (1) the traditional native speaker/non-native speaker dichotomy, (1) the raters’ English proficiency ratings, and (3) the raters’ native (first) languages. The second and third groupings derive from the idea that a native speaker can be seen either as a highly proficient, ideal speaker of the language or alternatively as a person who learned the language as a first or childhood language from birth (Davies, 2003). This analysis resulted in the following research questions:

1. Do native and non-native speaker raters assign comparable ratings on speaking tests?
2. Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?
3. Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?

4. Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

Since previous research is inconclusive, finding differences between native and non-native speaker raters (Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979) as well as no differences between the rater groups (Barnwell, 1989b; Eckes, 2008; Kim, 2009a; Kim, 2009b; Ludwig, 1982), the null hypothesis for each research question is adopted. There will be no differences between the rater groups (i.e., the rater groups’ ratings will be comparable).

There was no conclusive evidence that native speakers, defined by any definition, gave significantly different final ratings than non-native speakers with any large effect size. However, there appeared to be evidence of significant differences in the ratings given to the linguistic category sub-ratings between the rater groups, implying that there were differences in how the rater groups were analyzing the language.

6.2 Explanation of findings

6.2.1 RQ 1: Native speaker and non-native speaker rater comparison

The first and overarching issue to be discussed is whether or not NS raters and NNS raters assigned significantly different holistic final ratings. In a sense, this question is more important than the evaluation of the linguistic category ratings because final ratings are the score of record, meaning that this is the score that is reported and used to make decisions about job qualifications, promotions, and incentive awards of US Federal Government employees. The
linguistic category ratings are assigned as a part of the analytical process of arriving at the final rating and are not publicly reported. The final ratings, therefore, have much greater consequence for the examinees than the linguistic category ratings, which are only intermediate steps in the rating process. The issue of final rating comparability is the issue addressed in the first research question.

1. Do native and non-native speaker raters assign comparable ratings on speaking tests?

To answer the first research question, this study divided the thirty raters into two groups, native and non-native speaker raters, with six English testers in the native speaker group, and twenty-four testers of other languages in the non-native speaker group. There was no statistical difference between the ratings of native and non-native speakers’ final ratings of all twenty-five exams. In fact, there was little average difference in the ratings at all; the native speaker raters assigned a mean rating of 3.45 and the non-native raters assigned a mean rating of 3.44. When the ratings were grouped by exam level (Level 2/2+, Level 3/3+, and Level 4/4+/5), there were again no significant differences found between the native and non-native rater groups at any of the exam levels.

One explanation for the result that may be offered is that there is no fundamental difference in the English speaker rating abilities of native and non-native speakers. While that very well may be the case, an exploration of the unique characteristics and intervening factors of this data set is warranted. First, the rater participants are not average, randomly selected native and non-native speakers. They are highly proficient, trained raters, and most of them are bilinguals at least. Second, the process of rating according a detailed set of criteria, such as the
ILR Skill Level Descriptions, narrows raters’ reasonable choices for ratings, which leads to increased rater uniformity.

First, careful consideration should be paid to the characteristics of the raters other than being a native or non-native speaker of the language. The raters are not defined simply by whether or not they came from the native speaker population; they are also all members of the trained rater population. The raters have to undergo over one hundred hours of training and prove their skills through a rigorous certification process before being allowed to conduct official tests. If new raters are not able to show uniformity with previous raters, they are not used. Individualized interpretations of the ILR Skill Level Descriptions are trained out of the rater pool, as the goal is rater consistency. Rater consistency is sought for not only raters within a language group, but also across language groups.

It is not only important that the English exam raters are consistent and interchangeable, but also that the interpretation of a particular level is consistent across languages. For example the same principles apply to being a Level 3 in English as being a Level 3 in Vietnamese. The training also means that the raters are practiced at dissecting and analyzing language. All the participants are used to thinking about all aspects of language, including grammar, fluency, vocabulary, organization, etc., whereas; in contrast, many native speakers, and non-natives who have learned the language in an immersion setting, may have had little formal training in these concepts. The training is determinative to how raters would actually evaluate the tests, it is designed to minimize differences occurring naturally in the raters.

The other characteristic that the raters in this study have that differentiates them from the average native or non-native speaker is that they are all at least ILR Level 5 in one language. The raters’ high level of proficiency means that they also have a highly developed cognitive ability,
which may give them a better ability to notice subtle aspects of the examinees’ language abilities and inabilities. Moreover, many of the raters work in fields related to language, such as translation, language teaching or writing. This background gives the raters an awareness of language features and norms that the average speaker may not have.

Additionally, the non-native speakers are bilinguals, as they have a Level 5 proficiency in their first/native languages, and at least a Level 2+ proficiency in English. The non-native raters have experience thinking about how languages work in the process of becoming bilingual, heightening their awareness of linguistic features. The non-native speakers also have been living in the United States for a long time (many of them for much longer than they lived in a country where their first languages were spoken natively). At such a point in one’s life, it is difficult to determine which language is actually their dominant language, as it is likely that they spend the majority of their time interacting in or processing English. The dominance of the first or second language can impact the comparability of native and non-native speakers, according to Muñoz and Singleton (2011).

The favorable comparison made between the L2 proficiency of early L2 acquirers and that of native speakers of the language in question is explicable in many cases by the fact that these acquirers’ erstwhile L2 has in fact switched its status to that of L1—understood not in the classic sense of the language of infancy, obviously, but in terms of dominance, functionality and users’ perceptions. There is a sense, therefore, in which making a comparison between the proficiency of such language users in their dominant language and that of native speakers is not a matter of comparing L2 attainment with an L1 baseline, because for such language users their dominant language for all practical
purposes is ITSELF their L1. As for late acquirers, the continuing dominance (and influence) of their language of infancy, if viewed in the terms proposed by Jia & Aaronson (2003) and/or by Flege (Flege at al., 1999), emerges in most cases as an age-related phenomenon, but not strictly a maturationally determined phenomenon.

Munoz and Singleton assert that early L2 acquirers may actually have the L2 as their dominant language, making it essentially their L1, and therefore should not be expected to be significantly different from native speakers. However, their argument assumes that a person can only have one dominant language. The speaking proficiencies of the non-native speaker raters seem to contradict this principle, as these adults have maintained a Level 5 proficiency in their native languages, or their L1s, while acquiring English as well as (or better than) the average native speaker of English, as determined by their English proficiency ratings. It is unclear which language is internally dominant, but the non-native speakers certainly spend the majority of their time in English and may have become balanced bilinguals, with both languages being equally dominant.

Where the data suggested any differences between the native and non-native speaker raters’ ability to evaluate the language samples, the results from the first research question nonetheless suggest that the differences can be overcome through training or having raters that are balanced or high-achieving bilinguals. As a result, the fact that a person is a non-native speaker does not reflect a low level of proficiency in a second language, even if it is acquired later in life. The fact that the non-native speakers performed comparably to the native speakers suggests that this is not the rater characteristic that is important in the selection of raters. Instead of focusing on whether a person is a non-native speaker, more relevant may be the fact that the
raters have completed extensive training, have high cognitive awareness of the languages, and are proven advanced speakers. These qualifications are also based on accomplishment or ability measures, instead of biography, making the measures less prejudicial.

Since the results of the first research question indicate that there is no significant difference between native and non-native speaker rater groups in the ability to evaluate English speaking tests, it may be natural to jump to the conclusion that the groups are therefore equal and comparable. However, in order to determine whether or not “no significant difference in final ratings” is synonymous with interchangeability, a closer evaluation of the factors important to being a rater must be accurately examined.

The first important measure of rater quality is inter-rater reliability. In this study, the native speaker raters have an inter-rater reliability coefficient of 0.77\(^{38}\), measured using Krippendorff’s alpha. Although this coefficient is below the standard of 0.8, it is above the minimally acceptable mark of 0.7 (Krippendorff, 2004; Lombard et al., 2003). The non-native speaker raters, however, have a Krippendorff’s alpha coefficient of 0.59, well below the level of acceptability. It would be risky therefore to accept a member of the non-native speaker group into an acceptable pool of raters without determining that individual rater’s inter-rater reliability with already admitted (i.e., native speaker) members of that pool. If both native and non-native

\(^{38}\) The native speaker raters in this study regularly undergo indices of inter-rater reliability as a part of the program’s ongoing quality assurance process. Typically the English raters have a Krippendorff’s alpha coefficient of 0.92 or higher. Although this study was constructed to replicate real-world rating processes, the English raters usually administer and rate the test simultaneously, meaning that their attention is divided between the two tasks. This deviation from normal protocol is potentially the reason that their inter-rater reliability coefficient is lower in this study than in regular practice.
speaker raters were included, the inter-rater reliability coefficient would be 0.62. The native speakers do appear to be more homogeneous than the non-native speaker raters.

Standard deviation is another measure by which the homogeneity of the native speakers and non-native speakers can be compared. Native speaker raters’ ratings have a smaller standard deviation (0.79) than the ratings of the non-native speaker raters (0.90), indicating that their ratings are distributed more closely around their group mean. This may also indicate that the native speaker raters rate more homogeneously than the non-native speaker raters. Interestingly, the native speakers have a slightly higher standard error than the non-native speakers, 0.06 versus 0.04. This may be due to the fact that the non-native speaker group is larger than the native speaker group, but overall the difference of 0.02 in standard errors is very small, indicating that it is very unlikely that ratings of a given individual performance by each group would differ even by a half (plus) level.

These inter-rater reliability statistics suggest that the native speaker raters rate more homogeneously than the non-native speaker group. However, it is certainly not clear that the native speaker group should be considered homogeneous, especially with an inter-rater reliability coefficient of 0.77. Even though the native speaker group appears to be more consistent and homogenous in their ratings, it does not make the differences statistically significant. Overall, the analysis of this data set shows no difference between the two groups.

6.2.2 RQ 2: Comparison by rater English speaking proficiency

The first research question deals with the native speaker/non-native speaker dichotomy. When establishing such a dichotomy, one must consider what it means to be a member of each group and whether or not the dichotomy is relevant or justified. Research has already indicated
that, at a minimum, the native speakers of a language are not homogeneous in their ability to perform on English language tests (Hamilton et al., 1993). Also, there is little agreement as to what makes a native speaker: being born in an area where the language is spoken, being raised there, being educated there, and until what age? In fact, as Davies states, it is not even certain whether or not the native speaker is a myth or reality (2003).

Not surprisingly, previous research indicates that non-native speakers are no more homogeneous a group than native speakers (Hamilton et al., 1993). Nevertheless, studies that pit native speaker groups against non-native speaker groups treat non-native speakers as if they are a unified group with common characteristics. Still the questions remain, are non-native speakers really unified or is their common trait merely the fact that they are not native, a characteristic that may or may not have any quantifiable definition?

In the literature review are discussed two ways of defining what it means to be a native speaker, the native acquirer and the ideal speaker (Davies, 2003; Maher, 2001; Paikeday, 1985). By consequence these two methods can be used to define what it means to be a non-native speaker, a person who is not a native speaker. The first method for labeling a person as a native speaker derives from the idea that native speakers have a high or even ideal level of speaking proficiency (Davies, 2003). This notion leads to an analysis of rater groups not along the native speaker non-native speaker dichotomy, but rather analyzing the raters’ speaking proficiency, in this case English speaking proficiency. For this analysis, the raters were divided according to their English speaking proficiency ratings: Level 2/2+ (n=6), Level 3/3+ (n=14), Level 4/4+ (n=4), and Level 5 (n=6). This analysis of rater language answers the second research question.
2. Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?

Unlike the comparison in the native speaker/non-native speaker dichotomy, the English-speaking proficiency of the raters does make a significant difference in the overall ratings, although as mentioned before, these differences must be interpreted with caution. Across all exams, a significant difference is found between the ratings of the four English speaking proficiency groups. However, there is a low eta squared, or effect size, for this difference (0.02). Significant difference was again found in the Level 3 range exams, although again with a small effect size of 0.04. One of the reasons that more significant results are discovered in the Level 3 exams may be that the majority of the exams are at that level. There are 14 exams rated at the Level 3, and 11 exams rated at Levels 2, 4, and 5 combined.

Interestingly, no significant differences across ratings are found in the Level 2 exams or the Level 4-5 exams. The lack of significant difference in the Level 2 exams may be explained by the fact that all of the different rater groups had enough competence in English speaking proficiency to be able to accurately rate exams at this level. The lowest English proficiency raters have an English speaking proficiency rating of ILR Level 2+. It appears that non-native speaker raters even at the lowest proficiency levels in this study are able to apply their rating skills from their native language to assess examinees at this level in the same way as native speaker raters.

In the Level 4-5 exams, a different explanation is posited for the lack of significant difference. It is possible that there is a ceiling effect at the top of the scale, meaning that there are fewer possible choices among which the raters can select. Although it is possible that some raters could rate overly harshly and giving a low rating and causing additional variation in the ratings,
raters cannot err by overrating at these high levels. Reducing the number of options for the rater means that differences between rater groups are less likely.

Even in mid range Level 3 exams, there are only a few plausible choices for a well-trained rater. At all levels, the rater should be able to pick up on clues by the level of the tasks being offered to the examinee as to what the original testers think that the final rating should be. As the rater listens to the exam, he considers the level at which the elicitations (usually questions) are presented to the examinee. Raters are trained to be able to identify the task of any question posed, and therefore, have an anticipated level of response in mind for each question before the examinee begins to respond to the question. In effect, the rater then chooses between two possible ratings: at level, or below level. The elicitation/response process continues throughout the test until the raters establish the level at which the examinee can display a consistent performance, and the level above that where the examinee cannot display consistent performance.

The rater training contrasts to the practices in previous studies, where raters were given a familiarization of the scale or only a copy of it without explanation at all. Training the raters controls the variable of the raters’ idiosyncratic beliefs about language and language proficiency that may have been acquired from their environment and experience learning or using the language. Trained raters should give more uniform evaluations, making them interchangeable and less prone to bias. Using trained raters is more appropriate for the research questions, because the question is whether or not non-native speakers can rate like native speakers if properly trained. The result of these circumstances is that the bar is set high for the examination of differences, as the training is likely to, if anything, suppress differences present between the rater groups, so significant differences may be harder to identify. In reality, a rater should only
be choosing between a half to a whole level higher or lower, meaning that the rater is actually selecting between three to five possible final ratings. When the exam level is at the top of the rating scale, the raters are selecting among about three possible ratings, making it more likely that there is agreement among the rater groups.

When differences are found between the English speaking proficiency rater groups, Bonferroni post hoc tests identify the pairs of groups that significantly differed. Across all exams significant differences are found only between Level 2 raters and Level 3 raters. The Level 2 raters are the strictest of the four rater groups, and the Level 3 raters are the most lenient of the four rater groups. It is difficult to tell whether or not the significant difference is a result of the fact that the Level 2 raters are exceedingly strict, or that Level 3 raters are exceedingly generous, or potentially a combination of the two factors. When the Level 3 exams are further analyzed through Bonferroni post hoc tests, differences appear between the Level 2+ and the Level 3 rater groups and between the Level 2 and Level 4 rater groups. In both cases, the Level 2 rater group gave lower ratings than the Level 3 or Level 4 rater groups. The Level 2 raters have the lowest mean rating of 3.22 and the Level 3 raters again have the highest mean rating of 3.57. The Level 4 raters’ mean rating is 3.51.

The FBI’s policy of requiring raters to be native speakers and have Level 5 proficiency reflects the philosophy that “it takes one to know one.” There has previously been the idea that only a Level 5 speaker would be able to discern the subtle differences between Level 4 and Level 5 examinees. This philosophy is not supported by the current study’s results. The Level 4 non-native speakers gave almost identical mean ratings to the Level 5 native speaker raters, 3.46 and 3.45 respectively. There are never any significant differences between the Level 4 and Level 5 raters. It is often the case that the Level 3 raters are the most generous in their ratings, and the
Level 2 raters are the strictest. The tendency to be strict may be a result of overcompensation for low language ability. These raters made comments on occasion that they had a hard time rating because they did not know English well enough. The Level 3 raters may be lenient because they were not able to pick up on errors made committed by the high-level examinees.

One of the reasons why it may be more relevant to define non-native speakers by their English proficiency rather than their first language is because there is more direct construct relevance between the trait of their English speaking ability and the task of evaluating an examinee’s English proficiency than the relationship between their first language and the language being evaluated. As mentioned previously, the assumption at the FBI is that a person must be at a particular speaking proficiency level to be able to evaluate that level and any level below that level, based on the idea that competence precedes performance. If the person is able to display performance at a particular level, than he or she has the competence at that level. The NNS rater’s secondary bilingual competence certainly impacts his or her ability to evaluate a language such as English, even if it is not their strongest or first language.

6.2.3 RQ 3: Comparison by rater first language

The third research question addressed the rater’s native, or first, language, as native speaker can be defined as a person who was born into an environment in which that language was spoken and accordingly raised speaking that language (Davies, 2003). This way of defining the native speaker leads to the analysis where the raters are categorized by the first language that they learned.
3. Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?

For the other analysis, the effect of first language of the raters on their ratings is examined. Here, significant differences were discovered not only in the ratings of all tests combined, but also at each level of exam (Level 2, Level 3, and Levels 4-5). Again the partial eta squared, or effect size, for the analysis of all exams combined is very small at 0.02. The partial eta squared in the Level 3 exams is also small at 0.05. However, in both Level 2 and Level 4 exam groups, the partial eta squared is 0.12, which is large enough to note. A partial eta squared of this size represents a large enough portion of the range of all possible ratings to potentially change a final exam by a plus level.

The Bonferroni post hoc t-tests reveal that there is a significant difference between the Arabic and Mandarin native language groups, with the Arabic raters giving higher scores and the Mandarin raters giving lower scores. However, none of the ratings of the non-native language groups are significantly different from those of the English group. In the Level 2 and Level 3 exams, the Arabic testers seem to be particularly lenient, giving higher ratings than the other first language groups, but significantly higher than French and German raters on the Level 2 exams and Farsi raters on the Level 3 exams. In the Level 3 exams, the Farsi group appears to be particularly strict, as their ratings were significantly lower than the Arabic and Vietnamese rater groups. Perhaps most interestingly, the German rater group seems to be particularly strict in the Level 2 exams, whereas they are lenient when rating Level 4-5 exams. In the Level 4-5 exams, the German raters assign significantly higher scores than the Mandarin raters.
Although analyzing the native and non-native speaker data by rater first language shows more instances of significant difference among ratings than analyzing it by the English speaking proficiency of the raters, the post hoc tests show no clear pattern of which group may or may not be more biased, however thinking about the language families of the groups gives more insight into the results. The Arabic and the Mandarin language group mean ratings were significantly different, with the Arabic raters’ mean being higher than the Mandarin raters’ mean.

Interestingly, the language group with the mean rating closest to the mean of the entire group (3.44) was the English group (3.45). English is an Indo-European, Germanic language (Lewis, 2012). The three language groups with ratings that are the most different from the English group are the three languages that were not Indo-European (Lewis, 2012). These three language groups have mean ratings that are approximately 0.2 different from the English raters. The Arabic group is 0.22 higher than English. The Vietnamese group is 0.19 higher than English. The Mandarin group is 0.19 lower than English. The three other language groups that have mean ratings that are approximately 0.1 different than the English group are all Indo-European languages, but not Germanic languages (Lewis, 2012). The French and Farsi rater groups each have mean ratings at 0.11 below the English rater group, and the Spanish raters’ mean is 0.09 below the English rater group. The one group that has a mean rating that was almost the same as the English raters was the German group, which was only 0.03 higher than the English raters. It seems, therefore, that the closer that the language is related to English, the more similar the ratings are to the English ratings. It should be noted that there was no significant difference between the English rater group and any other language group.

When all exams are analyzed level by level, there is no trend as to whether or not any first language group is particularly strict or lenient. In each exam group, different language
groups are closest or most distant to the English raters. For example, in Level 2 exams, the Spanish group’s mean is closest to the English group, but the German and French groups are the furthest from the English means. In Level 3 exams, the French and Spanish groups’ means are closest to the English group, but the Arabic and Vietnamese groups are the furthest from the English means. In Level 4-5 exams, the Farsi, Arabic, and Vietnamese groups’ means are closest to the English group, but the Spanish and Mandarin groups are the furthest from the English means. Since the pattern of distance between language families does not hold through the exam levels, a better explanation of differences among the non-native speaker group may be the English proficiency of the raters.

Additionally, whether or not speaking proficiency is the ability that most directly relates to the task of rating has been questioned. It is the job of the rater to listen to an exam and give an evaluation. It is therefore sometimes argued that the raters’ listening comprehension, rather than their English speaking proficiency, should be measured. However, this option is not appropriate because most measures of listening comprehension assess whether the examinee is able to absorb content, not whether they are able to assess someone’s performance in speaking. It seems then that it is more appropriate to have measures of the raters’ English speaking proficiency in selecting qualified rater candidates.

6.2.4 RQ 4: Linguistic category differences among rater groups

To answer the fourth research question, the raters’ linguistic category sub-ratings were analyzed instead of the final holistic ratings in order to determine if there were differences in how the rater groups were analyzing the English exams. The linguistic categories included
functions, organization, structures, vocabulary, fluency, pronunciation, and social/cultural appropriateness. This analysis addressed the fourth research question.

4. Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

When this study analyzed linguistic category ratings using MANOVAs to compare native speaker and non-native speaker raters, significant differences appeared. In fact they appeared not only across all exams, but also at each exam proficiency level. However, the effect sizes are mixed across the levels. There is a large effect size in the Level 2 exams and the Level 4-5 exams of 0.11 and 0.10 respectively. In the Level 3 exams, there is an effect size of only 0.05. In the overall exams the effect size 0.04, which is relatively low. A look at the post hoc tests shows that differences in ratings of the linguistic category of organization are significant at all levels of exams, but that there is no other linguistic category on which native speaker and non-native speaker raters differ consistently.

In the analysis of all exams together, organization approaches significance at $p = 0.08$. Organization appears again in the Level 2 exams analysis as a significant post hoc test. However, at Level 3 fluency is significant and at Level 4/5 there is technically no significant post hoc test, although vocabulary is the category that most closely approaches significance at $p = 0.09$.

Even though significant differences appear in the linguistic category ratings, they appear with mixed effect sizes and without a consistent trend in which linguistic category is causing the overall effect emerging among them. Overall, the differences that appear in the linguistic
categories are washed out in the final ratings and therefore at some level they do not hold consequence for the final result of the evaluation.

However, in Language Testing, the final rating is not the only matter of interest. Not only is it important for the raters to assign the correct rating, but also for them to assess the right ability; in other words, raters should also be rating the same construct (Bachman, 2007). For the raters to have the same construct of speaking proficiency means that they are valuing the same attributes of an examinee’s speech and have the same idea about what it means to be at each of the ILR Levels.

Hence, extensive training of the raters on the ILR Skill Level Descriptions proves to be important so that raters are comparable. Raters should function interchangeably with each other, so that an examinee would have the same testing experience no matter which two raters were assigned to the examination. The fact that there are significant overall differences in the linguistic category ratings provides some evidence that the rater groups are approaching the construct of rating in different ways. These rating differences indicate that the native speakers and non-native speakers may be giving different weight to a particular linguistic category. Raters who have different constructs of what they are measuring are not measuring the same ability.

When the linguistic subcategories were again examined with the non-native speaker raters divided by English proficiency, significant differences in linguistic subcategory ratings appeared at every level of analysis. Again, the significant results had mixed effect sizes in the different levels of exams. There was a significant effect size in the Level 2 exams and the Level 4-5 exams of 0.08 and 0.11, respectively. In the analysis of all exams together, there was only an effect size of 0.04. In the Level 3 exams, there was an effect size of 0.06.
As in the analysis of the final ratings, the analyses of the linguistic category ratings could be affected by the fact that the majority of the exams that were rated fall into the Level 3 range (n=14). Again here, the analysis of the Level 3 exams reveals more evidence of differences between the rater language groups than exams at the lower (Level 2) or higher levels (Levels 4-5).

The post hoc tests reveal many linguistic categories that are significantly different between the rater proficiency groups. In the analysis of all exams and the Level 3 exams, significant differences occur in all of the linguistic categories except structures. In the Level 2 exams, the differences only occur in fluency, organization, and vocabulary. In the Level 4-5 exams, the only linguistic category that is significant is pronunciation.

It is interesting that significant differences appear in every linguistic category at one point of analysis or another except for structures. The category of structures refers to grammar, both at the word and sentential levels (FBI, 2009). Two of the characteristics that Davies (2003) attributes to native speakers are intuitions of productiveness and acceptability of the grammar of his or her own idiolect and intuitions of the standard language grammar as distinct from idiolect grammar. Yet the non-native speaker raters in this study, at any level or by any definition, never have significant differences with their native speaker counterparts in grammar ratings. This finding echoes the results of Porte’s (1999) research on error gravity in essays. He concluded that there is little difference in the perception of errors in the native and non-native speaker groups. There is no evidence in these results to support Kobayashi’s (1992) conclusion that native speakers are harsher on grammar or the opposite conclusion, that non-native speakers are harsher on grammar (Fayer & Krasinski, 1987; van Maele, 1994).
White (2003) discusses how first language grammars constrain second language acquisition. In the case of the non-native raters, who come from a wide variety of first language grammars, there does not appear to be any impact from their first language on their ability to judge grammar in a second language. This further supports the idea that the non-native raters are in fact bilinguals with well-developed English grammars. The non-native speakers would be classified as late language learners, with most of them beginning beyond the critical period, yet they had no problem assessing structures in this study, in contrast to what DeKeyser and Larson Hall (2005) would have predicted. A possible explanation for this ability by non-native speakers to evaluate structures is that judging grammar may be less subjective than other linguistic features. Grammar has rules of acceptability that are definable and standardized. Other linguistic features, such as pronunciation, social/cultural appropriateness, organization, etc., are harder to quantify and are more impressionistic.

One of the linguistic categories in which is more subjective differences between native and non-native speakers would be social/cultural appropriateness. This category has to do with issues of pragmatics and politeness (FBI, 2009). Studies conducted by Hinkel (1994) and Brown (1995) found that native speakers were more aware of social/cultural appropriateness concerns and therefore would be more critical of examinees’ performance under this category in their ratings. However, the current study does not reach the same finding. The only place where there was a significant difference in the area of social/cultural appropriateness was in Level 3 exams, between the Level 2+ and Level 3/3+ raters, which are both non-native speaker groups. There is no indication that the Level 5/native speaker raters’ scores are more or less accurate than the Level 3 or Level 4 non-native speakers. This may be due to the fact that these high level non-native speakers have a high social/cultural ability themselves and are therefore able to accurately
rate this social/cultural abilities. The non-native raters’ high proficiency scores may also be indicative of the fact that they have lived in the United States for a long time, and have assimilated and grown to understand American culture.

When the final ratings were analyzed for the raters’ English proficiency levels, the Level 2 raters reported significantly low ratings in each instance. This trend continued in the linguistic category ratings. When all exams were combined for the analysis, six out of seven linguistic categories had significant differences between raters of different proficiency levels, and the Level 2 raters had significantly lower ratings in five of those categories, excepting structures and pronunciation. In fact, for three of the categories (functions, organization and vocabulary), the Level 2 raters assigned ratings that were significantly lower than each of the other rater groups. The same pattern appeared in the Level 3 exams analysis. The Level 2 raters were also implicated in the Level 2 exams analysis, again with lower ratings in fluency, organization, and vocabulary. In the Level 3 and Level 4-5 exams, the Level 2 raters had significantly higher ratings than the Level 3 raters in pronunciation.

Based on these results, as well as the results from the final ratings analyses, Level 2 raters are not likely to be successful in rating exams at all levels. In many linguistic categories, Level 2 raters give stricter ratings than many of the other rater groups. The Level 2 raters seem to have a stricter perception of examinees at all levels, except for pronunciation in Level 3-5 exams. Perhaps this is because the Level 2 raters’ English language proficiency is not sufficient to allow them to make the rating distinctions that they need to make. As a result, it is possible that they compensate by rating low, erring on the side of caution.

One of the principles taught in speaking proficiency rater training is that if a rater is unsure about how to assess a performance—that is, if a rater has trouble deciding between two
ratings—then the rater should err on the side of caution and assign a conservative rating (FBI, 2009). This way the rater can say with 100% assurance that the examinee merits at least the rating assigned, a statement that is more reliable to the agency using the rating for high-stakes decisions. The deviation from the pattern of the Level 2 raters’ conservative ratings is in the area of pronunciation.

Although the results mostly indicated that there were not differences or that the differences found had small effect size between the native speaker/Level 5/English raters and the other rater groups, two measures of rating consistency, inter-rater reliability and standard deviation from means, indicated that the ratings of the native speaker group were somewhat more stable than the ratings of the other groups. The native speaker group always had a higher inter-rater reliability coefficient than any other group, no matter how the rater groups were divided. Additionally, the native speaker/Level 5/English rater group’s ratings almost always had the lowest standard deviation from the mean, perhaps because there was more variation among the non-native speaker raters.

6.3 Application to previous literature

The literature review laid the groundwork for the issues surrounding what a native speaker is, and how it is used in Language Testing research as well as Second Language Acquisition research. The native speaker concept has social and political consequences. The current study validates the two definitions of the native speaker: the ideal speaker of the language and the native learner (Davies, 2003, 2004, 2011; Paikeday, 1985), but difference in the definitions indicates that there is not a universal understanding of what it means to be a native speaker (Escudero & Sharwood-Smith, 2001). There is a burden, therefore, to use the term
appropriately and specifically to avoid prejudice and bias (Shohamy, 2001). By explicating what a tester or researcher means when they report using native speakers in their research or practices, one can avoid incorrect assumptions about judgments of language proficiency being the same as evaluations of native identity (McNamara, 2007) and the fairness of the test can be maintained (Elder, 1997; Kunnan, 2000, 2004).

6.3.1 Application to Language Testing research

This study, like several recent Language Testing studies that compared native and non-native speaker raters, contradicts previous research that found differences between native and non-native speakers’ abilities to accurately rate (Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979). Most of these previous studies compare the raters groups using descriptive statistics and not inferential statistical measures. The results in this study complement more recent research that has been conducted on the comparison of native and non-native rater groups that concludes that there is no difference in final ratings (Brown, 1995; Hill, 1996; Kim, 2009b; Ludwig, 1982; Zhang & Elder, 2011). Although Barnwell (1989b) does not find any overall differences between native and non-native speaker groups, he finds that native speakers are more critical on examinees, particularly on those who have higher proficiency levels. This is also a finding in Galloway’s (1980) and Hill’s (1996) studies. The results of the current study do not indicate that the native speakers are harsher either in final ratings or in linguistic category ratings. If anything, the evidence shows that non-natives with lower levels of language proficiency regularly evaluate examinees more harshly. Hill (1996) also finds non-natives to be more consistent in their ratings. Measures of inter-rater reliability and the standard deviation from the means indicate that the native speakers were more consistent than the non-native speakers.
Shi (2001) had results that were consistent with this study. Shi found native speakers to be more positive and non-native speakers to be more focused on what the examinee did wrong. Additionally, she found the holistic scores to be the same between the two rater groups, but the analysis to be different. Kim (2009) also found the overall ratings to be the same between the native and non-native speaker rater groups, but that the rater groups do not handle the samples the same way. She found native speakers to be more elaborate on pronunciation, grammar and accuracy. The raters in this study did not find a significant difference in any of those linguistic features. Zhang and Elder (2011) found that the non-native speakers focused on linguistic resources in their rating process, although there were no overall differences in final scores between the native and non-native rater groups.

The conclusion that may be suggested based on the results of this study is that it is appropriate and justified to use non-native speaker raters to evaluate speaking proficiency exams of English at Levels 2 through 5. Non-native testers would, of course, still need to pass the same training and qualifying exams as the native speaker testers, and be able to meet any other prerequisite qualifications, such as a minimum language proficiency in the test language. The results do mean that there is no categorical basis for the exclusion of a non-native speaker rater candidate from applying and attempting to qualify as an English rater.

The current examination of differences between native and non-native speaker raters aims to fill in gaps left by previous studies examining how rater nativity, potentially redefined as rater speaking proficiency or rater first language, affects the ratings assigned on oral proficiency exams. The rationale behind the specialized structure of this study is two-fold. First, the study is intended to test whether the conclusions of previous research are valid when applied to a high-stakes, large-scale operational speaking testing context instead of a classroom assessment.
environment. Second the design of the study is meant to minimize intervening factors that may contribute to finding differences between the rater groups, and test, as directly as possible, whether or not those intervening factors could be overcome. Intervening rater variables might include insufficient language proficiency, the lack of an ability to evaluate language, or fuzzy understanding of the rating scale.

Another improvement over previous studies is the range of the examinee levels included in the sample to be rated. Since much previous research has been conducted on performances of language students, it cannot be determined whether the same results would be found in the testing of adult, high-proficiency learners, who have acquired the language through a variety of methods. Therefore, speech samples for this study were chosen at the highest levels of language proficiency, ILR Levels 2-5, where flaws in speaker performances are not as obviously perceivable as at lower levels. It is possible that in studies where there was no difference found between native and non-native speaker raters it was due to the fact that flaws in the examinee’s speech were obvious, often interfering with comprehension. In many cases, the flaws in the examinees’ speech from performances of ILR Levels 2 - 5 are quite subtle, and comprehension of the examinee’s main message is not impacted. It would therefore take a more attuned, trained ear to discern those differences. Interviews from non-native speakers were chosen from test-takers who have high levels of language competence and performance (ILR Levels 2+ - 4+) in English, which are equivalent to the levels typically accomplished by lay native speakers. The use of high-level samples is more likely to uncover differences between native and non-native speaker rating abilities because the task is more difficult.

The current study concludes that being a native speaker, and its complement: being a non-native speakers, is not a sufficient reason to be able to perform the duties of a rater
accurately, in contrast to previous studies that drew the contrary conclusion. These previous studies may have drawn the contrary conclusion because the only qualification examined prior to the selection of raters was whether they were native speakers (Escudero & Sharwood-Smith, 2001). Additionally, in previous studies, there was little to no consideration of the candidate’s ability to perform the rating task reliably (Barnwell, 1989b; Brown, 1995; Fayer & Krasinski, 1987; Galloway, 1980; Hill, 1996; Kim, 2009b; Ludwig, 1982; Ross, 1979; Zhang & Elder, 2011). Since being a native speaker was the only qualification for rater considered, it is understandable how finding no significant difference between the ratings of native and non-native speakers led to the acceptance of both speaker groups outright as raters. These previous studies considered one rater variable in isolation—the rater’s nativity—and as a consequence it led to a misinterpretation of results.

The decision of whether to include a rater candidate within a pool of raters is not a decision that should be based on whether the candidate is a member of a particular group such as a native speaker or non-native speaker. A rater should be considered for inclusion based on his or her individual combination of qualifications. Likewise, because the simple fact that a group has a high inter-rater reliability statistic does not mean that any individual member of that group should automatically be included in a rater cadre. If an individual rater candidate is to be admitted into a rater pool, they must align with the dynamic of the current raters (Bachman, 1990).

The results did show that the native speaker/English/Level 5 rater group was more homogeneous than any of the other rater groups when considering their inter-rater reliability statistic as well as the standard deviation of their ratings. However, the results also support the opportunity for non-native speakers to be allowed the opportunity to join the active rater pool for
English, assuming that they meet the certification requirements for English raters. Therefore, all of the rater participants have demonstrated that they have a certain level of evaluative competence. They have the ability to listen to a language sample, analyze it for its strengths and weaknesses, describe it in terms of linguistic categories, and give it an accurate holistic rating. Evaluative competence must be demonstrated at the end of the rater training workshop and before raters are permitted to give actual ratings on tests. Additionally, the FBI conducts regular, random quality control exercises to evaluate the consistency of the ratings being assigned. Interrater reliability is calculated, and the testers are given feedback and retraining when needed.

Previous literature on raters from different first language backgrounds indicates that raters favor examinees with accents similar to their own (Gass & Varonis, 1984; Major et al., 2002) or familiar accents (Carey, Mannell, & Dunn, 2011; Fayer & Krasinski, 1987). The results of the analyses of the rater’s grouped by native language at first indicated that similarity of first language to the language being assessed did impact the reliability of the ratings; however, it does not mean that these raters favored examinees with similar first languages. Even so, this analysis was not conclusive, as this pattern of results was not sustained when the exams were analyzed by level. Similarly Elder and Davies (1998) showed that rater first language distance to the assessed language had some impact on the ratings, but not in a conclusive manner as it cannot be separated from other rater characteristics. Hamp-Lyons and Davies (2008) and Johnson and Lim (2009) also had similar research studies finding that raters’ bias by first language was minimal or without pattern. The only study of this type that had conclusive results was Winke, Gass, and Mayford (2012), which showed that Spanish and Chinese raters were more lenient with raters from their own background, although Korean raters were not.
The inconclusive and unpatterned nature of the results leads to the conclusion that if non-native speakers were to be used as raters, testing program managers should periodically review test results to examine for potential bias trends with raters whose native language is either close to the language tested or close to the native language of the examinees.

Most of the research that was conducted on prerequisite proficiency levels assessed the prerequisite proficiency not of raters, but of teachers. Elder (1993, 2001) examined this issue and found such a qualification useful. Certainly, the rater's English proficiency level emerged as the most salient factor for rating ability, and therefore, it should be considered instead of the native speaker label.

A recommendation that is more founded in literature and supported by the results of this study is that non-native speakers who are ILR Level 3 or above speakers of English should be allowed the opportunity to serve as English testers, as most of the differences that occurred between the rater groups divided by English proficiency level were between the Level 2 English speakers and the other groups. Such a result indicates that the Level 3 and Level 4 raters may have acceptable prerequisite English speaking proficiencies.

### 6.3.2 Application to Second Language Acquisition

This study also has implications for Second Language Acquisition and ultimate attainment research, which is concerned with the end state of language proficiency in learners. This study indicates that it is possible for language learners to acquire the language well enough to have an equivalent competence, as defined by the ability to rate speaking tests, as native speakers. The non-native speaker population in this study is truly an exceptional one, as they are high achieving bilinguals. The non-native raters have a high level of linguistic and underlying
cognitive competence in another language. These competencies likely aid the ability to give holistic speaking ratings accurately, even if the bilingual does not have an ILR Level 5 proficiency in English. Ortega (2010, 2013) would argue that this makes the non-native speakers more appropriate raters, as bilingualism is the goal of a foreign language learner. It is not apparent from this data if monolinguals who do not have a high level of linguistic or cognitive competencies would be able to accurately rate high-level speaking proficiency exams, as all of the participants in this study were ILR Level 5 in some language.

The lack of any significant differences between the native and non-native speaker raters supports Corder’s (1967, 1981) and Selinker’s (1972) ideas about an independent interlanguage. All of the non-native speakers were able to overcome any assumed restraints from their native language, possibly because they were not trying to reshape their first language into the second language, but rather were relying on the rules and principles of the interlanguage.

Furthermore, the results do not support the idea of the Critical Period Hypothesis (Lenneberg, 1967). This hypothesis proposes the idea of maturational constraints on language learners who begin language acquisition post adolescence. The non-native speaker raters mostly began their language acquisition as teenagers and adults after moving to the United States. The linguistic achievement of the non-native speakers supports Birdsong’s (1999) idea that ultimate attainment in the second language has the potential to equal that of the first language. Again, the non-natives were Level 5 achievers in their first language, giving them the potential to reach a high level of proficiency in the second language, or English. The data collected has mostly to do with the raters’ competence rather than performance. There is no sign from this data set that the speakers’ Universal Grammar has deteriorated with age, as posited by the Fundamental Difference Hypothesis (Bley-Vroman, 1990) and Failed Formal Features Hypothesis (Hawkins
Studies that claim that there is evidence of the critical period hypothesis (Coppieters, 1987; DeKeyser & Larson-Hall, 2005) often cite grammar as an area where these differences appear. However, this is the one area where there were no differences in the competencies of the native and non-native speaker groups.

Perhaps one reason for these results that showed native and non-native speaker differences is that the native and non-native speaker populations were low to mid proficiency language learners and native speakers. When Montrul and Slabakova (2003) conducted their research, they used near native learners, and found no differences between the groups. This scenario is much closer to the context of the current study, which also includes very high-level learners.

The comparability between native and non-native raters in this research suggests that the concept of the native speaker alone is not appropriate as a yardstick for language learners (Muñoz & Singleton, 2011). The monolingual ideal native speaker is not what a language learner is trying to attain (Cenoz & Genesee, 1998). What is more appropriate is the idea of a balanced bilingual speaker as the goal (Ortega, 2010) or to simply analyze the language itself (Cook, 2002) without reference to an abstract person.

### 6.3.3 Application to native speaker construct

The results of this study speak to the fact that the categorical grouping of people into native speakers and non-native speakers is not useful when determining the qualifications for speaking proficiency test raters. Instead, one of the many considerations that must be made in selecting a speaking proficiency test rater is a minimum speaking proficiency ability, which is the trait that people are often referring to when they require a native speaker for those positions.
The raters in the current study were grouped and regrouped for analysis according to three distinguishing factors: (1) the language in which the speaker was natively raised, (2) the native speaker ideal, and (3) native speaker as a demographic characteristic. The first distinguishing factor was the native speaker/non-native speaker distinction, which refers to people who were born and raised speaking English in one group and people who were born and raised speaking another language in the other group. The second and third distinguishing factors divided the raters according to what is implied by the native speaker according to works by Davies (2003), Maher (2001), and Paikeday (1985). For the native speaker ideal (second factor), the raters were divided into groups by their actual English speaking proficiency rating. For the native speaker as a demographic characteristic, (third factor), the raters were divided by their native languages. The results of these analyses showed that for rating, it did not matter if you met the traditional definition of a native speaker or what your native language was at all. The most relevant trait was what your English proficiency rating was. Paikeday’s native speaker is the arbiter of grammatical acceptability (1985, p. 74); however, the native speaker, or Level 5 speaker for that matter, showed no significant difference in the area of grammar. These non-native raters did not meet Paikeday’s ideal notion of what a native speaker is, performing without a slip of the tongue and highly educated, yet they were able to rate language comparably to the native speakers. Maher (2001) notes that native speakers appropriate language and they are fully accepted as members of the native speaking community. These speakers have both biological competence and competence coming from being raised in the native speaking culture. However, the non-native speakers have none of these, and yet they are judged by some to be equivalent to natives.
Chomsky does not have a problem with the notion that the non-native speaker can perform as well as the native, as he does not believe that the language faculty changes through puberty to its end state. Instead Chomsky believes that the concept of nativeness is cultivated by influences from society. Paikeday extends Chomsky’s belief to say that each person is a native speaker of his or her own dialect, and no one is a native speaker of a standard language (Paikeday, 1985). Evidence received from the results of this study only support this idea. Here we have a set of highly proficient native speakers that have lived in the United States for the majority of their lives and have acquired the English language as well as if not better than most native speakers. They are in a sense just as much native speakers as people who only speak English or have lived in the United States all of their lives. Again, there is no statistical difference in the competence of the two groups.

Davies (2003) describes four knowledge types of the native speaker: metalinguistic, discriminating, communicational (relating to competence), and skills (relating to performance). As mentioned previously, communicative competence distinguishes levels of native speakers, as experiences with language and culture are expressed through language in a range of situations, paying careful attention to notions of appropriateness beyond personal experience. Metalinguistic knowledge supports a native speaker’s intuitions about language. Discriminating knowledge describes the native speaker discernment about language, allowing them to make judgments and evaluations about language ability. Communicational knowledge refers to the ability to understand the language and skills. Performance skills refer the ability to express thoughts orally. Native speakers that are performing as raters must prove their abilities in metalinguistic knowledge and discriminating knowledge. As raters, and not test administrators, they do not have to prove their abilities in communicative competence and communicational
knowledge. Therefore, according to this framework, the current study is only examining half of the four areas of knowledge that a non-native speaker would need to prove to ensure comparability with a native speaker.

Davies (2004) finds that differences between native and non-native speakers are more typically psycholinguistic rather than sociolinguistic. He finds that non-nativeness becomes apparent in speed and confidence in grammaticality judgments. That is certainly the case in this study. Some non-native speakers expressed unsolicited concerns about their ability to perform the task, even though they rated comparably to native speakers. Additionally, the non-native speakers did show some differences in fluency, particularly in the Level 3 tests. Davies does mention that many of these items are difficult to test, so usually nativeness is not a measure of ability, but rather is usually a measure of identity and confidence (2003).

The lack of difference between native and non-native speakers in this study’s results only further supports the claim that the term native speaker is ambiguous, misinterpreted, and potentially unfounded as a reference to speaking proficiency ability. It is not a term that is useful to represent a level of language proficiency, but would be better maintained as a referent for a method of language acquisition, referring to the first language learned. The Oxford English Dictionary (OED, n.d.) defines native speaker in this way, and that should be all it entails. Case in point, the non-native speaker raters used in this study had a speaking proficiency in English that met, if not exceeded, many typical native speakers. Furthermore, whether a person is a native speaker should not be used as a qualification for expertise in evaluating language proficiency.
6.3.4 Application to social and political issues

As there are no significant differences between native and non-native speaker raters at any of the exam levels, it would be unfair to categorically disallow a non-native speaker from applying for a speaking proficiency rater position solely on the basis of their nativity. Non-native speakers should be given equal opportunity to compete for speaking tester positions in languages other than their native/first language if they can demonstrate ability in the requisite rater competencies. Those are the qualifications that a rater needs to perform the job consistently. Alternatively, the results of this study should not be interpreted as a group entitlement for non-native speakers to be admitted as raters, just as it should not be a group entitlement for native speakers to be admitted as raters. Both groups have the right to compete for the rater position, and each individual candidate must demonstrate their ability to perform the job accurately and reliably. With the right qualifications and competencies, both the native and non-native speakers have demonstrated the ability to discern measures of correctness uniformly and provide a fair test rating in this study.

The concept of the native speaker as the measuring stick and the usefulness of the native speaker as a rater are intertwined. As the two concepts are closely linked, much confusion can occur in the use of the term and it can have very personal implications. The top rating of the ILR Skill Level Descriptions, ILR Level 5, is referred to commonly as the native speaker level. Additionally, a rater must be both an ILR Level 5 and a native speaker, by language acquisition, in order to qualify for the position. These native speaker (by acquisition)/native speaker (by ideal language user) raters evaluate examinees, some of whom are native speakers and some of whom are non-native speakers, by acquisition. When a native speaker, by acquisition, receives a rating that is less than ILR Level 5 (the native speaker level), it can often be frustrating or even hurtful,
as the expectation is that they will receive the native speaker level score, since after all, they are a native speaker, though only by acquisition. In fact, less than 5% of native speakers by acquisition receive an ILR Level 5 rating. The native speaker examinee then interprets the rating of ILR 3 or 4 (which are excellent ratings) as disappointing and an evaluation of the examinee’s identity. The native speaker identifies with the language, and defines himself or herself by the language, only to discover that they are not a native speaker after all.

Davies (2011) comments that native speaker by acquisition really refers to the native speaker of an idiolect, not of a standard language, which is what is assessed in an exam. So the examinee thinks that the language rating in his or her native language is an evaluation of his idiolect rather than the standard language, of which no one is a native speaker. McNamara also discusses how language tests are about identity, both at the macro and micro level (2007). Perhaps, though, the results of this study can help to disentangle this problem. According to the analyses, it is not necessary for a rater to be either a native speaker or an ILR Level 5 speaker to be an accurate rater. Therefore, the label “native speaker” should be removed from the ILR Level 5 and the native speaker requirement removed from the rater position. This solution would be an application of McNamara’s theory of the social context of testing, which involves thinking though the implications of the tests and ratings given and considering how issues outside the test, such as identity are impacted.

Which English is applied to the test is of concern in such a high stakes test as the FBI’s Speaking Proficiency Tests. Indeed, the form of English that is used is a standard American version, although the testers are instructed to permit acceptable uses of other standard forms. As Davies (2011) says, tests are evaluations of standardized forms of the language, not of idiolects. Jenkins (2006a) argues that English as a Lingua Franca should also be considered as a
standardized form rather than just the “Inner Circle” Englishes. However, the decision to use the American version of English to administer the test is a practical and appropriate one, according to Bachman’s Assessment Use Argument. American English is the variety used in the US Federal Government, so it is appropriate that it is tested. As Taylor (2006) states, it is not only the variety that is imposed by the administering agency, but also of the examinees taking the test. The political decision to use American English is tempered by the decision not to require ILR Level 5 for positions within US Federal Government. At present, a score of ILR 2+ is sufficient for most positions, except for that of language testers and raters. If this requirement were removed, then an examinee could speak non-standard forms of English and still be successful on the exam, as long as the examinee is able to communicate the message. The emphasis of the evaluation would then change from form and content to content alone.

The ratings issued in the Speaking Proficiency Test are concerned with correctness, but not just grammatical correctness. The exam deals with determining that language is both correct and appropriate (Ohlander, 2012). In some senses, it is easier to deal with the concept of grammatical correctness, as there are rules and standards for what is acceptable. The matter of the appropriateness of language use is much more subjective and less prescribed, yet still important. These two correctness issues help to explain the lack of difference in the linguistic category of structures, which is largely prescriptive. More often differences were seen in rater groups in linguistic categories such as organization, fluency, vocabulary, and social/cultural appropriateness. Shohamy (2006) warns against one criterion for correctness, as it perpetuates uniformity. However, overall the raters are trained to listen not only for accuracy of the language, but also effectiveness of the message. With the passing rating set at about 2+, wide
variability is permissible for most positions and purposes for Language Testing. The removal of
the term “native speaker” and its ambiguity will help to clarify this notion.

Messick (1981, 1989, 1994, 1996) warns how tests have the tendency to separate people
into groups, reflecting policy and priorities of the administering organization. Shohamy (2001)
advises test stakeholders to be aware of the power of the test, and its impact on the life of the
examinee, as it controls access to opportunities. One way to mitigate such power is to make
informed decisions about appropriate test usage and interpretation of scores. This requires that
education be given to the stakeholders. The results of this study help to inform such policy
decisions by suggesting the removal of the native speaker concept from the highest level of the
ILR Skill Level Descriptions, and to not require a Level 5 rating for raters. Policies such as these
being a native speaker, or being labeled as a native speaker, carries power and authority, so the
elimination of this misunderstood and misused term can help reduce that problem.

After properly considering all of these issues surrounding a test (such as the magnitude of
the testing body, the composition of the final score, the sample of examinees in test development,
and the format of the test) (Shohamy, 2006), then a test’s fairness can be properly evaluated. Test
fairness is concerned with the use and subsequent implications of test scores (Elder, 1997;
Kunnan, 2000, 2004). The issue of test fairness was a driving factor for this research, as it would
not be fair to categorically eliminate a group of examinees based on a demographic
characteristic, namely being a non-native speaker. The SPT strives to follow an appropriate
Assessment Use Argument (Bachman, 2007) and it designs test batteries that include the SPT to
cover the skills necessary to do a job. The ILR Skill Level Descriptions are designed for the US
Federal Government context, and are appropriate in that setting (Lowe, 1983). Considering the
local use of the test at the FBI (Chalhoub-Deville & Wigglesworth, 2005), the SPT accomplishes its purpose among other tests in the test battery, and is constantly undergoing quality evaluations and improvements so that it meets the needs of its agency.

Since the terms native speaker and non-native speaker have been proven not to be relevant in the context of rater candidate qualifications, those terms should therefore be abandoned for use in the rater selection context. For it is not how the language was acquired that matters, but how well a person was able to rate speaking proficiency samples. What matters are the other rater competencies and a minimum English proficiency of at least ILR Level 3, according to the results of this study. Since the relationship between being a native speaker and identity are so critically intertwined, Language Testing organizations could avoid unnecessary complications and misunderstandings by not using the term native speaker or non-native speaker to refer to qualifications. If a person were to end up not being qualified due to a lack of other prerequisite linguistic, cognitive, or evaluative competencies, the candidate might misinterpret the rejection of their rater candidacy as a rejection them being a native speaker and consequently a reflection on their identity as a native speaker. Of course, the testing organization has no intention to make a statement about a person’s identity as a native speaker, but it would be useful to recognize that these two concepts are inextricably intertwined and that a misinterpretation like this is common. Luckily, refraining from the use of native speaker as a qualification for position would help to avoid this offense.

6.4 Implications

The results of this study, and in particular research question 1, have several practical implications for testing programs. The first implication is that non-native speakers of English
should be allowed to serve as English test raters, if they can demonstrate a reliable ability to accomplish those tasks. Beyond practical implications, there are further theoretical implications of this study. The results show that non-native speakers of English, with proper training, can acquire English well enough to rate English-speaking proficiency exams as well as trained native speaker raters of English. Additionally, the study shows that non-native speakers can discriminate between high-level samples of English as well as native speakers, if they’ve had the same training and have a high level of proficiency in English themselves. This result has ramifications for the question of ultimate attainment—whether or not there are differences between non-native speaker competence at the highest levels of language proficiency and native speaker competence. At least in rating of oral proficiency, this study presents no evidence of a difference between non-native speaker and native speaker competencies.

The results from research question 2 do have some practical implications for the testing program at the FBI. As is the practice currently, it is not appropriate to use Level 2+ raters as raters of exams at all levels. It is important to consider that when raters are assigned the examinees to evaluate, they do not know at what level the examinee will be. So even though Level 2 raters are successful at rating Level 2 examinees, there is no operational way to guarantee that they will only ever encounter Level 2 examinees if they were put into the general rating pool.

Based on the ANOVAs from the study, it appears that Level 4 raters do not show a significant difference with Level 5 (NS) raters, and therefore should be allowed to test for certification as raters of English tests. Level 3 raters only ever show significant differences with Level 2 raters and not Level 5 (native speaker raters), so potentially they could be used as raters.
of English exams at all levels as well and should be allowed to test for certification as raters of English tests.

Beyond the local testing context at the FBI, the Interagency Language Roundtable should consider that the Skill Level Descriptions should not just apply to non-native speakers, but should be equally applied to native speakers of a language. By default, speaking ability should not be referred to as native versus non-native, which implies the assumption that native is superior to non-native, but rather that a particular person is able to speak the language at a particular level, with the understanding that the level is or is not appropriate for a particular task.

If a system based on rater proficiency level as a prerequisite for rating ability were used rather than notions such as native speaker, non-native speaker, or fluent or proficient (in their categorical senses), all of which have no universal meaning or interpretation, then proficiency level and being a native speaker would be less intertwined and confused. Native speaker could simply be used to refer to the language(s) and/or culture(s) that a person learned primarily and through their childhood. The native speaker would then be totally extracted from any assumption of language ability or proficiency. This would also help to solve the problem of confusing identity with being a native speaker. Understandably, it is very difficult for a native speaker to accept the fact that he or she may not have the highest proficiency level of language. Receiving a rating of less than the perfect rating of Level 5 often feels like a critique of a person’s identity, as if it means that they are not classified as a native speaker. This confusion would be reduced if being a native speaker could be tied to identity instead of being tied to proficiency level and therefore indirectly tied to identity.

The results of this study also have implications for native speaker rater candidates. Currently the qualifications to be a rater of English include being both a native speaker and an
ILR Level 5 speaker of English. If non-native speakers of English that have a lower proficiency than ILR Level 5 are to be allowed to compete as English rater candidates, then also native speaker candidates that are less than Level 5 should be equally considered. Each candidate must individually demonstrate their ability to rate reliably at any level of exam. This notion is supported by the fact that not every native speaker ILR Level 5 rater candidate that attends the rater training workshop is able to successfully meet the certification requirements. Being a successful rater involves a combination of different competencies that must all be met, regardless of rater background or proficiency level.

A successful speaking proficiency test rater is a person who is able to combine a high-level of linguistic competence, cognitive competence, evaluative competence, and relevant language proficiency (FBI, 2009). Linguistic competence is an awareness of and ability to discern the features of the language. It includes an ability to dissect and analyze not just the structures that a person is able to use but also the strategies and effectiveness of their message across all contexts. When an individual achieves a high level of language proficiency in any language, they must have behind it a high-level linguistic competence. This linguistic competence helps them to discern how to tailor their message to a particular audience to be maximally effective, but also be able to be sensitive to the subtle features of language that may have an impact. Linguistic competence has as a prerequisite cognitive competence. In order to deliver a high level message that weaves together complex, abstract ideas, and then later articulate them clearly to an intended audience, a speaker must be able to handle a high cognitive load. Additionally, an effective rater must have a high level of evaluative competence. He or she must be able to discern when a speaker is using the right forms, the appropriate vocabulary, the correct style, and other linguistic features that require a discerning ear. A person who has a high
level of evaluative competence is reliable both with their co-evaluators and within themselves over time. They are able to absorb instruction on how to interpret rating criteria and apply those concepts consistently. Finally, an effective speaking proficiency test rater also has to have a certain level of language proficiency in the language being rated. They must understand how the language functions in the different forms (structures, words, pronunciation features, styles) that are used in that language.

The participants in this study, both native and non-native speakers, are all current raters. They have all consistently demonstrated linguistic, cognitive, and evaluative competencies in a language, their native languages. Additionally all of them have a good to extremely high level of language proficiency in English. Because the linguistic, cognitive, and evaluative competencies were controlled across all participants, the only true variable was the participants’ language proficiency. However, when considering other candidates for the rater position, they will not have proven linguistic, cognitive, and evaluative competencies. Therefore non-native rater candidates, just like native rater candidates, need to demonstrate those abilities in addition to a certain level of language proficiency before they would be able to qualify for a rater position. However, this does not take away from the fact that non-native speakers need to be afforded the same opportunity to be raters as native speakers.
Chapter 7: Conclusions

7.1 Overview

The comparison of native and non-native speakers’ abilities has existed in both Second Language Acquisition (Corder, 1967; Selinker, 1972) and Language Testing (Galloway, 1980) research for over 40 years. A variety of approaches have been taken to examine the comparison of these two different groups with mixed results. In some studies, no differences are found (Barnwell, 1989b; Eckes, 2008; Kim, 2009a; Kim, 2009b; Ludwig, 1982). In others, differences are apparent (Fayer & Krasinski, 1987; Galloway, 1980; Ross, 1979). The current study builds on the foundation of the previous literature by offering a research design that looks at highly proficient English speakers’ (both native and non-native speakers) evaluations of high-level English exams within the context of an operational testing program. The expansion of the ranges of abilities of the examinees (ILR Levels 2 through 5) as well as more precise measurements of rater characteristics (English proficiency level and native language) help to supplement research to date.

The results from this study determine that there are no significant differences between the final, holistic ratings of the native and non-native speaker raters in rating English SPTs ranging from ILR Levels 2 through 5. As a result, non-native speaker raters should be permitted to undergo training to serve as English oral proficiency raters. However, it appears that not all non-native speakers are equally able to evaluate the English speaking exams. Analyses suggest that ILR Level 3 to 4+ proficiency non-native speakers of English (who were also trained SPT raters of another language) would be able to rate tests as well as ILR Level 5 proficiency native speakers of English. The other conclusion that can be made from the research is that ILR Level
2+ proficiency is insufficient for rating tests at all ILR Levels, even when the Level 2+ speaker is a current rater who has a high-level of linguistic, cognitive, and evaluative competencies. This conclusion is supported by significant differences between the Level 2+ raters and Level 3/3+ raters, and between Level 2+ raters and Level 4/4+ raters in post hoc tests where the Level 2+ raters gave final ratings that were consistently lower than the other rater groups.

Analyses of linguistic category sub-ratings indicate numerous significant differences among the rater groups on the linguistic category ratings. Even though such differences exist, the effect of these differences does not translate into differences in the final ratings. Differences in the linguistic category ratings between the ILR Level 2+ raters and the other rater groups’ further support the conclusion that the Level 2+ raters were rating on the whole consistently lower than the other English proficiency rater groups.

In addition to the immediate consequences that these results have on the testing program at the FBI, this research considers larger questions about the definition of the native speaker, the appropriateness of using the term native speaker in Second Language Acquisition and Language Testing research, assumptions about proficiency requirements for raters, and the applicability of the results in other testing contexts that are not holistic in nature. Each of these issues is discussed further as well as the implications of the current study’s results. Although this study was designed to build upon the current research on the comparisons of native and non-native speaker raters specifically and the necessary qualifications to serve as a rater in general, it is in some ways limited by its operational context. Therefore additional research would be beneficial to continue to add to the understanding of what it is to be a native speaker.
7.2 Conclusions

7.2.1 (Re)defining the native speaker

Many of the problems in the use of the ‘native speaker’ term in Language Testing, Second Language Acquisition, and in general come from the fact that the term is both ambiguous (Davies, 2004) and it is layered with assumptions on the parts of both the term’s users and the audience (Marsh, 1859; Escudero & Sharwood-Smith, 2001). Native speakers are assumed to have intuitions about the language, to be arbiters of its correctness, and display ideal language (Davies, 2004). Even though the term is defined by sources such as the Oxford English Dictionary (OED) as, “a person who has spoken a specified language since earliest childhood, as opposed to a person who has learnt it as a second or subsequent language in later life” (n.d.), it has also been used to refer to an ideal speaker of the language, either directly (ILR, 1985), or indirectly (Whitney, 1875; Wyld, 1906-7). The Interagency Language Roundtable’s (ILR) association of the native speaker at the top of the scale was not intended to refer to the average native speaker as the OED defines it, a native acquirer of the language, but rather an ideal speaker of the language. There are attempts to clarify this distinction by modifying the native speaker as the “highly articulate, well educated native speaker;” however, this specification often gets lost, as the Level 5 is commonly referred to as the native speaker level. Unfortunately, this shorthanded speech adds to the confusion.

Therefore, when the term native speaker is used in any research or testing context, it should be clearly defined. Defining the term would allow for comparability between studies where the native speaker term is used and avoid potential misunderstandings about what is intended by the use of the speaker. The native speaker should clearly either refer to how the speaker learned the language or to the fact that the speaker is an ideal speaker by some standard.
Additionally, how the person who is labeled as native qualified for that label needs to be explicated. Are native speakers qualified to use that term by the fact that they were born in an area where the language is natively spoken? Perhaps it is because they were raised speaking that language or because it was their first language (Barnwell, 1989)? For some researchers, native speakers (Zhang & Elder, 2011) must also be educated in that language? Speakers of a language may consider themselves native because they learned the language through an immersion context or because it is now their primary language (Jenkins, 2006a)?

The other way that a native speaker may be defined is as the ideal speaker (Davies, 2004; Maher, 2001). It is just as important to specify that term native speaker refers to an ideal speaker as if native speaker referred to the method of acquisition. However, simply stating that by using the term native speaker what the author means is an ideal speaker is not sufficient. What does the author mean by the ideal speaker? If ideal refers to a measure of speaking proficiency, then how that speaking proficiency was measured should be explained in detail. This may include which test was used, which rating scale was employed, and how to interpret what it means and does not mean to have the highest score possible on that particular scale.

The use of the term native speaker in the Language Testing context is particularly problematic, especially when a native speaker implies an ideal speaker of the language instead of someone who just acquired the language in the native-like manner. Research has shown that native speakers are not a homogeneous population (Hamilton et al., 1993). The variation among the speaking proficiencies of native speakers is almost as diverse as those of non-native speakers. Research conducted at the FBI (Brooks & Brau, 2007) demonstrates that native speakers score anywhere between ILR Level 2 and ILR Level 5 in speaking proficiency. Non-native speakers can score anywhere in the entire range of this scale, from Level 0 to 5. Although it is uncommon
to see native speakers at the lower ranges, between Level 0 and Level 2, there is a great deal of overlap between the proficiencies of native and non-native speakers. For that reason, it is better that native speaker does not refer to a particular ability, but rather method of acquisition of the language.

The ambiguity of the native speaker term has made it difficult to interpret the results of previous studies uniformly due to the fact that it is often unclear whether the researcher is referring to someone who acquired the language in a native manner or someone who is at a high level of proficiency. As a result, finding clear trends among multiple articles and arriving at clear conclusions that may lead to a better understanding of the native speaker and Language Testing principles is challenging. Since the native speaker term is not clearly defined, it is possible that the author’s conclusions could be misinterpreted or potentially altered depending on what was meant by the native speaker.

For example, Barnwell’s study investigates that comparability of native and non-native speakers of Spanish in their evaluations on American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interviews (OPI) (1989b). Barnwell’s (1989b) article documents one of the seminal studies within the context of Language Testing that examines rater group differences according to whether or not they are native speakers. Barnwell compares the ratings given by two trained ACTFL Spanish speaking raters (non-native speakers) to the ratings that are given by fourteen untrained native speakers from Spain. The native speaker group are not clearly defined other than they are from born and raised in Spain, and are therefore native speakers. Barnwell posits that as the purpose of learning Spanish is to be understood by native speakers in a native-speaking context such as Spain, the trained ACTFL raters are act as native speaker surrogate judges and should therefore be comparable to native speakers. Barnwell
reports similar patterns in how the native and non-native speaker rater groups ranked the four examinees. More specifically, Barnwell concludes that the native speaker raters are more severe than the non-native speaker raters in high-level non-native speaker examinees. Barnwell hypothesizes that native speakers react negatively to non-native speaker examinees with higher levels of proficiency (Barnwell, 1989b).

Barnwell characterizes the native speakers simply as that, native speakers with the assumption that the reader understands without fully entails. What Barnwell probably means is that these raters are native acquirers of the language since he characterizes them as being people born in raised in Spain in Spanish is the language in question. Barnwell does not address the speaking proficiency level of the native speaker raters; it is assumed that it is excellent or perhaps ideal. ACTFL raters are non-native speakers of Spanish, but may have demonstrated a certain level of Spanish proficiency on the ACTFL speaking scale in order to be in that position.

One of the Barnwell’s conclusions is that the native speaker raters are more severe on the high proficiency examinees, particularly those who were non-native speakers of Spanish. Barnwell’s finding appears to at first glance disagree with a finding of the current study, that non-native speaker ILR Level 2+ raters were more severe in their ratings than any of the other proficiency rating groups. Barnwell finds that native speakers assign more severe ratings, and the current research finds that the rater group that was significantly more severe is the Level 2+ non-native speakers. It could be that the results of this study and Barnwell’s 1989 study are comparable. However, the way that Barnwell is conceptualizing the native speaker makes the conclusions appear to be at odds. It may be that if Barnwell had defined the rater groups by proficiency level rather than native versus non-native speaker groups, he may have found that the native speakers were about the Level 2+ (or ACTFL’s Advanced High Level); therefore
comparable to the Level 2+ speakers in this research. The tendency to be harsh with ratings may be caused more by the proficiency of the rater than whether or not the rater was a native speaker. If Barnwell had required his native Spanish speaker raters to be highly proficient raters, equivalent to ILR Level 4 or 5, he may not have found distinctions between the native and non-native speaker groups, reversing his conclusion. This goes to show that the discrepant nature of conclusions in native versus non-native speaker research within the context of Language Testing maybe due to ambiguous definitions of native speaker.

Although it would be interesting to consider how the conclusions of a study that defined native speaker as a highly proficient speaker of the would be different if native speaker were defined as a native-acquirer of the language, there are not many Language Testing studies where the native speaker is defined explicitly with a level proficiency. In Zhang and Elder (2011), the native speaker rater group is characterized by the fact that they acquired the language in a native manner. However the researchers were careful to select native speakers and non-native speakers both who had qualifications that led them to believe that they were highly proficient in English, the language of interest in their study. Like in the current study, there were additional similarities between the native and non-native speaker rater groups, such as the raters’ background and experience. Both the native and non-native speaker groups had experience teaching English as a second language and some members of each of the groups had experience being formal raters. The rater training was a factor in this study as well.

Perhaps since the two groups came from such similar populations, both were English teachers to speakers of other languages, experience raters, and highly proficient speakers, it is not entirely surprising that no overall differences were discovered between the two groups, as in the current study. Like in the current study, there were differences in the ways that the two
groups approached or analyzed the rating samples. It is possible that if Zhang and Elder had divided the raters by measures of language proficiency rather than just being a native speaker, differences may appear between proficiency levels equivalent to those in the current study. The researchers may have found that not all non-native speakers performed equivalently, and that non-natives with higher proficiency levels were more comparable to native speaker raters with high proficiency levels, but that raters who had lower proficiency levels assigned ratings that were significantly different.

It may be inappropriate in many cases to use the native speaker as a yardstick for measuring speaking proficiency. Instead, it is likely more appropriate to actually measure the speaking proficiency of an individual and report the ability on a rating scale then to give a person the label, such as native speaker, and assume that that means a measurable criterion against which an examinee’s performance can be measured.

Native speakers have often been used as raters because it was assumed that they would have a better ability to judge the differences between examinees (Maher, 2001). It has been posited that being a native speaker gave a person some sort of intuition about a language that would be difficult if not impossible to acquire (Brown, 1995; Davies, 2003). The current study, as well as many of the studies that preceded it, has reconﬁrmed that a clear distinction between the ability of a native speaker and a non-native speaker rater is not evident even when making distinctions between high-level proficiencies, deﬁned as ILR Levels 3 through 5. The results support the theory that not being a native speaker does not preclude a person from being able to make distinctions between rating levels as accurately as native speakers. The current study was conducted within an operational testing context, which means that the rater participants are not laypeople. They are all trained raters of speaking tests. Consequently, it is possible that
differences between the ratings given by native and non-native speakers on high level exams can be overcome by extensive training and adhesion to detailed criteria.

Since there is a difference between a native speaker and the high proficiency, ideal speaker, perhaps an alternative term for the latter should be adopted. Terms such as language expertise, language inheritance, and language affiliation (Rampton, 1990) have been suggested in previous literature, although these seem more appropriate when referring to the method of acquisition. Perhaps something such as highly articulate speaker would be more appropriate. The use of such an alternative term would help disambiguate what was meant by the language testers when referring to the native speaker. Therefore, when person is labeled a native speaker, it would refer to their manner of language acquisition only. The result would, hopefully, be less confusion.

Additionally, the general public, or more specifically the examinees, should be further educated by the test administrators that the skill of speaking is a learned skill, and therefore not fully developed by simple being a native speaker. Although language acquisition may be an innate and natural process for children, at a certain point\(^39\), even a native speaker by acquisition must become a learner of the language in order to improve their rating. This high level proficiency may come through higher-level learning and/or exposure to a variety of professional contexts where such language is used. The responsibility then lies on the testing bodies to better educate the examinee population on the levels of the scale. For example, in the US Federal

\(^{39}\) In the ILR Skill Level Descriptions, evidence has shown that many native speakers naturally acquire the language to the Level 2+, but must intentionally develop the language further to reach the Levels 3 and above (Brooks & Brau, 2007).
Government context, it should be understood that language learners generally deal with lower levels of acquisition, equivalent to ILR Levels 0 to 2. Being a native speaker may only naturally result in a rating in the ILR 2 range. For the ILR professional levels, Levels 3 and above, the amount of complex structures, vocabulary, fluency, and appropriateness of speech or pragmatic usage of the language requires extensive practice, flexibility, experience, exposure to such language (FBI, 2009; ILR, 1985). With that understanding, examinees have a better chance of interpreting their rating and understanding the justifications. Language test education, or at least appreciation, can help avoid disappointment because examinees expected higher ratings, such as ILR Level 4 or 5, instead of mid-range ratings, in ILR Levels 2+ and 3.

In conclusion, a native speaker should a term reserved for those who acquired the language in a native manner, as defined in the *Oxford English Dictionary* (n.d.). When referring to an ideal speaker or conceptualization of someone who is at a high level of proficiency, the individual should not be identified as a native speaker, but rather identified as a person who has a speaking proficiency at a particular level. Such a distinction, if uniformly adhered to, should help to disambiguate the term native speaker. The ideal speaker should be referred to by a level of proficiency, an actual measurement, instead of hypothetical ideal.

### 6.2.2 Justifying the native speaker

More important than defining, or redefining, the native speaker is justifying the usage of native speaker concept at all. It should not be assumed that it is in all cases appropriate to use the native speaker concept. The improper or inappropriate use of the native speaker term can lead to confusion (Shohamy, 2006). Additionally, when examinees that consider themselves native speakers are not given “native speaker ratings,” the result can be hurtful and alienating because
the rating is implying that the examinee should not be associated with language as a native, which may be the examinee’s identity (McNamara & Roever, 2006). Before defining what the native speaker is in a Language Testing or research context, one must consider whether or not it is appropriate to use the term at all. First, one must consider why the term is an appropriate model or concept. Researchers and test developers must ask themselves whether they are looking for a language acquisition ideal or are they looking for the way that the language was learned? What is the goal or what is trying to be expressed by the use of native speaker?

According to Bachman’s Assessment Use Argument (2007), a test developer needs to have a proper justification for the use of different elements of the test, as well as help the examinee to properly interpret what the results mean. In some cases, it may be most appropriate to use native speaker to refer to how the language was acquired. This approach would be particularly useful in Second Language Acquisition research, as it is important to categorize language learning methodologies to better understand how language acquired. In Language Testing, it may be more appropriate to categorize speakers by proficiency level for both examinees and raters rather than by how the language was learned, thereby assuming a high proficiency level.

In both instances, it may be most appropriate to use a third concept, a bilingual speaker. It has been argued by Ortega (2010, 2013) that the use of a monolingual native speaker as an acquisition goal is inappropriate. In Second Language Acquisition, what a learner is actually aspiring to be is not a monolingual native speaker, but rather a bilingual speaker. The language learner already has a native language as an asset when beginning the second (or third) language acquisition process. Therefore, the goal or target for language acquisition should be a bilingual speaker since that is what the learners looking to become. When testing a second language, the
bilingual speaker may also be the most appropriate model. If the examinee is a language learner, the context is not that dissimilar from second language acquisition.

Since testing has to do more with measurement, the construct of the ideal speaker may be more appropriate in some cases; whereas, the bilingual speaker may be appropriate yardstick for language testing in other scenarios. For learners in an academic context, the bilingual speaker may be the most appropriate model. In US Federal Government language testing, however, there are circumstances where the purpose of language testing is to ascertain whether or not the person is equivalent in every way to a monolingual native speaker. For example, if a person is being asked to act in a covert mission, it may be important that they sound and use the language in exactly the same way as a monolingual native speaker would. It is important that they covert operator be imperceptible from a native speaker (FBI, 2009).

However, in most US Federal Government testing situations, it is important to understand that the high level speaker is not the goal. It would be useful to remind examinees that although the ILR Skill Level Descriptions range from Level 0 to Level 5, Level 5 is the maximum possible level, but it is not required or even expected from most examinees. Most of the positions in the US Federal Government context require only ILR Level 2 or Level 3 ratings, not the highest possible rating of ILR Level 5. There are few positions in the US Federal Government that require ILR Level 5, and one of those positions is the speaking proficiency tester at the FBI (FBI, 2009). The research in this study has shown, however, that this requirement is not
necessary for rating speaking exams, considering that the rater has adequate training, is highly competent, and (potentially) highly proficient in another language\textsuperscript{40}.

Not having a proper justification for the use of the native speaker term can have consequences far beyond the test context itself (Shohamy, 2006). Test developers must consider not only a justification for why the native speaker is an important part of the test construct, they must also consider that the social and political implications of their decision will persist for years after the test is administered and rated, a theoretical carbon footprint of the test. Insufficient consideration for how the native speaker is employed can severely impact the fairness of the test, in much the same ways other parts of the construct of the test (Elder, 1997; Kunnan, 2000, 2004). It may not be fair to judge the ability of a learner against that of a monolingual native speaker, when a learner can never become in monolingual native speaker of the language. Research designed to determine the significant differences between the performances of a high proficiency bilingual speaker in a monolingual native speaker would help reveal the impact of this type of difference.

If the native speaker concept is to be used, it is important to consider which variant of native speaker (or which variant of the native speaker’s language) would be appropriate for the test. Arguments for using Inner Circle (Kachru, 2006) or standard variants of a language like English, such as American or British English, include the fact that this is typically what the learner desires to acquire (Taylor, 2006). In the case of the United States Federal Government, it may be most appropriate to use the American English variant as a model. The FBI does make ________________

\textsuperscript{40} This third point would need to be verified in a separate study.
considerations for speakers who use different, but equally comprehensible, variants of English as the requirements for most positions do not require a particular variant of English. If being imperceptible from a native speaker is the goal of the test, than it may be important to judge whether or not there is any perceptible difference between the examinee in the native speaker.

Test developers must also consider how they’re going to treat the issue of correctness in relation to their use of the native speaker in the test (Hutton, 1999). There is wide variation among native speakers even within the same standardized dialect or regional dialect within a language. Test developers should make deliberate choices as to whether or not usage by any native speaker justifies a particular term or phrase as correct in the test, which may be the case if the average native speaker if the appropriate model (Jenkins, 2006a). In cases where an articulate native speaker is the model, it may be more appropriate to have a more strict philosophy on correctness. This might be the case when a person is being tested to represent the country in diplomatic tasks, where the way of person speaks reflect on a country in general (FBI, 2009). There may even be cases where the use of a regional variant is desirable. If a person is not speaking in public and representing the country a strict interpretation of language correctness may be inappropriate. If a person is working in a regional context it may be desirable for the examinee to display familiarity with and ability to use that variant. Whatever choice is made, the choice will inevitably perpetuate a certain set of ideals (Shohamy, 2006).

If the test developer decides to use the native speaker as a yardstick to measure other native speakers, they should consider the impact it would have on the examinee is they do not receive a “native speaker” equivalent score. This directly addresses the issues of identity (McNamara & Roever, 2006), as the examinee may be a native speaker and then determined not to be equivalent to native speaker. The examinee is then left wondering what this test result
means for them as a person. Are they not a native speaker? Is the examinee now excluded from the characteristic which they previously considered central to their identity? The result may be that the examinee rejects the score and loses confidence in the test. In these cases it seems better just to avoid the use of native speaker in a rating scale context. This is further evidence that the ideal notion of a native speaker should be defined by speaking proficiency level and not a general term.

Finally, the test developer should consider the unintended policies that might be supported from the decisions that they make regarding native speakers (Hawkey, 2006; Shohamy, 2001, 2006). For example, using the term in itself implies that there is a significant reason to distinguish this group of people when in fact there is little evidence that says they are different from non-native speakers. Using a native speaker ideal favors those who acquired the language in a native manner, which is something that cannot be changed later in life. This may unfairly disadvantage those who learned the language later in life or as a second language.

7.2.3 It does not take one to know one

Another issue that is raised though the course of this study is whether or not “it takes one to know one.” In other words, does a rater need to be an ILR Level 5 in order to accurately rate at ILR Level 5 and below? By the same theory, a rater would be required to be an ILR level for in order to rate exams at Level 4 or below, and so on. This theory comes from the idea that in order to be able to make the distinctions between two levels, one must fully comprehend and be able to produce speech at the higher level. There is very little research to support this hypothesis; however, it has been in practice for years by several testing organizations (FBI, 2009). What this study shows is that this hypothesis is not universally true. The results show that it is possible to
display a production competency at lower level of proficiency and still accurately distinguish differences between levels higher than a rater’s performance ability. Non-native speaker raters who have an ILR Level 3 and Level 4 proficiency in English were able to rate exams at all levels comparably to native speaker ILR Level 5 raters.

There are several possible explanations or mitigating factors that would affect the broad application of this conclusion. In this study, the scoring model that was used is a holistic rating in a test over approximately 45 minutes. Whether or not this same distinction persists when a test is shorter or if scoring is more analytic rather than holistic is unknown and would require further research. Another possible explanation for the rejection of this hypothesis is that all of the participants have undergone equivalent rater training conducted largely in English, and this rater training was able to overcome differences that existed between the English proficiency levels of the participants.

A possible explanation for the rejection of the “it takes one to know one” hypothesis is that there is a considerable difference between the performance competence needed to score a particular level on the speaking proficiency test and the receptive competence that a person would need in order to rate the test (Chomsky, 1965). Again, the group of participants in the study was not composed of laypeople, they were highly-trained raters. The raters were also ILR Level 5 in another language, which means that they had a high level of linguistic and evaluative competence already established. These competencies may be transferred from their ability to evaluate in one language to their ability to evaluate English speaking tests.

Overall, the theory of “it takes one to know one” is anything but certain. This idea needs to be further tested with different rates or populations and in different testing contexts. However, it is encouraging to see that additional resources may be available for raters, as the requirements...
for becoming a rater may allow for applicants who have lower levels of proficiency than ILR Level 5, and not necessarily a native speaker. Additionally, the Second Language Acquisition notion of performance versus competence (Davies, 2003; Hawkins & Chan, 1997; Sorace, 1993) seems to be reassured through the data in this study. There seems to be a difference in a person’s ability to evaluate a language such as English and their ability to perform in English. Still, there may be some relationship between these two abilities, which is supported in literature (Chomsky, 1965), as Level 2+ the speakers were not able to rate comparably with speakers of other English and proficiencies. It is perhaps the level 2+/3 boundary that is the tipping point for an evaluator to have enough competence in the language on average to be able to perform rating duties.

7.2.4 Differences in holistic rating versus analytical rating

One final issue that may affect the broad application of the results is that the FBI’s SPT is a holistic proficiency test. It may be that the results are not applicable to an analytically-rated language test. The current study, like many of the Language Testing studies conducted in the last fifteen years that focus on differences between native and non-native speaker raters did not find significant differences between native and non-native speakers of the language in holistic ratings. However, there were significant differences when analyzing linguistic sub-ratings assigned in the test. In studies that examined differences in evaluations of particular linguistic features or comments given to justify the overall rating assigned, it is quite common for significant differences to appear between the native and non-native rater groups (Zhang & Elder, 2011). If this analytic method of rating is appropriate for a test, then perhaps the test developers and administrators should consider more carefully whether or not to use both native and non-native speakers as raters.
For example, differences appeared in this study between native and non-native speaker raters in the linguistic subcategory of social/cultural appropriateness. The category of social/cultural appropriateness refers to the pragmatic usage of language. Since pragmatic language use is closely tied to culture and not taught explicitly, it may be that a higher level of speaking performance is needed to be able to rate accurately.

If this particular element of the speaking test were of importance to the construct of the test given, then the test developers or test administrators should be aware that there is a potential for native speaker and non-native speaker raters (or raters of different speaking proficiency levels) to have significant differences in how this category is interpreted. If raters are interpreting this factor differently, than they may be operating on a different construct of proficiency, which may have an impact on the ratings. Even though differences that appeared in this subcategory did not transfer to the holistic, overall rating, it is important to note the native and non-native speaker raters appear to have different approaches or values of different rating elements in their path to a final score. This may imply some lack of uniformity among the raters if both native and non-native rater speaker groups were used without discretion.

In a similar vein, Zhang and Elder (2011) found significant differences in native and non-native speaker raters comments about the ratings that they gave, even though there were not significant differences in the overall ratings themselves. This study found differences in the number of times native and non-native speakers made comments about the examinee’s linguistic resources, including vocabulary and general linguistic resources, interaction, demeanor, and compensation strategies. In most cases native speaker raters had more to say than non-native speaker raters, but in some particular sub-categories the non-native speakers made more comments than native speaker raters when justifying their ratings. Non-native speakers were
more prolific on topics such as the completeness and appropriateness of the response and group participation, as well as a number of comments that couldn’t be more specifically characterized.

In these cases, the test developer and/or administrator must consider which procedures are appropriate for the test context. If the linguistic subcategories are important to the construct of what is being measured overall, then perhaps is appropriate to use native speakers only. If the holistic score is the only important conclusion that is made from the test, then native speakers as well as non-native speakers may both be included in the rater population. In any case, each of these decisions should be expressly justified in the test specifications. It cannot be taken for granted that native speakers are ideal speakers of the language, that they have any particular levels speaking proficiency, or that they are automatically qualified to evaluate even their own language. Assumptions such as the native speaker has appropriate intuitions to evaluate language or that a speaker with a certain level of proficiency is necessary in order to be able to rate accurately may make sense on the surface, but there is little data to support these theories. In conclusion, if the native speaker concept is used and justified, than the native speaker must be clearly defined so that there is no ambiguity in understanding the research or the test construct, and so that the test or research study can be comparable to other like studies.

7.3 Local implications

Several practical local implications result from the conclusions made from this research study. The most immediate and obvious of these is that the FBI may be able to use some currently trained, active non-native speaker raters with an English proficiency level of ILR 3 or above to rate English speaking proficiency tests. This policy decision could greatly expand the number of English rater resources available to the language testing program at the FBI. With the
present requirement for rater candidacy being a native speaker and demonstrating an ILR Level 5 proficiency in English, it is extremely difficult to find qualified candidates, even in an English-speaking country such as the United States. Cross training additional raters from other current language rater pools may save the FBI time and resources in recruiting and training efforts.

The fact that no differences were found between the native on non-native speaker groups supports the position that non-native speakers can acquire the language well enough to perform evaluative tasks, at least with the assistance of extensive training. What may be more important than demonstrating language proficiency in order to be an evaluator is the development of a high-level cognitive and linguistic competence which would support being able to make accurate discriminations in high-level language samples. The non-native speaker raters (who were highly competent in at least two languages) were in many cases able to draw upon these other competencies to overcome any potential lack of language proficiency in English to be able to rate as accurately as native speakers. This held true for ILR English proficiency Levels 3 – 4+, but not for the ILR Level 2+ raters, who were unable to perform the rating task accurately.

Another practical implication for the FBI is that it may be useful to consider changing the testing format from two people both giving and rating the test simultaneously to having a format where one person administers the test and another person only rates the test. If this test protocol were in place, the FBI could make better use of the qualified non-native speaker raters. It is important to note that the results of this research did not speak to whether or not non-native speakers or ILR Level 3 to 4+ speakers would be able to administer English speaking proficiency tests, but only to rate them. Until further research can be done, testing protocol could be altered so that one English native speaker tester administers the tasks (and rates the examinee’s
responses) and a certified English rater, either native or non-native speaker, evaluates the examinee’s responses.

A couple of negative practical implications should also be considered in making decisions based on conclusions from this research. Although it is positive that less effort might be needed to find qualified rater candidates to attend rater training, broadening the pool of rater candidates might result in an influx of ‘false positives’ at the workshop, that is rater candidates that meet the prerequisite requirements to attend the training, but are not successful during the training. By allowing ILR Level 3 or 4 speakers of English to attempt to qualify as English raters, which requires attendance at the two-week long workshop and successful completion of the English rater certifying exams, the FBI will be expending significant additional resources to potentially gain only a small number of additional qualified raters.

Another issue that must be dealt with is whether or not the examinees will have confidence in the ability of the non-native speaker rater, and whether the non-native speaker rater would have confidence in their own abilities. In the past examinees have mistakenly believed that they were being evaluated by a non-native speaker, and took great issue with the fact that they did not believe that it would be a fair assessment. There have been several instances where examinees have appealed their ratings at the FBI with the complaint that their evaluators were not native speakers, and therefore not qualified to assess them. This type of complaint is even more prevalent when the examinee is a native speaker. The examinees had the perception, like the perception of researchers and testing organizations that have used native speaker as a model or as evaluators, that only a native speaker is qualified to conduct language evaluations. If these new rater policies are adopted, the FBI will have to do its best to educate its examinees to understand that their evaluation will not be impacted by the fact that a non-native speaker of the
language is rating them. This information could be supported by statistics from quality assurance reviews on speaking tests.

In addition to the examinees having confidence in the abilities of the non-native raters, the non-native raters need to develop confidence in themselves in several cases. During the course of the study, there were several occasions where the non-native raters, all experienced raters in other languages and highly-proficiency speakers of English, emailed the researcher or commented in the rating justification that they did not feel qualified to evaluate English because they were not a native speaker. It could be that the FBI’s practice of not allowing non-native speakers to be used as raters created a de facto policy and belief that the FBI did not have confidence in a non-native’s ability, and that perception permeated the rater group. Even if that were not the cause of the raters’ concerns about their own abilities, Davies (2004) discusses that this sentiment is not uncommon among non-native speakers. He supports the idea that there are psycholinguistic rather than sociolinguistic differences between native and non-native speakers, particularly in speed and grammaticality judgments. When linguistic features ratings were examined among the rater groups, there were never any significant differences in grammaticality judgments (the structures ratings) and only a few differences between Level 2 raters and higher proficiency raters in speed (the structures ratings). This lack of significant differences supports Davies’ idea that nativeness is not a measure of ability, but of identity and confidence (2003).

In the end, it was found that the dichotomous categorization of raters as native or non-native speakers was not a useful distinction to be made in the determination of overall rater ability. In fact the use of the term native speaker appears to do more harm than good in the context of rater candidate qualifications. Any time a candidate who considers himself a native speaker of the language is rejected as a rater because of a deficiency in some other area of rating
ability, the candidate could potentially misinterpret the decision as reflecting on his or her native speaker identity. Native speaker ability is closely linked to identity, and when this type of miscommunication happens, rater candidates are likely to take offense. The candidate feels that his or her ability to use the language is insufficient and therefore a piece of his or her identity is insufficient. It would be better, therefore, to avoid the use of native speaker in determining the qualifications of a rater candidate and to rely more on judgments made of other rater competencies or language proficiency, reserving the term native speaker to refer to the method of language acquisition.

7.4 Limitations and future research

There are several limitations in the current research study that naturally lead to the need for additional research. The first limitation is that this study does not address whether or not native speakers who have proficiency levels less than ILR Level 5 could be effective raters of English. This question could not be addressed here because the parameters of participant selection were limited to currently available FBI testers. At present, there are not any native speaker raters with less than and ILR Level 5 rating who are trained because it is not permitted to expend resources to train or certify someone who, it is believed, would not be operationally useful. The researcher would have to administer extensive training to the new native speaker ILR Level 4 and lower raters to create a broader sample. These new rater participants would have to undergo extensive practice administering and rating speaking tests to have the experience that other raters in the sample currently have. This would have taken more financial and time resources than were available.
As was discussed in the discussion, it takes more than a level of speaking proficiency to be able to be an effective rater. It takes high level of linguistic, cognitive, and evaluative competencies, not to mention cultural competency as well. Due to the fact that all of the non-native speaker participants in this study are current raters of their first language, they all have high levels of these competencies, giving them the bilingual advantage. It is yet to be determined whether or not a person who does not have these rater competencies and is not a Level 5 speaker of the rated language would be successful as a rater.

In addition, this study was done entirely in the context of English speaking proficiency tests. It cannot be determined at this point if the results of this study would apply to other languages. Further research needs to be carried out to see if this study’s conclusions are particular to English or if they are cross-linguistic. Also, the context of this research that of a native English speaking environment, the United States. Most of the participants have been livening in the United States for some time, if not most of their lives. Almost all of the non-native speaker participants are naturalized United States citizens. Moreover, most of the participants work with Americans in English speaking professional environments, and many work for the United States Federal Government full time. With this background, it is likely that the participants have favorable opinions to American culture and the English language as it is spoken in the United States.

Finally, there was significant additional data collected as a part of this research study, but because of time and scope limitations, they were not analyzed and included. For example, on each of the rater reports, full justifications were written out, explicating the rationale the raters used to arrive at their final decisions. All 25,000 of these comments were coded for analysis. However, analysis of the comments was beyond the scope of the current project. In addition,
transcripts are available from discussions between rater pairs after evaluating language tests. These post-test discussions could give insight into the rating process and the particular weight raters give to different aspects of language. These rater perspectives on the rating process and construct could be cross-linguistically compared, and would potentially result in a better understanding of the differences in rater perspectives among the different groups of language testers.
Appendix A: Interagency Language Roundtable Language Skill Level Descriptions -

Speaking

Preface The following proficiency level descriptions characterize spoken language use. Each of the six "base levels" (coded 00, 10, 20, 30, 40, and 50) implies control of any previous "base level's" functions and accuracy. The "plus level" designation (coded 06, 16, 26, etc.) will be assigned when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level." The "plus level" descriptions are therefore supplementary to the "base level" descriptions. A skill level is assigned to a person through an authorized language examination. Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively. Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details. Unless otherwise specified, the term "native speaker" refers to native speakers of a standard dialect. "Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education; however, in cultures where formal higher education is common, the language-use abilities of persons who have had such education are considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language.
**Speaking 0 (No Proficiency)** Unable to function in the spoken language. Oral production is limited to occasional isolated words. Has essentially no communicative ability. (Has been coded L-0 in some nonautomated applications.) [Data Code 0]

**Speaking 0+ (Memorized Proficiency)** Able to satisfy immediate needs using rehearsed utterances. Shows little real autonomy of expression, flexibility or spontaneity. Can ask questions or make statements with reasonable accuracy only with memorized utterances or formulae. Attempts at creating speech are usually unsuccessful. **Examples:** The individual's vocabulary is usually limited to areas of immediate survival needs.

Most utterances are telegraphic; that is, functors (linking words, markers and the like) are omitted, confused or distorted. An individual can usually differentiate most significant sounds when produced in isolation but, when combined in words or groups of words, errors may be frequent. Even with repetition, communication is severely limited even with people used to dealing with foreigners. Stress, intonation, tone, etc. are usually quite faulty. (Has been coded S-0+ in some nonautomated applications.) [Data Code 06]

**Speaking 1 (Elementary Proficiency)** Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics. A native speaker must often use slowed speech, repetition, paraphrase, or a combination of these to be understood by this individual. Similarly, the native speaker must strain and employ real-world knowledge to understand even simple statements/questions from this individual. This speaker has a functional, but limited proficiency. Misunderstandings are frequent, but the individual is able to ask for help and to verify comprehension of native speech in face-to-face interaction. The individual is unable
to produce continuous discourse except with rehearsed material. **Examples:** Structural accuracy is likely to be random or severely limited. Time concepts are vague. Vocabulary is inaccurate, and its range is very narrow. The individual often speaks with great difficulty. By repeating, such speakers can make themselves understood to native speakers who are in regular contact with foreigners but there is little precision in the information conveyed. Needs, experience or training may vary greatly from individual to individual; for example, speakers at this level may have encountered quite different vocabulary areas. However, the individual can typically satisfy predictable, simple, personal and accommodation needs; can generally meet courtesy, introduction, and identification requirements; exchange greetings; elicit and provide, for example, predictable and skeletal biographical information. He/she might give information about business hours, explain routine procedures in a limited way and state in a simple manner what actions will be taken. He/she is able to formulate some questions even in languages with complicated question constructions. Almost every utterance may be characterized by structural errors and errors in basic grammatical relations. Vocabulary is extremely limited and characteristically does not include modifiers. Pronunciation, stress, and intonation are generally poor, often heavily influenced by another language. Use of structure and vocabulary is highly imprecise. (Has been coded S-1 in some nonautomated applications.) [Data Code 10]

**Speaking 1+ (Elementary Proficiency, Plus)** Can initiate and maintain predictable face-to-face conversations and satisfy limited social demands. He/she may, however, have little understanding of the social conventions of conversation. The interlocutor is generally required to strain and employ real-world knowledge to understand even some simple speech. The speaker at this level may hesitate and may have to change subjects due to lack of language resources. Range
and control of the language are limited. Speech largely consists of a series of short, discrete utterances. **Examples:** The individual is able to satisfy most travel and accommodation needs and a limited range of social demands beyond exchange of skeletal biographic information.

Speaking ability may extend beyond immediate survival needs. Accuracy in basic grammatical relations is evident, although not consistent. May exhibit the more common forms of verb tenses, for example, but may make frequent errors in formation and selection. While some structures are established, errors occur in more complex patterns. The individual typically cannot sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited. Person, space and time references are often used incorrectly. Pronunciation is understandable to natives used to dealing with foreigners. Can combine most significant sounds with reasonable comprehensibility, but has difficulty in producing certain sounds in certain positions or in certain combinations. Speech will usually be labored. Frequently has to repeat utterances to be understood by the general public. (Has been coded S-1+ in some nonautomated applications.) [Data Code 16]

**Speaking 2 (Limited Working Proficiency)** Able to satisfy routine social demands and limited work requirements. Can handle routine work-related interactions that are limited in scope. In more complex and sophisticated work-related tasks, language usage generally disturbs the native speaker. Can handle with confidence, but not with facility, most normal, high-frequency social conversational situations including extensive, but casual conversations about current events, as well as work, family, and autobiographical information. The individual can get the gist of most everyday conversations but has some difficulty understanding native speakers in situations that require specialized or sophisticated knowledge. The individual's utterances are minimally
cohesive. Linguistic structure is usually not very elaborate and not thoroughly controlled; errors are frequent. Vocabulary use is appropriate for high-frequency utterances, but unusual or imprecise elsewhere. **Examples:** While these interactions will vary widely from individual to individual, the individual can typically ask and answer predictable questions in the workplace and give straightforward instructions to subordinates. Additionally, the individual can participate in personal and accommodation-type interactions with elaboration and facility; that is, can give and understand complicated, detailed, and extensive directions and make non-routine changes in travel and accommodation arrangements. Simple structures and basic grammatical relations are typically controlled; however, there are areas of weakness. In the commonly taught languages, these may be simple markings such as plurals, articles, linking words, and negatives or more complex structures such as tense/aspect usage, case morphology, passive constructions, word order, and embedding. (Has been coded S-2 in some nonautomated applications.) [Data Code 20]

**Speaking 2+ (Limited Working Proficiency, Plus)** Able to satisfy most work requirements with language usage that is often, but not always, acceptable and effective. The individual shows considerable ability to communicate effectively on topics relating to particular interests and special fields of competence. Often shows a high degree of fluency and ease of speech, yet when under tension or pressure, the ability to use the language effectively may deteriorate. Comprehension of normal native speech is typically nearly complete. The individual may miss cultural and local references and may require a native speaker to adjust to his/her limitations in some ways. Native speakers often perceive the individual's speech to contain awkward or inaccurate phrasing of ideas, mistaken time, space and person references, or to be in some way
inappropriate, if not strictly incorrect. **Examples:** Typically the individual can participate in most social, formal, and informal interactions, but limitations either in range of contexts, types of tasks or level of accuracy hinder effectiveness. The individual may be ill at ease with the use of the language either in social interaction or in speaking at length in professional contexts. He/she is generally strong in either structural precision or vocabulary, but not in both. Weakness or unevenness in one of the foregoing, or in pronunciation, occasionally results in miscommunication. Normally controls, but cannot always easily produce general vocabulary. Discourse is often incohesive. (Has been coded S-2+ in some nonautomated applications.) [Data Code 26]

**Speaking 3 (General Professional Proficiency)** Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations in practical, social and professional topics. Nevertheless, the individual's limitations generally restrict the professional contexts of language use to matters of shared knowledge and/or international convention. Discourse is cohesive. The individual uses the language acceptably, but with some noticeable imperfections; yet, errors virtually never interfere with understanding and rarely disturb the native speaker. The individual can effectively combine structure and vocabulary to convey his/her meaning accurately. The individual speaks readily and fills pauses suitably. In face-to-face conversation with natives speaking the standard dialect at a normal rate of speech, comprehension is quite complete. Although cultural references, proverbs and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation. Pronunciation may be obviously foreign. Individual sounds are accurate: but stress, intonation and pitch control may be faulty. **Examples:** Can typically
discuss particular interests and special fields of competence with reasonable ease. Can use the language as part of normal professional duties such as answering objections, clarifying points, justifying decisions, understanding the essence of challenges, stating and defending policy, conducting meetings, delivering briefings, or other extended and elaborate informative monologues. Can reliably elicit information and informed opinion from native speakers. Structural inaccuracy is rarely the major cause of misunderstanding. Use of structural devices is flexible and elaborate. Without searching for words or phrases, the individual uses the language clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers. Errors occur in low-frequency and highly complex structures. (Has been coded S-3 in some nonautomated applications.) [Data Code 30]

**Speaking 3+ (General Professional Proficiency, Plus)** Is often able to use the language to satisfy professional needs in a wide range of sophisticated and demanding tasks.

**Examples:** Despite obvious strengths, may exhibit some hesitancy, uncertainty, effort or errors which limit the range of language-use tasks that can be reliably performed. Typically there is particular strength in fluency and one or more, but not all, of the following: breadth of lexicon, including low- and medium-frequency items, especially socio-linguistic/cultural references and nuances of close synonyms; structural precision, with sophisticated features that are readily, accurately and appropriately controlled (such as complex modification and embedding in Indo-European languages); discourse competence in a wide range of contexts and tasks, often matching a native speaker's strategic and organizational abilities and expectations. Occasional patterned errors occur in low frequency and highly-complex structures. (Has been coded S-3+ in some nonautomated applications.) [Data Code 36]
**Speaking 4 (Advanced Professional Proficiency)** Able to use the language fluently and accurately on all levels normally pertinent to professional needs. The individual's language usage and ability to function are fully successful. Organizes discourse well, using appropriate rhetorical speech devices, native cultural references and understanding. Language ability only rarely hinders him/her in performing any task requiring language; yet, the individual would seldom be perceived as a native. Speaks effortlessly and smoothly and is able to use the language with a high degree of effectiveness, reliability and precision for all representational purposes within the range of personal and professional experience and scope of responsibilities. Can serve as in informal interpreter in a range of unpredictable circumstances. Can perform extensive, sophisticated language tasks, encompassing most matters of interest to well-educated native speakers, including tasks which do not bear directly on a professional specialty.

**Examples:** Can discuss in detail concepts which are fundamentally different from those of the target culture and make those concepts clear and accessible to the native speaker. Similarly, the individual can understand the details and ramifications of concepts that are culturally or conceptually different from his/her own. Can set the tone of interpersonal official, semi-official and non-professional verbal exchanges with a representative range of native speakers (in a range of varied audiences, purposes, tasks and settings). Can play an effective role among native speakers in such contexts as conferences, lectures and debates on matters of disagreement. Can advocate a position at length, both formally and in chance encounters, using sophisticated verbal strategies. Understands and reliably produces shifts of both subject matter and tone. Can understand native speakers of the standard and other major dialects in essentially any face-to-face interaction. (Has been coded S-4 in some nonautomated applications.) [Data Code 40]
**Speaking 4+ (Advanced Professional Proficiency, Plus)** Speaking proficiency is regularly superior in all respects, usually equivalent to that of a well educated, highly articulate native speaker. Language ability does not impede the performance of any language-use task. However, the individual would not necessarily be perceived as culturally native. **Examples:** The individual organizes discourse well, employing functional rhetorical speech devices, native cultural references and understanding. Effectively applies a native speaker's social and circumstantial knowledge; however, cannot sustain that performance under all circumstances. While the individual has a wide range and control of structure, an occasional nonnative slip may occur. The individual has a sophisticated control of vocabulary and phrasing that is rarely imprecise, yet there are occasional weaknesses in idioms, colloquialisms, pronunciation, cultural reference or there may be an occasional failure to interact in a totally native manner. (Has been coded S-4+ in some nonautomated applications.) [Data Code 46]

**Speaking 5 (Functionally Native Proficiency)** Speaking proficiency is functionally equivalent to that of a highly articulate well-educated native speaker and reflects the cultural standards of the country where the language is natively spoken. The individual uses the language with complete flexibility and intuition, so that speech on all levels is fully accepted by well-educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms and pertinent cultural references. Pronunciation is typically consistent with that of well-educated native speakers of a non-stigmatized dialect. (Has been coded S-5 in some nonautomated applications.) [Data Code 50]
# Appendix B: Speech Evaluation Guide

## LEVEL 0 AND LEVEL 1

<table>
<thead>
<tr>
<th>Category</th>
<th>LEVEL 0 (0+)</th>
<th>LEVEL 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>• No proficiency.</td>
<td>• Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics.</td>
</tr>
<tr>
<td><strong>Hallmarks</strong></td>
<td>• Essentially no communicative ability.</td>
<td>• Listener must strain and employ real-world knowledge to understand even simple statements/questions.</td>
</tr>
<tr>
<td></td>
<td>• Even with repetition, communication is severely limited even with people used to dealing with foreigners.</td>
<td>• Misunderstandings are frequent.</td>
</tr>
<tr>
<td><strong>Functions</strong></td>
<td>• Unable to function in the spoken language.</td>
<td>• Can maintain very simple conversations on familiar topics.</td>
</tr>
<tr>
<td></td>
<td>• Usually limited to areas of immediate survival needs.</td>
<td>• Can satisfy predictable, simple, personal and accommodation needs.</td>
</tr>
<tr>
<td></td>
<td>• Can generally meet courtesy, introduction, and identification requirements.</td>
<td>• Can give simple information.</td>
</tr>
<tr>
<td></td>
<td>• Can ask for help and verify comprehension</td>
<td>• Can formulate some questions.</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>• Displays none.</td>
<td>• Cannot produce continuous discourse except with rehearsed material.</td>
</tr>
<tr>
<td><strong>Structures</strong></td>
<td>• Speaks in isolated words.</td>
<td>• Has accuracy that is random or severely limited.</td>
</tr>
<tr>
<td></td>
<td>• Speaks with memorized utterances or formulae.</td>
<td>• Uses time concepts that are vague.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Has structural errors and errors in basic grammatical relations in almost every utterance.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Uses structures that are highly imprecise.</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>• Uses occasional isolated words.</td>
<td>• Has vocabulary that is extremely limited and characteristically does not include modifiers.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Has vocabulary that is inaccurate, and very narrow.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• May have encountered quite different vocabulary areas.</td>
</tr>
<tr>
<td><strong>Fluency</strong></td>
<td>• Has attempts at creating speech that are usually unsuccessful.</td>
<td>• Often speaks with great difficulty.</td>
</tr>
<tr>
<td><strong>Pronunciation</strong></td>
<td>• Has stress, intonation, tone, etc. that are usually quite faulty.</td>
<td>• Has pronunciation, stress, and intonation that are generally poor, often heavily influenced by another language.</td>
</tr>
<tr>
<td><strong>Social/Cultural Appropriateness</strong></td>
<td>• Has little to no awareness, but may be evident in memorized utterances or formulae.</td>
<td>• Can generally meet courtesy, introduction, and identification requirements.</td>
</tr>
</tbody>
</table>

---

256
## Level 1 and Level 2

<table>
<thead>
<tr>
<th>Category</th>
<th>Level 1</th>
<th>Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>• Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics.</td>
<td>• Able to satisfy routine social demands and limited work requirements.</td>
</tr>
<tr>
<td><strong>Hallmarks</strong></td>
<td>• Listener must strain and employ real-world knowledge to understand even simple statements/questions. • Misunderstandings are frequent.</td>
<td>• Speaks with confidence, but not with facility. • In work-related tasks that are more complex than routine interactions, language usage generally disturbs the native speaker. • Errors are frequent.</td>
</tr>
<tr>
<td><strong>Functions</strong></td>
<td>• Can maintain very simple conversations on familiar topics.</td>
<td>• Can handle routine work-related interactions that are limited in scope.</td>
</tr>
<tr>
<td></td>
<td>• Can satisfy predictable, simple, personal and accommodation needs. • Can ask for help and verify comprehension.</td>
<td>• Can participate in personal and accommodation-type interactions with elaboration and facility.</td>
</tr>
<tr>
<td></td>
<td>• Can generally meet courtesy, introduction, and identification requirements.</td>
<td>• Can handle most normal, high-frequency social conversational situations including extensive, but casual conversations about current events, as well as work, family, and autobiographical information.</td>
</tr>
<tr>
<td></td>
<td>• Can formulate some questions. • Can give simple information.</td>
<td>• Can typically ask and answer predictable questions in the workplace. • Can give straightforward instructions to subordinates. • Can give complicated, detailed, and extensive directions.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>When performing the functions/tasks for Level 1, the examinee...</th>
<th>When performing the functions/tasks for Level 2, the examinee...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organization</strong></td>
<td>• Cannot produce continuous discourse except with rehearsed material.</td>
<td>• Uses utterances that are minimally cohesive.</td>
</tr>
<tr>
<td><strong>Structures</strong></td>
<td>• Has structural errors and errors in basic grammatical relations in almost every utterance.</td>
<td>• Uses simple structures and basic grammatical relations that are typically controlled; however, there are areas of weakness</td>
</tr>
<tr>
<td></td>
<td>• Has accuracy that is random or severely limited. • Uses time concepts that are vague. • Uses structures that are highly imprecise.</td>
<td>• Uses linguistic structure that is usually not very elaborate and not thoroughly controlled. • Uses errors that are frequent.</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>• Has vocabulary that is extremely limited and characteristically does not include modifiers. • Has vocabulary that is inaccurate, and very narrow. • May have encountered quite different vocabulary areas.</td>
<td>• Can participate in personal interactions with elaboration and facility. • Uses vocabulary that is appropriate for high-frequency utterances, but unusual or imprecise elsewhere.</td>
</tr>
<tr>
<td><strong>Fluency</strong></td>
<td>• Often speaks with great difficulty.</td>
<td>• Speaks with confidence, but not with facility.</td>
</tr>
<tr>
<td><strong>Pronunciation</strong></td>
<td>• Has pronunciation, stress, and intonation that are generally poor, often heavily influenced by another language.</td>
<td>• Has mispronunciations that sometimes result in miscommunication.</td>
</tr>
<tr>
<td><strong>Social/Cultural Appropriateness</strong></td>
<td>• Can generally meet courtesy, introduction, and identification requirements.</td>
<td>• Can handle most normal, high-frequency social conversational situations including extensive but casual conversations.</td>
</tr>
</tbody>
</table>
# Level 2 and Level 3

<table>
<thead>
<tr>
<th>Category</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>• Able to satisfy routine social demands and limited work requirements.</td>
<td>• Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations in practical, social and professional topics.</td>
</tr>
<tr>
<td><strong>Hallmarks</strong></td>
<td>• Speaks with confidence, but not with facility.</td>
<td>• Commits errors that virtually never interfere with understanding and rarely disturb the native speaker.</td>
</tr>
<tr>
<td><strong>Functions</strong></td>
<td>• Can handle routine work-related interactions that are limited in scope.</td>
<td>• Professional contexts including matters of shared knowledge and/or international convention.</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>• Uses utterances that are minimally cohesive.</td>
<td>• Uses discourse that is cohesive.</td>
</tr>
<tr>
<td><strong>Structures</strong></td>
<td>• Uses simple structures and basic grammatical relations that are typically controlled; however, there are areas of weakness</td>
<td>• Can effectively use structures to convey meaning accurately.</td>
</tr>
<tr>
<td></td>
<td>• Uses linguistic structure that is usually not very elaborate and not thoroughly controlled.</td>
<td>• Has structural inaccuracy, but it is rarely the major cause of misunderstanding.</td>
</tr>
<tr>
<td></td>
<td>• Commits frequent errors.</td>
<td>• Commits errors in low-frequency and highly complex structures.</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>• Can participate in personal interactions with elaboration and facility.</td>
<td>• Can effectively use vocabulary to convey meaning accurately.</td>
</tr>
<tr>
<td></td>
<td>• Uses vocabulary that is appropriate for high-frequency utterances, but unusual or imprecise elsewhere.</td>
<td>• Uses the language clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers without searching for words or phrases.</td>
</tr>
<tr>
<td><strong>Fluency</strong></td>
<td>• Speaks with confidence, but not with facility.</td>
<td>• Speaks readily and fills pauses suitably.</td>
</tr>
<tr>
<td><strong>Pronunciation</strong></td>
<td>• Has mispronunciations that sometimes result in miscommunication.</td>
<td>• Commits errors that virtually never interfere with understanding and rarely disturb the native speaker.</td>
</tr>
<tr>
<td><strong>Social/Cultural Appropriateness</strong></td>
<td>• Can handle most normal, high-frequency social conversational situations including extensive but casual conversations.</td>
<td>• Can participate effectively in most formal and informal conversations in practical, social and professional topics. Although cultural references, proverbs and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation.</td>
</tr>
<tr>
<td>Category</td>
<td>When performing the functions/tasks for Level 3, the examinee...</td>
<td>When performing the functions/tasks for Level 4, the examinee...</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>• Uses discourse that is cohesive.</td>
<td>• Organizes discourse well, using appropriate rhetorical speech devices.</td>
</tr>
<tr>
<td><strong>Structures</strong></td>
<td>• Can effectively use structures to convey meaning accurately.</td>
<td>• Uses structures accurately on all levels normally pertinent to professional needs.</td>
</tr>
<tr>
<td></td>
<td>• Has structural inaccuracy, but it is rarely the major cause of misunderstanding.</td>
<td>• Uses structures with a high degree of effectiveness, reliability and precision.</td>
</tr>
<tr>
<td></td>
<td>• Commits errors in low-frequency and highly complex structures.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Uses structural devices that are flexible and elaborate.</td>
<td>• Uses sophisticated verbal strategies.</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>• Can effectively use vocabulary to convey meaning accurately.</td>
<td>• Uses vocabulary with a high degree of effectiveness, reliability and precision.</td>
</tr>
<tr>
<td></td>
<td>• Uses the language clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers without searching for words or phrases.</td>
<td>• Can make unfamiliar concepts clear and accessible.</td>
</tr>
<tr>
<td></td>
<td>• Speaks readily and fills pauses suitably.</td>
<td>• Can perform extensive, sophisticated language tasks.</td>
</tr>
<tr>
<td><strong>Fluency</strong></td>
<td>• Speaks readily and fills pauses suitably.</td>
<td>• Speaks effortlessly and smoothly.</td>
</tr>
<tr>
<td><strong>Pronunciation</strong></td>
<td>• Commits errors that virtually never interfere with understanding and rarely disturb the native speaker.</td>
<td>• May not be perceived as a native.</td>
</tr>
<tr>
<td><strong>Social/Cultural Appropriateness</strong></td>
<td>• Can participate effectively in most formal and informal conversations in practical, social and professional topics.</td>
<td>• Can set the tone of interpersonal official, semi-official and non-professional verbal exchanges with a representative range of speakers (in varied audiences, purposes, tasks and settings).</td>
</tr>
<tr>
<td></td>
<td>• Although cultural references, proverbs and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation.</td>
<td>• Reliably produces shifts of both subject matter and tone.</td>
</tr>
<tr>
<td></td>
<td>• Uses appropriate rhetorical speech devices and native cultural references.</td>
<td>• Uses sophisticated verbal strategies.</td>
</tr>
<tr>
<td>Category</td>
<td>LEVEL 4</td>
<td>LEVEL 5</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Overall</td>
<td>• Able to use the language fluently and accurately on all levels normally pertaining to professional needs. The individual's language usage and ability to function are fully successful.</td>
<td>• Speaking proficiency is functionally equivalent to that of a highly articulate well-educated native speaker and reflects the cultural standards of the country where the language is natively spoken.</td>
</tr>
<tr>
<td>Hallmarks</td>
<td>• Uses the language with a high degree of effectiveness, reliability and precision.</td>
<td>• Functionally equivalent to that of a highly articulate well-educated native speaker.</td>
</tr>
<tr>
<td>Functions*</td>
<td>• Can perform extensive, sophisticated language tasks, encompassing most matters of interest to well-educated native speakers, including tasks which do not bear directly on a professional specialty.</td>
<td>• The individual uses the language with complete flexibility and intuition, so that speech on all levels is fully accepted by well-educated native speakers in all of its features, including breadth of vocabulary and idioms, colloquialisms and pertinent cultural references.</td>
</tr>
<tr>
<td>Organization*</td>
<td>• Organizes discourse well, using appropriate rhetorical speech devices.</td>
<td>• Uses the language with complete flexibility and intuition.</td>
</tr>
<tr>
<td>Structures*</td>
<td>• Uses structures accurately on all levels normally pertinent to professional needs.</td>
<td>• Controls structural on all levels is fully accepted by well-educated native speakers.</td>
</tr>
<tr>
<td>Vocabulary*</td>
<td>• Uses vocabulary with a high degree of effectiveness, reliability and precision.</td>
<td>• Speech on all levels is fully accepted by well-educated native speakers in all of its features, including breadth of vocabulary.</td>
</tr>
<tr>
<td>Fluency*</td>
<td>• Speaks effortlessly and smoothly.</td>
<td>• Uses the language with complete flexibility and intuition.</td>
</tr>
<tr>
<td>Pronunciation*</td>
<td>• May not be perceived as a native.</td>
<td>• Has pronunciation that is typically consistent with that of well-educated native speakers of a non-stigmatized dialect.</td>
</tr>
<tr>
<td>Social/Cultural Appropriateness*</td>
<td>• Can set the tone of interpersonal official, semi-official and non-professional verbal exchanges with a representative range of speakers (in varied audiences, purposes, tasks and settings).</td>
<td>• Uses speech that reflects the cultural standards of the country where the language is natively spoken.</td>
</tr>
<tr>
<td></td>
<td>• Reliably produces shifts of both subject matter and tone.</td>
<td>• Uses speech on all levels that is fully accepted by well-educated native speakers in all of its features, including idioms, colloquialisms and pertinent cultural references.</td>
</tr>
<tr>
<td></td>
<td>• Uses appropriate rhetorical speech devices and native cultural references.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Uses sophisticated verbal strategies.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix C: Individual Tester Report (ITR)

INDIVIDUAL TESTER REPORT (ITR)
Speaking Proficiency Test (SPT)

SPT Language | English | TAP # | Examinee’s Initials
---|---|---|---
Test Date | 6/29/11 | Examinee’s SSN: xxx-xx-
Time | 10:00 | 2546
Field Office | Testing Assistant
Tester 1 | M. Bell | Tester 2

ILR Base Level | 3 | Individual ILR Rating | 3

FUNCTIONS and CATEGORY LEVELS

<table>
<thead>
<tr>
<th>Functions</th>
<th>Organization</th>
<th>Structures</th>
<th>Vocabulary</th>
<th>Fluency</th>
<th>Pronunciation</th>
<th>Social/Cultural Appropriateness</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

DESCRIPTION of EXAMINEE’S PERFORMANCE

GENERAL
Examinee was quite fluent and exhibited sophistication in lexicon but in no other areas of the language.

SPECIFIC

- Functions: able to support opinion, describe in detail, deal with complexities, narrate, manage abstractions and discuss topics on a global level;
- Organization: discourse was cohesive, and ideas were presented in a clean manner; no strategic organizational techniques, though
- Structures: controlled basic structures and complex ones, as well; no marks of sophistication; sporadic article omissions, NS errors
- Vocabulary: lexicon was broad and occasionally sophisticated (e.g., done tastefully, gamut, tangible results, well-maintained, interaction, nebulous, gratuitous, repulsive, dehumanizing, committed atrocities, coherent); occasional awkward expression or imprecision (e.g., ‘we’re now turning back into phonics-based;’ ‘action’ for ‘reaction;’ ‘keeping strings to their homeland’); idiomatic expressions incorporated with accuracy (e.g., ‘tie the line, stick with it, hang out, blew his chances, naysayer, foot the bill’); had a tendency to relay personal experiences as part of his responses, though it didn’t seem to be for lack of lexicon, but rather it seemed a mark of his conversation style
- Fluency: speaks readily and with ease; occasional fillers, though nothing disruptive to communication; at times, rapidity occasionally resulted in collapsed syllables or false starts
- Pronunciation: native
- Social/Cultural Appropriateness: able to manage an interactive situation requiring persuasion; able to present an argument in a formal setting with appropriate register

I affirm that I have read, in full, the ILR Skill Level Description for Speaking for the final level assigned and that, based on my assessment, the examinee’s performance at least meets, if not exceeds, every feature of that description.
INDIVIDUAL TESTER REPORT (ITR)
Speaking Proficiency Test (SPT)

TAP #
Examinee’s Initials CH
Examinee’s SSN: xxx-xx-2546

TEST INFORMATION

TOPICS (at least five)
1. music and education
2. changes in classroom over 20 yrs.
3. home neighborhood
4. his life compared to daughter’s growing up
5. impact of play life environment on daughter's life
6. labeling of games and music
7. winning strategy for presidential candidates
8. Kerry win, impact on war on terror
9. Bush decision to withdraw troops in former Easter Bloc
10. Muslim population in the U.S.
11. American way of life

SITUATIONS or INFORMATION GATHERING TASK (IGT)

<table>
<thead>
<tr>
<th>Situation/IGT Topic</th>
<th>Role of Tester</th>
<th>Role of Examinee</th>
<th>Task</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop smoking</td>
<td>smoker</td>
<td>friend</td>
<td>persuade friend to stop smoking</td>
<td>register was appropriate, though did not set tone; persuasion technique was gently; used a concerned but frank tone</td>
</tr>
<tr>
<td>school board speech</td>
<td>parent rep</td>
<td></td>
<td>make formal argument re: music programs</td>
<td>introduction was brief, did not acknowledge board members explicitly; did not set tone, though register was appropriately formal; content was appropriate and thorough; speech progressed logically, but organization was not strategic</td>
</tr>
</tbody>
</table>

BREAKDOWNS

<table>
<thead>
<tr>
<th>Probe</th>
<th>Task</th>
<th>Level</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>both situations</td>
<td>tone-setting</td>
<td>4</td>
<td>did not strategically set tone in either situation</td>
</tr>
</tbody>
</table>

OTHER COMMENTS ABOUT TEST or EXAMINEE

Testers did not probe, so it was difficult to identify breakdown. Rating is based primarily on performance and overall weaknesses/shortcomings.

☐ I affirm that I have completed this ITR prior to any discussion with my co-tester.
References


265


268


273


Ortega, L. (2010, March 6). The bilingual turn in SLA. Plenary presented at the American Association of Applied Linguistics, Atlanta, GA.


