NATIVE AND NON-NATIVE RATERS OF L2 SPEAKING PERFORMANCE: ACCENT FAMILIARITY AND COGNITIVE PROCESSES

By Valeriia Bogorevich

A Dissertation

Submitted in Partial Fulfillment

of the Requirements for the degree of

Doctor of Philosophy in Applied Linguistics

Northern Arizona University

May 2018

Approved:

Soo Jung Youn, Ph.D., Chair

Okim Kang, Ph.D., Co-chair

Vedran Dronjic, Ph.D.

Luke Plonsky, Ph.D.

Sara Abercrombie, Ph.D.

ABSTRACT

NATIVE AND NON-NATIVE RATERS OF L2 SPEAKING PERFORMANCE: ACCENT FAMILIARITY AND COGNITIVE PROCESSES

VALERIIA BOGOREVICH

Rater variation in performance assessment can impact test-takers' scores and compromise assessments' fairness and validity (Crooks, Kane, & Cohen, 1996). Rater variation can also undermine a test's validity and fairness; therefore, it is important to investigate raters' scoring patterns in order to inform rater training. Substantial work has been done analyzing rater cognition in writing assessment (e.g., Cumming, 1990; Eckes, 2008); however, few studies have tried to classify factors that could contribute to rater variation in speaking assessment (e.g., May, 2006).

The present study used a mixed methods approach (Tashakkori & Teddlie, 1998; Greene, Carcelli, & Graham, 1989) to investigate the potential differences between native English-speaking and non-native English-speaking raters in how they assess L2 students' speaking performance. Kane's (2006) argument-based approach to validity was used as the theoretical framework. The study challenged the plausibility of the assumptions for the evaluation inference, which links the observed performance and the observed score and depends on the assumption that the raters apply the scoring rubric accurately and consistently.

The study analyzed raters' scoring patterns when using a TOEFL iBT speaking rubric analytically. The raters provided scores for each rubric criterion (i.e., Overall, Delivery, Language Use, and Topic Development). Each rater received individual training, practice, and calibration experience. All the raters filled out a background questionnaire asking about their teaching experiences, language learning history, the background of students in their classrooms, and their exposure to and familiarity with the non-native accents used in the study.

For the quantitative analysis, the two groups of raters 23 native (North American) and 23 non-native (Russian) raters graded and left comments for speech samples from Arabic (*n* = 25), Chinese (*n* = 25), and Russian (*n* = 25) L1 background. Students' samples were in response to two independent speaking tasks; the students' responses varied from low to high proficiency levels. For the qualitative part, 16 raters (7 native and 9 non-native) shared their scoring behavior through think-aloud protocols and interviews. The speech samples graded during the think-aloud included Arabic (*n* = 4), Chinese (*n* = 4), and Russian (*n* = 4) speakers.

Raters' scores were examined using the Multi-Faceted Rasch Measurement using FACETS (Linacre, 2014) software to test group differences between native and non-native raters as well as raters who are familiar and unfamiliar with the accents of students in the study. In addition, raters' comments were coded and also used to explore rater group differences. The qualitative analyses involved thematical coding of transcribed think-aloud sessions and interview sessions using content analysis (Strauss & Corbin, 1998) to investigate the cognitive processes of raters and their perceptions of their rating processes. The coding included such themes as decision-making and re-listening patterns, perceived severity, criteria importance, and non-rubric criteria (e.g., accent familiarity, L1 match). Afterward, the quantitative and qualitative results were analyzed together to describe the potential sources of rater variability. This analysis was done employing side-by-side comparison of qualitative and quantitative data (Onwuegbuzie & Teddlie, 2003).

The results revealed that there were no radical differences between native and non-native raters; however, some different patterns were observed. Non-native raters also showed more lenient grading patterns towards the students with whom their L1 matched. In addition, all raters, regardless of the group, demonstrated several patterns of rating depending on their focus while listening to examinees’ performance and interpretations of the rating criteria during the decision-making process. The findings can motivate professionals who oversee and train raters at testing companies and intensive English programs to study their raters’ scoring behaviors to individualize training to help make exam ratings fair and raters interchangeable.

Acknowledgements

Many individuals made important contributions to my dissertation project. First and foremost, I thank my dissertation committee, Dr. Soo Jung Youn, Dr. Okim Kang, Dr. Vedran Dronjic, Dr. Luke Plonsky, and Dr. Sara Abercrombie, for many useful comments and suggestions made during the process. I sincerely appreciate their time, support, and constructive feedback that made each draft better. In addition, I am grateful to the Applied Linguistics faculty at Northern Arizona University for sharing their wealth of knowledge with me. Also, I would like to thank the Program in Intensive English for providing access to archived research data.

The study would not have been possible without financial support. Data collection was directly supported by the National Federation of Modern Language Teachers Associations Dissertation Support Grant and the TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment. I thank NFMLTA and ETS for their generous support that greatly facilitated the timely completion of my dissertation.

Although they must remain anonymous, I would like to thank all the individuals who participated as raters, transcribers, and coders in the study. I appreciate their efforts and time spent completing this lengthy process with minimal compensation. In addition, I extend my gratitude to a number of individuals who assisted in recruiting participants for the study, including Julia Kachanova, Anna Kationova, Julia Galichanina, and Maria Pustovalova.

I am grateful to many colleagues and friends who have supported me before and throughout the Ph.D. program. It is not possible to mention everyone, but special heartfelt thanks go to Diana Proskuriakova, whose example encouraged me to start this endeavor; Aleksei Andreyevich Koshelev and Yelena Alekseyevna Kosheleva whose guidance as advisors at Voronezh Regional Educational Advising Center was indispensable; Anya Sonnikova, Rita

Privalova, Roman Guzovskiy, Katya and Dima Panykovi, Dima Cherezov, and many other VREAC volunteers for their advice and help; Elnaz Kia, Ally Kermad, Panjanit Chaipuapae, Nina and Roma Lesnovi, Diana Solano, Alex Novikov, and Alina Kurnosova for their encouragement and support throughout my Ph.D. studies.

Lastly and most importantly, completion of this dissertation would have been impossible without the support and love of my husband, Alex, and my son, Sasha. I am greatly thankful for the happy moments and their never-ending optimism and reassurance.

Table of Contents

ABSTRACT ii
Acknowledgements vi
List of Tables xiii
List of Figures xiv
Chapter 1: Introduction 15
Background of the Problem 15
Statement of the Problem 23
Purpose of the Study 24
Significance of the Study 25
Definitions of Terms 25
Summary 27
Dissertation Organization 28
Chapter 2: Literature Review 29
Challenges of Testing L2 Performance 29
Theoretical Framework: Argument-Based Approach to Validity 31
Strengths of Kane's validity framework. 35
Inferences and assumptions in performance assessment. 36
Rater Variation as a Validity Threat 38
Common validity threats for evaluation inference. 38
Rater cognition. 39
Native and non-native raters. 43
Accent familiarity. 48
Present Study 52
Research Questions 54
Chapter 3: Method 56
Research design 56
Pilot Study 58
Participants 59
Raters. 60
Examinees 62
Coders 63

Instruments 63
Speaker background questionnaire 63
Rater background questionnaire 64
Accent familiarity scale 64
Speaking prompts 65
Rating rubric 66
Procedures 66
Data collection. 66
Examinee data preparation. 66
Chinese and Arabic. 66
Variability of Arabic and Chinese accents 67
Russian 68
Selection process 68
Noise, length, and content screening 69
Benchmarking 69
Recordings for individual ratings 70
Recordings for reflective think-aloud 70
Rating Procedures 70
Recruitment and screening 71
Individual rater training Skype sessions 71
Rubric and task familiarization 72
Norming 72
Calibration practice 73
Individual rating 73
Raters' comments 74
Individual think-aloud and interview Skype sessions 75
Rater selection for qualitative inquiry 75
Benchmark review 77
Reflective rating with think-aloud protocols 77
Interview questions 79
Accent familiarity 80
Data Analyses 81

Quantitative analyses. 82

MFRM analyses. 82

Sub-question a. 83

Sub-questions b. 84

Sub-question c. 84

Sub-question d. 84

Sub-question e. 85

Content Analysis. 86

Coding scheme for raters' comments. 86

Coding of raters' comments. 88

Sub-question f. 88

Analyses of think-aloud protocols and interviews. 89

Chapter 4: Results 91

Overall Facets Results 92

Facets summary 92

Explanation of measure and fit statistics 95

Examinee measurement report 95

Rater measurement report 97

Rating criteria measurement report 99

Scale functioning 100

Comparison of NS and NNS Rater Groups Based on the Scores 101

Raters' internal consistency and severity 102

Raters severity and consistency by examinee L1 104

Bias analyses 107

Accent familiarity of NS and NNS raters 109

Relationship between familiarity, severity, and consitency 112

Comparison between NS and NNS Rater Groups Based on the Comments 115

Number of overall comments 115

Comment direction: negative, positive, and neutral 116

Comment direction: Delivery, Language Use, Topic Development, and General 117

Rater types: negative and positive 117

Summary 122

Qualitative Results ... 123
NS and NNS raters ... 124
Listening Strategies ... 124
Note-takers and hard listeners. ... 124
Multitaskers. ... 128
Re-listeners. ... 130
Grading Strategies ... 132
Top-down and bottom-up. ... 134
Changing a scoring decision. ... 136
References to Non-Rubric Criteria ... 141
Non-Rubric Criteria for Delivery. ... 142
Voice quality. ... 142
Accent familiarity. ... 146
Familiarity and unfamiliarity. ... 146
Ability to recognize accents. ... 150
Whose bias? ... 152
Non-Rubric Criteria for Topic Development. ... 156
Finished or unfinished responses. ... 156
Organization. ... 158
Reading the prompt. ... 159
Making a choice. ... 160
Idea quality. ... 161
Perceived severity ... 163
Perceived rating criteria importance ... 165
Summary ... 169
Mixed Methods Results ... 170
Statistical and Perceived Severity ... 171
Rater Consistency and Decision-Making Strategies ... 173
Summary ... 175
Chapter 5: Discussion ... 176
NS and NNS raters ... 177
Consistency and severity. ... 177

Rubric criteria utilization. 180
Raters' Cognitive Processes 189
Raters' accent familiarity 194
Accent familiarity of NS and NNS. 195
Familiar and unfamiliar groups of raters. 198
L1 match. 200
Chapter 6: Conclusion 204
Implications of the Study 204
Methodological. 204
Practical. 206
Limitations and Directions for Future Research 209
Concluding Remarks 212
References 214

List of Tables

Table 1. Rater Demographic Information 60
Table 2. Total Number of Recordings by Each L1 63
Table 3. Recordings for Rater Training and Calibration Practice per Rater 69
Table 4. Recordings for Individual Rating 70
Table 5. Recordings for Reflective Rating with Think-Aloud 71
Table 6. Rater Severity and Infit Measures 76
Table 7. Measurement Report for Speakers (Arranged by Infit Mean Square) 97
Table 8. Measurement Report for Raters (Arranged by Infit Values) 98
Table 9. Measurement Report for Rating Criteria (Arranged by Difficulty Measure) 100
Table 10. Measurement Report for NS Raters (Arranged by Infit Values) 102
Table 11. Measurement Report for NNS Raters (Arranged by Infit Values) 103
Table 12. Measurement Report for NS and NNS Groups (Arabic L1) 105
Table 13. Measurement Report for NS and NNS Groups (Chinese L1) 105
Table 14. Measurement Report for NS and NNS Groups (Russian L1) 105
Table 15. Measurement Report for Criteria for NS (Sorted by Difficulty) 107
Table 16. Measurement Report for Criteria for NNS (Sorted by Difficulty) 107
Table 17. Measurement Report for Rating Criteria (Arranged by Difficulty Measure) 109
Table 18. Rater Familiarity with Examinees' L1s (With Identification) 110
Table 19. Rater Familiarity with Examinees' L1s (Without Identification) 111
Table 20. Measurement Report for Familiar and Unfamiliar Groups (Arabic L1) 113
Table 21. Measurement Report for Familiar and Unfamiliar Groups (Chinese L1) 114
Table 22. Measurement Report for Familiar and Unfamiliar Groups (Russian L1) 114
Table 23. Number and Percentage of Negative, Positive and Neutral Comments by Rater Group 116
Table 24. Number and Percentage of Comments by NSs and NNSs (by Criteria) 117
Table 25. Number and Percentage of Negative, Positive and Neutral Comments Provided by NS Raters 118
Table 26. Number and Percentage of Negative, Positive and Neutral Comments Provided by NNS Raters 119
Table 27. Number and Percentage of Comments Provided by NS Raters (by Criteria) 120
Table 28. Number and Percentage of Comments Provided by NNS Raters (by Criteria) 121
Table 29. Raters' Perceived Severity 164
Table 30. Raters' Perceived Criteria Importance 166
Table 31. Raters' Statistical and Perceived Severity 172

List of Figures

Figure 1. Types of inferences in an interpretive argument (left) and an inference structure (right). 33
Figure 2. Overview of the mixed methods research design. 57
Figure 3. Variable map 93
Figure 4. Probability curves for the 4-point scale. 101
Figure 5. Rater group familiarity with examinees' L1s (in percentages). 111
Figure 6. Accent familiarity scales that formed composite accent familiarity score. 112

Chapter 1: Introduction

Language testers have always been interested in rater variation that occurs in raters scoring L2 performance assessment (i.e., writing and speaking). Research has shown that raters differ in their scores for the same written essay (e.g., Barkaoui, 2007) or spoken sample (Orr, 2002). Language assessment research studies have shown that raters can differ in their approaches to scoring, interpreting scoring criteria, focus on rubric criteria, and employed non-rubric criteria based on raters' background characteristics, such as language background, rating experience, or amount of rater training (e.g., Kim, 2015; Zhang & Elder, 2014). The fact that raters differ undermines the assumption for the evaluation inference of the validity argument for performance assessment (Crooks, Kane, & Cohen, 1996). If the evaluation inference is at stake, the whole validity argument is undermined (Kane, 2006). The present dissertation challenged the assumption of the evaluative inference viz. that raters utilize scoring criteria appropriately by taking into account raters' first language (L1) background, accent familiarity, and raters' cognitive processes while scoring.

Background of the Problem

Unlike reading and listening assessments, which are more objective, speaking and writing tests, which are examples of performance assessment, are susceptible to subjectivity due to the nature of humans. The potential rater subjectivity can be a source of error affecting reliability and validity of performance assessments (Dunbar, Koretz, & Hoover, 1991). For example, if speaking proficiency is measured through fluency of speech delivery, then variance due to fluency of speech delivery in this ability of examinees might be construct-relevant variance because it is directly related to the measured construct. On the other hand, variance due to rater background, training, decision-making processes, accent familiarity, and attitudes can be considered construct-irrelevant because these characteristics are not part of examinees' fluency of speech delivery. According to Haladyna and Downing (2004), the construct-irrelevant variance can exist in any test to what extent? Such type of variance threatens the validity and fairness of evaluations and successively causes unfair uses of scores (Hoyt & Kerns, 1999; McNamara, 1996; Weigle, 2002; Weir, 2005). Thus, it is important to detect any potential sources of construct-irrelevant variance and address them appropriately through rater training.

One important issue outlined by the assessment research is individual rater differences. Researchers have often been interested in differences in rater characteristics such as rater variation and rater bias when scoring L2 performance assessment. This topic has been scrutinized from different angles: decision-making processes, individual characteristics, and familiarity with a test-taker accent (e.g., Brown, 2000; Eckes, 2009; Wei & Llosa, 2015). Though all of these topics are important, the present research has focused on the rater differences regarding speaking performance assessment; however, studies on the topics of rater characteristics and behavior in writing performance assessment are also mentioned where relevant.

Differences in scores that raters give to the same writing or speaking sample can be roughly subdivided into two categories: *how* raters differ and *why* raters differ. Studies in the first category have examined the extent to which raters agree or differ in the scores they give to the same writing (e.g., Barkaoui, 2007, 2010; Cumming, 1990; Vaughan, 1991) or speaking sample (e.g., Ang-Aw & Goh, 2011; Orr, 2002). Studies in the second category have looked at the possible reasons for such differences (e.g., Davis, 2012, 2015; Kim, 2011, 2015; Kim, 2009; Zhang & Elder, 2011, 2014). The studies in the first category attempted to answer the question of how raters differ by analyzing differences in raters' approach to scoring, interpreting scoring criteria, and examining raters' focus on the rubric and non-rubric criteria. The studies in the second category attempted to answer the question of why raters differ by analyzing rater language background, rating experience, and the amount of rater training. To date, there has been more research done in writing assessment trying to answer how raters differ (e.g., Cumming, 1990; Eckes, 2008), whereas research in speaking assessment has been more focused on answering why raters may differ (e.g., Davis, 2015; Kim, 2011).

Studies in writing and speaking assessment have usually employed two research methods to answer the aforementioned questions: quantitative Multi-Faceted Rasch Analysis (MFRM) and qualitative analysis of raters' verbal reports or written comments. MFRM studies examined differences in rater severity, consistency, and interactions with other aspects of rating such as gender, rubric criteria, and rater group (i.e., native vs. non-native raters or experienced vs. novice raters) (e.g., Bonk & Ockey, 2003; Brown, 1995; Eckes, 2005; Hiseh, 2011). First, MFRM allows detection of potential rater characteristics from a statistical perspective, including rater leniency/severity, centrality, inaccuracy, and differential dimensionality or, in other words, differential rater functioning (DRF) (Wolf & McVay, 2004). Among these, DRF is one of the most difficult rater traits to study statistically. For example, DRF related to test-takers' gender or L1 background is known to easily be statistically expressed; however, DRF related to examinees' handwriting legibility in writing or speech comprehensibility in speaking cannot be easily detected and statistically expressed (Wolf & McVay, 2004). Even though statisticians have tried to address this problem, qualitative analyses can provide richer data.

Qualitative studies have scrutinized what raters attend to while grading performance tasks not only to reveal rater differences but also to create a new rubric or to validate an existing one (e.g., Bown, 2005; Brown, Iwashita, & McNamara, 2005; Wei & Llosa, 2015). There have been far more qualitative studies regarding rater scoring processes in writing than in speaking. Thus, there is still a need for more qualitative studies in assessing speaking since the quantitative approach cannot account for psychological or cognitive processes that underlie the rating process; as Connor-Linton (1995) stated, “if we do not know what raters are doing ... then we do not know what their ratings mean” (p. 763). Also, the cognitive research that has focused on raters’ verbal reports can help researchers advance our understanding of the influence of the rater decision-making processes on performance assessment.

Given the complex nature of rater variation, a mixed methods approach is deemed to be the most appropriate for the current study. Mixed methods designs allow researchers to analyze the problem from both quantitative and qualitative perspectives and thereby compensate for the weakness of one method by the strength of the other (Creswell & Piano Clark, 2007; Hesse-Biber & Leavy, 2006; Teddlie & Tashakkori, 2006). The present study will employ the concurrent complementarity design (Greene, Carcelli, & Graham, 1989), where the qualitative results will enhance, clarify, broaden, elaborate, and increase the interpretability and meaningfulness of the quantitative results.

Performance assessment is a demanding field for researchers to study because it is hard to measure the appropriateness of individual performance and ascertain which rater’s score is closer to the ideal true score. One concern raised by Lindemann and Subtirelu (2013) is that consistency of raters’ scores can stem from the similar attitudes, beliefs and biases; therefore, there is a need for “further investigation of the validity of assessments of L2 speech accuracy based on listeners’ ratings, even when those ratings are consistent” (p. 547). This statement was made in the speech perception context where scoring rubrics are not used to accompany numeric scale points, and L2 speech assessment might be less affected in such a way. Nevertheless, it may be possible that groups of raters share certain rating processes to give scores, which may be contaminated by construct-irrelevant variance caused by raters' beliefs. Thus, a qualitative approach is indispensable for investigating such possibilities.

One area of research on rater variability has addressed the possible group differences that might be caused by raters' native- or non-native-speaker status. Language testers have raised concerns that native and non-native raters may differ in terms of their understanding of certain aspects of rating, for example, cultural communication norms (e.g., Brown, 1995) or written rhetorical patterns (e.g., Kobayashi & Rinnert, 1996), which may cause differences in scores. Another argument that is given to support the prospective differences is that non-native raters can have very diverse backgrounds or come from an area with established English dialects. Such backgrounds of non-native raters can affect their ability to evaluate language performances. Studies comparing native and non-native raters have been done in writing assessment (e.g., Johnson & Lim, 2009; Shi, 2001), speech perception and pronunciation (Fayer & Krasinski, 1987; Kang, 2012; Saito & Shintani, 2016), and speaking assessment (Kim, 2009; Zhan & Elder, 2014). Some of the studies showed that non-native speakers are more severe (e.g., Zhang & Elder, 2014; Brown, 1995; Fayer & Krasinski, 1987; Kang, 2012) or that native speakers are more severe (e.g., Barnwell, 1989) whereas other studies showed no differences (Xi & Mollaun, 2009; Wei & Llosa, 2015).

Studies comparing native and non-native raters differ in their findings and contradict one another. One explanation is that the differences in the outcomes of the studies may be attributed to differences in rater populations and research designs. The studies that have compared native and non-native raters have been done in speaking or writing; with or without a rubric; grading mono or multi-lingual students; looking at English, Spanish, or Arabic; with or without rater training; involving naïve and experienced raters as well as teachers and non-teachers. Some studies (e.g., Zhang & Elder, 2011) showed that the quantitative difference could not be seen, but some differences can be uncovered using a qualitative approach. Zhang and Elder (2014) pointed out that native and non-native "raters may arrive at their judgments via somewhat different pathways and show different degrees of tolerance of breakdowns in relation to particular features of speech" (p. 318).

The differences that may occur when comparing *native* to *non-native* raters are also seen when comparing *native* to *native* raters grading speaking (e.g., Chalhoub-Deville 1995; Chalhoub-Deville & Wigglesworth, 2005). Chalhoub-Deville and Wigglesworth's study compared native speakers from four native speaker backgrounds. The results illustrated that the U.S. raters were most lenient, U.K. raters most severe, and Canadian and Australian raters were in-between. An interesting explanation for these results was offered in the area of educational measurement by Suto (2012) who stated that rater agreement or disagreement could depend on their "community of practice" or "school of thought." The author suggested that "it is likely that raters of equal experience and eminence would hold different understandings of what constitutes a good response and interpret the scoring criteria slightly differently, despite common training on those questions" (p. 23). Another research study also showed the discrepancies among native raters due to their L2 background, because they were heritage speakers, or communicated with non-native speakers of a similar L1 background on a regular basis (Winke & Gass, 2013).

The studies discussed above suggest that there may be a difference in raters that is not only driven by the native or non-native affiliation but also based on raters' familiarity and exposure to other people who speak similarly (or with a similar accent) due to the same L1 background. In support of this idea, some studies have suggested that raters' familiarity expressed through knowledge of test-takers' L1 would impact their ratings (e.g., Winke, Gass & Myford, 2013). In another study, Carey, Mannell, and Dunn (2011) looked at the impact of raters' residence in the examinees' country. Both studies revealed that familiarity and exposure affected raters' scores.

The fundamental issue that needs to be explored is rater variation, and what can cause it. The fact that some studies have shown that several raters can give the same score to the same candidate driven by their own performance perceptions and scoring rubric interpretations shows variability in raters' cognitive processes. Such latent rater variability questions the validity of the rendered test scores. Research into the cognitive processes of human raters can introduce improvements for rubric development, rater selection, and rater training in order to support the validity and fairness of performance assessments (Bejar, 2012). According to Bejar, existing research on rater cognition in various assessment contexts has been focused on two major areas: "the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (p. 2). There are multiple rater attributes that can cause variation in raters' cognitive processes such as raters' gender, age, educational background, ESL/EFL teaching experience, language background (e.g., native/non-native raters, matches between rater and examinee L1 or L2 language background), and rater training experience. The current study is interested in investigating raters' decision-making processes during the assessment of test-takers' speaking performance taking into account raters' language background (native vs. non-native raters) and raters' familiarity with test-takers' way of speaking.

The current study focuses on semi-direct monologic tests in order to explore rater differences without any additional variations that can be added by grading paired conversations or interviews. The performance-based speaking ability of language learners can be tested through an interview, a dialog, or a monolog. Speaking assessment can be administered *directly* when test-takers are communicating with a person and *semi-directly* when examinees are recording their answer using technology (Qian, 2009). These two types of administration differ in lexical density, but both have been claimed to yield highly correlated scores (O'Loughlin, 1995). The interlocutors in a direct speaking test may add more variance that might be difficult to interpret (Stansfield & Kenyon, 1992a, 1992b); therefore, the semi-direct method is more preferable. Moreover, the semi-direct method of testing speaking is more practical because it enables raters to not be on the testing site, and allows recruitment of raters from all over the world. Although lacking interactivity, the semi-direct testing of speaking has multiple advantages such as cost-effectiveness and efficiency; therefore, it is widely used for L2 speaking language exams.

The speaking test responses and tasks that will be used in the present study are from an Intensive English Program (IEP) placement test; however, they are also typical for proficiency tests (e.g., TOEFL). Proficiency tests and the placement test used in the current study are similar because, based on the scores from a placement test, a decision can be made about whether a student is proficient enough to start mainstream university classes bypassing more English as a second language instruction. In addition, Xi (2010) explicitly describes the two-folded purpose of TOEFL iBT as a proficiency as well as a placement test, "For placement purposes, the TOEFL iBT scores can be used alone or along with an in-house English placement test to exempt international students from taking English classes or to place them into English support classes" (p. 156). Thus, the concerns raised in the current study can be applicable for high-stakes proficiency exams as well as for the in-house placement exams administered by Intensive English Programs.

Statement of the Problem

Construct-irrelevant variance can have a significant effect on test-takers' scores. Such variance can be caused by the differences between NS and NNS, differences in the level of accent familiarity, and decision-making differences between NS and NNS.

Overall, the studies on how and why raters differ have yielded mixed findings. First, studies addressing the differences of NS versus NNS have yielded different results in speaking assessment. Second, there are still relatively few studies on how NS and NNS ESL/EFL teachers approach the scoring of speaking exams using a well-established rating rubric. Thus, there is a need for a well-designed study grounded in the common practices of speaking assessment. Evidence from such an investigation will help clarify the differences/similarities between NS and NNS raters' scoring of speaking performance. This issue is important because the testing companies are already using NNS speakers as raters or are planning to expand their rater pool by hiring NNS (Winke & Gass, 2013).

The second concern that needs more attention in L2 speaking assessment research is the role that the level of accent familiarity plays. Unlike writing, speaking is susceptible to an additional source of variation that is caused by raters' level of familiarity with a test-taker's accent, which can lead to a positive or negative bias. Raters who are familiar with test-takers' L1s or share test-takers' L1s might give higher scores to such L1 test takers. The other side of this perspective can also be true: raters can assign lower scores to test-takers whose L1s are completely unfamiliar to them (Wei, 2015). In other words, it is possible that raters who are very familiar with some interlanguage phonologies or who share the interlanguage with examinees can show positive bias towards those examinees (e.g., more lenient scoring). On the other hand, the occurrence of negative bias (e.g., more severe scoring) can be present when raters are unfamiliar with some interlanguage phonology of the examinees to be scored. This issue is important because it can compromise the fairness of test results; however, research addressing this issue has been rare.

In addition to the aforementioned issues, there can also be strategic differences between NS and NNS in terms of the process of arriving at analytic or holistic scores. The research literature on performance language assessment reveals that raters tend to interpret rating rubrics differently because they may assign different importance to some criteria or some aspects of examinees' performance and that their interpretations of the rubric may interact with the task variable (Moere, 2014). It is possible that NS and NNS speakers can show systematic variation in terms of their focus on some specific criteria, but there has been little research addressing this topic in the field of speaking assessment.

Ultimately, it is crucial to identify and minimize any unwanted construct-irrelevant sources of variance that can have a significant effect on test-takers' scores. Thus, the present study analyzes the differences between NS and NNS in terms of quantitative scores, explores the effects of rater accent familiarity on scores, and compares the cognitive processes of NS and NNS raters while rating L2 speaking performance.

Purpose of the Study

The purpose of the current study is to focus on the differences between native and non-native raters' cognitive decision-making processes while scoring examinees' speaking performance and to analyze any emerging patterns of other factors that could contribute to rater variability such as the use of non-rubric criteria. The present research helps to compile more specific recommendations for improving rater training and rater monitoring for performance assessment. These recommendations and guidelines would be an asset in ensuring the development of unbiased rating ability. The purpose of the study is also explained in more details in the Present Study section.

Significance of the Study

The study adds to the accumulating body of literature on rater bias in assessment contexts and provides valuable insights for the field of performance assessment, specifically, assessment of L2 speaking. The study broadens the understanding of rater variation that may have an effect on raters' ability to rate accurately, consistently, and with a uniform severity level. The study categorizes potential non-rubric factors that affect raters' scoring ability, which deepens the knowledge about the effects of rater linguistic background on rating scores assigned, specifically the effects of accent familiarity or unfamiliarity. The study also provided backing to the fact that proficient non-native speakers can be used for scoring speaking exams. The study also outlines recommendations for improving rater training and rater monitoring for performance assessment as well as lays down the foundation for material development for training native- and nonnative-speaker groups of raters.

Definitions of Terms

A number of concepts are defined in this section in order to provide readers with a clear idea of their exact interpretation and use in the study. Definitions of major terms are provided below in alphabetical order.

Accent familiarity. Based on two subjective measures, raters' perceived accent familiarity was operationalized as a composite accent familiarity score (on the scale from 11 to 66). For the first measure, familiarity with L1 identification, the raters self-reported their familiarity with the examinee L1s used in the study (i.e., Arabic, Chinese, and Russian) in terms of general familiarity, communication, and teaching experience. For the second measure, familiarity without L1 identification, the raters listened to 24 unidentified 12-second recordings (eight from each L1) and reported their accent familiarity with each of them. For both measures, the 6-point scale was used to describe raters' perceived accent familiarity (No, Very Little, Little, Some, A Lot, Extensive).

Construct-irrelevant variance. Construct-irrelevant variance represents unexplained random or systematic variance which does not relate to the rated construct and is happening due to other unrelated reasons (Messick, 1993; Haladyna, Downing, 2004; McNamara, 1996).

Speaking ability. Speaking ability is defined as the ability of test-takers to use English for communication as measured by a semi-direct monologic speaking test and graded using the descriptors of TOEFL iBT independent rubric, which includes: delivery, language use, and topic development. In semi-direct speaking assessments, examines record their answers to questions using technology (e.g., a voice recorder) without an interlocutor (Qian, 2009).

Speaking assessment. The studies on speaking assessment are defined as research projects that involved raters qualified enough to be hired by a testing company to grade L2 speaking ability by using a well-established valid speaking rubric, and who were provided with a minimum amount of rater training before attempting this task. Speech perception and speech processing studies where raters make instantaneous judgments about pronunciation concepts (e.g., comprehensibility) using unguided Likert scale rubrics are not included as part of the speaking assessment definition.

Performance assessment. Performance assessment refers to a type of assessment which predicts test-takers real-life abilities based on their responses to a sample task (Miller, Linn & Gronlund, 2009).

Performance assessment is defined as the type of assessment that requires test-takers to apply knowledge and demonstrate skills by performing a simulation of a real-world task (e.g., a speech sample or an essay) rather than choose from pre-made answer choices. Based on the definition by the Standards for Educational and Psychological Testing, performance assessment is "product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCTM], 1999, p. 186)

Rater cognition. Rater cognition is a mental process of decision-making (e.g., rating scale application) that raters go through to arrive at a score (Bejar, 2012; Wood, 2014).

Rater bias. In language assessment research, the general term of bias has been described as examinees' unequal chances of answering an item correctly due to a construct-irrelevant factor (Angoff, 1993; Harding, 2012; Zumbo, 2007). In turn, rater bias can be defined as raters' unequal treatment of examinees due to construct-irrelevant factors. Rater bias can be subdivided into two levels – positive and negative; positive rater bias happens when examinees get higher scores than deserved, whereas negative rater bias results in lower scores.

Summary

This chapter contextualized the study by addressing the problems and gaps in the current literature on rater characteristics and bias in L2 speaking performance. The chapter also outlined the purpose of the dissertation and provided the necessary background information to support the need for the present study. Key terms as they were used in the dissertation were defined. In this chapter, it was argued that it is important to investigate the differences between native and non-native raters based on the potential differences in their decision-making processes while rating speaking performance. In addition, it emphasized the importance of singling out the prospective non-rubric factors (e.g., accent familiarity), which can contribute to rater variation.

Dissertation Organization

The current Introduction chapter is followed by Chapter 2 which provides the review of the relevant literature on the topics of performance assessment. First, it defines performance assessment and its challenges. Second, it provides detailed explanations and outlines the advantages of Kane's (1992, 2006) argument-based approach to validity, which was used as the theoretical framework for the study. Third, sources of variability in performance assessment are reviewed, such as rater cognitive processes, language background, and accent familiarity. Chapter 2 ends with the description of the present study and lists research questions. Chapter 3 provides details the research design, gives information about the pilot study, and then describes the participants, instruments, and data collection procedures. Chapter 4 provides the quantitative and qualitative results, and Chapter 5 discusses the research questions. Lastly, in Chapter 6, the dissertation ends by looking at the study's limitations, implications, and directions for future research.

Chapter 2: Literature Review

This chapter reviews the theoretical and empirical literature that is relevant to the present study. First, the chapter defines performance assessment, then describes the validity framework, and lastly overviews the potential sources of variability in rating that are rater cognitive decision-making processes, rater status (native/non-native), and accent familiarity.

Challenges of Testing L2 Performance

The need to measure language ability in target language use situations has been argued for since the 1960s when the concept of performance testing of language abilities was introduced. Performance assessment did not originate in the field of language testing per se, but in the broader field of general education contexts; however, it appeared to be relevant to language testing. Researchers emphasized that testing separate elements such as pronunciation, grammar, and vocabulary does not account for the extent to which learners could actually use them. On the other hand, assessment of examinees' L2 performance allows making inferences about test-takers' true language abilities in target situations (McNamara, 1996).

There have been attempts to emphasize the importance of performance in L2 assessment during the spread of the communicative approach. For example, Carroll (1961) and Davies (1968) argued that testing learners' performance has internal validity to make decisions based on the scores and external validity that allows generalization of the results suitable for language proficiency testing. However, the construct of performance assessment was not well-established. Spolsky (1968) and other researchers were also advocating for inclusion of performance assessment to language testing because this type of assessment taps the actual competence of learners. Later, Clark (1972) and Savignon (1972) elaborated on the construct of L2 performance assessment by describing practical suggestions.

Nowadays, performance assessment is defined as the type of assessment that requires test-takers to apply knowledge and demonstrate skills by performing a simulation of a real-world task (e.g., a speech sample or an essay) rather than choose from pre-made answer choices. For example, the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCTM], 1999) define the performance assessment as "Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied" (p. 186). Moreover, performance assessment is also sometimes referred to as direct, authentic, or alternative assessment (Miller, Linn & Gronlund, 2009). Performance assessment is considered to be a better estimate of learners' skills because it is characterized by more authentic and complex tasks; however, both critics and proponents of performance assessment question the objectivity of the obtained scores because they are provided by human raters.

Since the first implementation of performance-based assessment of speaking, researchers have been interested in the factors affecting test-takers' scores, specifically, rater effects have been holding researchers' attention in the field of L2 speaking assessment. Because of the fact that performance assessment usually utilizes human raters to score test-takers using a scoring rubric, there are several ways in which subjectivity of rater judgments can cause variations. For example, raters can bring in their own personal judgment standards and their own level of self-consistency. Moreover, differences in raters' practical rating experience with various tests or with a specific test, raters' training, educational background, and teaching experience can add variance to the scores given by different raters to students with the same ability level (Lumely, 2005). Bachman, Lynch, and Mason (1995) also emphasized that the potential sources of undesirable measurement error can be rater inconsistency or bias towards a task or a test-taker. There are multiple potential sources of rater variability including rater internal and external consistency, rater severity, quality of the rating scale, task demands, the occasion of rating, and interaction with other aspects of the rating process (Brown, 1995; Lumley & McNamara, 1995; Wigglesworth, 1993).

The fact that raters can bring in subjectivity undermines the soring quality that is critical to argue the validity and fairness of a test. Even though raters use a rubric to grade, the given scores can be unreliable due to the nature of performance assessment that is subjected to uncontrollable variation. This uncontrollable variance is usually called construct-irrelevant variance (Messick, 1993; Haladyna, Downing, 2004; McNamara, 1996) and is not random but systematic meaning that it is group- or person-specific. The construct-irrelevant variance occurs not only due to personal bias such as rater background, severity, and stereotypes, but it can also occur when raters do not follow the rubric or because of rater interactions with it. For example, raters can bring construct-irrelevant variance if they differ in their beliefs about which criterion on the rubric is more important than other criteria (Miller et al., 2009). In addition, raters might avoid lower and higher bands of the rubric or middle bands of the rubric (McNamara, 1996). As we can see, construct-irrelevant variance threatens the validity and fairness of the scores and must be controlled for in order to diminish its impact on test-takers' scores.

Theoretical Framework: Argument-Based Approach to Validity

Validity framework has been chosen as the theoretical framework in the current study, specifically, argument-based approach to validity, which was outlined by Cronbach (1988) and refined by Kane (1992, 2001, 2006). According to Kane (2013, 2016), a test itself or test scores cannot hold property of being valid; validity is the evaluation of plausibility and appropriateness of the proposed test score interpretations and uses. The kinds of evidence that are needed for validation are dependent on the test's interpretations and uses; therefore, a careful analysis of these interpretations and uses is needed to provide a clear basis for evidence selection.

Kane's (2006) argument-based validity framework approaches validation through the interpretive and validity arguments. The interpretative argument is the specification of the proposed interpretations and uses of scores, and the validity argument is the evaluation of their plausibility and appropriateness. The interpretive argument describes the interpretations and uses "by laying out the network of inferences and assumptions leading to the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, p 23.). In its turn, the validity argument evaluates the plausibility of the interpretive argument to justify the interpretation. In other words, the validity argument critically evaluates and challenges the plausibility and appropriateness of the inferences and assumptions stated in the interpretive argument. The interpretations and uses that are supported by the evidence are considered to be valid, and the interpretations and uses that are not supported are invalid.

The validity argument is based on the interpretive argument. The interpretive argument consists of inferences and assumptions as the building blocks. It is critical to understand what *inference* means in this context because inferences play a crucial role in the development of the interpretive argument (Kane, 1992, 2001; Mislevy, Steinberg, & Almond, 2003). The understanding of the notion of inference is based on the Toulmin's (2003) structure of an argument. An argument consists of several components: grounds or data, claim, warrant, backing, and rebuttal, where an inference helps to move from grounds to a claim (Chapelle, Enright, & Jamieson, 2008, 2010). A warrant is used to justify an inference, and any warrant needs backing. On the other hand, a rebuttal can be used to weaken an inference (Figure 1).

An example that is similar to one described in Chapelle et al. (2010) can help one to understand the structure of an argument. An example of an inference about an ESL student's speaking performance based on grammar structures can serve this purpose. In this example, an *inference* can be made that a student's speaking ability is low because the grammatical structures used by the student are incorrect. The observation of the incorrect grammatical structures is the *grounds*. The *claim* in this example is that the student's speaking ability is low. Nevertheless, the grounds do not directly lead to the claim, but the *inference* connects them. The inference must be justified using a warrant that states that one can infer the claim from the grounds. In the present example, the *warrant* would be that incorrect grammatical structures are characteristic of students whose speaking ability is low. The warrants themselves are usually not self-evident and need support. Thus, the *backing* is needed to provide the appropriate evidence to support the

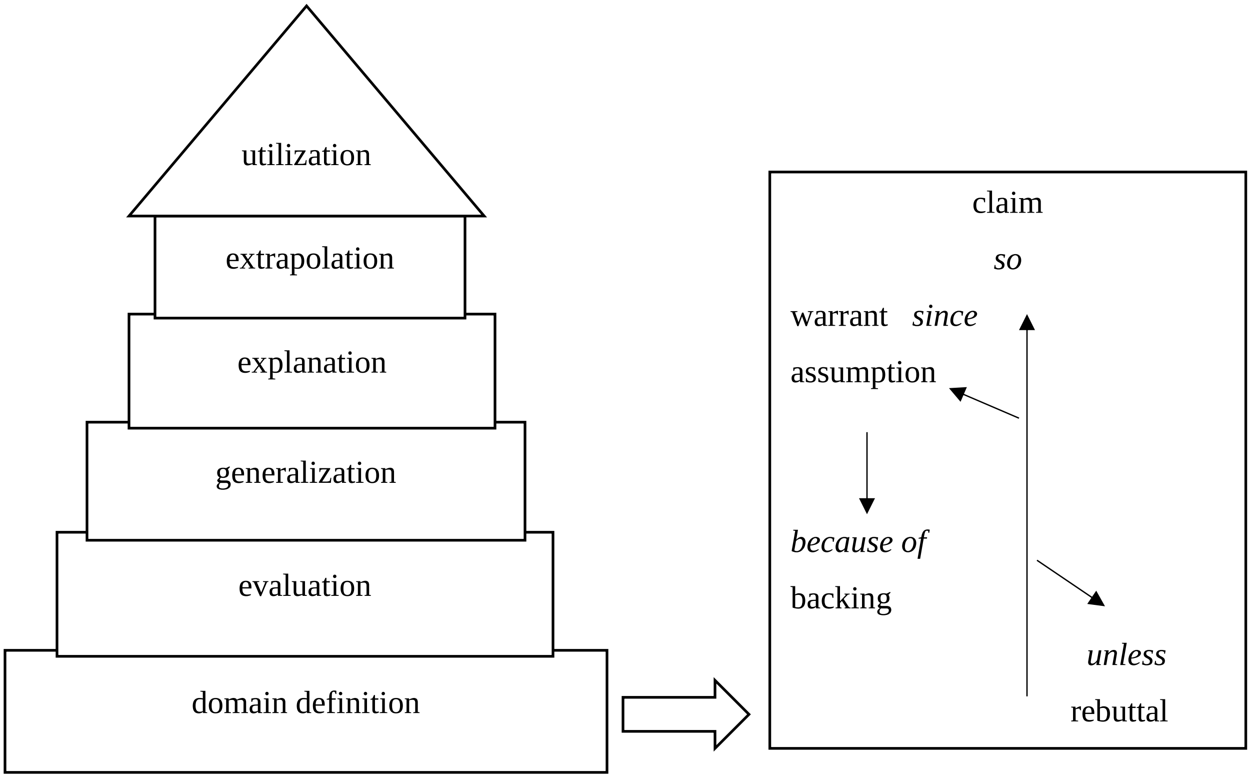


Figure 1. Types of inferences in an interpretive argument (left) and an inference structure (right).

warrant. From the example here, the backing is the fact that the warrant is based on teachers' experiences and research evidence. In sum, the bottom-up process of supporting the inference looks as follows: the backing provides support for the warrant, the warrant provides support for the inference, and the inference makes the connection between the grounds and the claim. On the other hand, the inference can be weakened by a *rebuttal* (Figure 1). The rebuttal provides additional facts to suggest that the inference might not be completely accurate. For the current example, a rebuttal can be that the observed incorrect grammatical structures may be representative of some specific dialect. Such a rebuttal would delimit the strength of the inferential link and call for more investigation and evidence. In addition to inference parts such as grounds, warrants, backing, and rebuttals, Kane (2012, 2010) explains the notion of an *assumption* in the argument structure (Figure 1). In Kane's terms, the warrant is the "simple if-then rule" that rests on assumptions, and backing for the warrant provides evidence in support of those assumptions. From the example above, an assumption can be that the rubric to measure the correctness of grammatical structures was developed appropriately.

Going back to the definition of an interpretive argument, we can see that an interpretive argument can consist of multiple inferences. In addition, now, it is clear that each inference has a specific structure that resembles an argument structure with well-developed logical links connecting grounds, claims, warrants, assumptions, backings, and rebuttals. To build the interpretive argument, all the inferences are connected together; subsequently, each inference becomes the grounds for the next inference in the interpretive argument pyramid.

The important inferences for a test interpretive argument can be explicated through work done by Chapelle et al. (2008) on building TOEFL iBT interpretive argument. The researchers adapted Kane's framework to explain the exam's interpretations and uses. Six important inferences that lead to test score interpretations and uses were singled out (bottom-up): domain definition, evaluation, generalization, explanation, extrapolation, and utilization (Figure 1). As it was mentioned before, each inference is built on the previous inference using it as its grounds. For example, domain definition inference serves grounds for evaluation inference, evaluation for generalization, etc. Kane's argument-based approach resembles a pyramid of inferences: the initial inferences are used to form all other inferences. Thus, even though weaknesses at any stage undermine the argument, flaws in the bottom inferences entail more ramifications because they affect all subsequent inferences.

Strengths of Kane's validity framework. Unlike other approaches to validity, Kane's validity framework has distinct advantages. Argument-based validity provides a more pragmatic approach to validation meaning that the approach is based on practical considerations (Kane, 2012). Evaluating the validity by stating the proposed interpretations and uses of an assessment improves the accuracy of conclusions, the appropriateness of the score uses, and the quality of the data-collection (Kane, 1992). In addition, explicitly stated logical interpretive arguments provide guidelines for what evidence is needed for validation.

Another advantage of the argument-based approach is the fact that "the interpretive argument also provides a basis for identifying the most serious challenges to a proposed interpretation – challenges that expose weaknesses (e.g., hidden assumptions) in the interpretive argument" (Kane, 1992, p. 9). The process of challenging the interpretive argument encourages validity research that, in turn, can increase the chances of making improvements in testing procedures.

Xi (2010) emphasized that Kane's argument-based approach allows integration of fairness into the validity framework. Fairness can be embedded in validity through the rebuttal part of the interpretive argument. Xi suggested a number of rebuttals for each inference. The researcher pointed out that inclusion of fairness as part of the validity will have a positive effect because it shows how unfairness detected at the beginning of the argument "may accumulate force and eventually become salient through biased score-based decisions and inequitable consequences" (p. 163).

Inferences and assumptions in performance assessment. The interpretive argument for performance assessment consists of at least three types of required inferences: evaluation/scoring, generalization, and extrapolation (Kane, 2006; Kane et al., 1999; Chapelle, 2012). These core inferences are critical for the validity of the interpretation assigned to performance assessment; therefore, weakness of any of the links undermines the validity of the interpretive argument as a whole (Crooks, Kane & Cohen, 1996). Each inference is connected to the specific process of assessment. First, students' performance is scored (evaluation/scoring), next, the score is generalized to the universe score (generalization), and then the universe score is extrapolated to the target score (extrapolation) (Kane, 1999). These three inferences represent the building blocks of the interpretive argument that validate the meaning of a test score by warranting the plausibility and appropriateness of the assumptions underlying these inferences.

First, the evaluation inference (sometimes referred to as a scoring inference) links the observed performance and the observed score (Kane, 2006). This inference depends on the assumptions, for example, that the scoring rubric is appropriate and is accurately and consistently applied by raters. The plausibility of these assumptions can be supported by evidence from empirical research into the soundness of rating criteria and quality of rater performance. It is important to note that administration of assessment tasks was separated from the evaluation inference in Crooks et al. (1996) but added to the evaluation inference in Kane et al. (1999).

Inclusion of administration to the scoring inference adds other important assumptions such as motivation of the students to do their best, exclusion of the equipment malfunction, and absence of inappropriate help from external sources or test administrators; however, the current study is not addressing these assumptions.

Second, the evaluation inference is followed by the generalization inference, which connects the observed score and the universe score, in other words, it provides evidence that a similar score can be obtained across administrations (Kane, 2006). The generalization inference relies on two basic assumptions: The sample is large enough to minimize the sampling error, and that the sample is representative of the population. Evidence in support of these assumptions is collected by psychometricians testing reliability and generalizability across samples of parallel tasks, forms, occasions, and raters. Substantial variability associated with any of the above would mean that the scores cannot be generalized beyond the specific set of tasks, raters, etc.

Third, the generalization inference is linked with the extrapolation inference, which joins the universe score and the level of skill that can be observed in the target domain (a real-life situation) (Kane, 2006). The main assumptions here would be that the test tasks are developed based on the tasks typically performed in the target domain, and that there are no skill-irrelevant (construct-irrelevant) sources of score variability that can undermine score interpretation in real life. These assumptions can be supported by investigations into how test scores correlate with sample performances from the target domain. In addition, it must be shown that the tasks cover most of the knowledge and skills needed for real-life situations.

Ultimately, the validity argument for performance assessment is constructed through the process of evaluation of the discussed core inferences that are part of the interpretive argument. Arguing the validity rather than stating it keeps it open to be questioned by critics. The interpretive argument can be evaluated using three criteria: clarity of the argument, the coherence of the argument, and plausibility of the inferences and assumptions (Kane, 2012). The current study challenges the validity argument for speaking performance tests by questioning the plausibility of the assumptions for the evaluation inference. The next section will describe how rater variability, which can be a source of construct-irrelevant variance, is relevant within Kane's validity framework.

Rater Variation as a Validity Threat

The present study investigates variability in raters scoring L2 speaking assessments. Rater variation can undermine the strength of the evaluation inference in the validity argument for speaking performance assessment. In Kane's validity framework, the evaluation inference is the building block of the generalization inference, and, subsequently, the extrapolation inference. Thus, flaws in the evaluation inference affect and weaken all the other inferences, and, as a result, the validity overall. Thus, it is important to study rater variability as understanding the reasons behind it can help us solve practical problems regarding test validity and improve rater training. In this section, first, the common validity threats for evaluation inference are indicated. Second, it is explained how rater variation stemming from (a) raters' cognitive processes, (b) raters' native/non-native status, and (c) raters' accent familiarity can be an example of a validity and fairness threat.

Common validity threats for evaluation inference. As discussed in the previous section, evaluation inference rests on several assumptions. To build the validity argument, the appropriateness of these assumptions is questioned. Crooks, Kane and Cohen (1996) outlined the following possible threats to validity from the perspective of the evaluation inference: (1) Scoring rubrics do not capture important qualities; (2) Raters show undue emphasis on specific rubric criteria, or unfairly favor particular response forms or styles; (3) There is insufficient intra-rater/inter-rater consistency; (4) Scoring may be too analytic; (5) Scoring can be too holistic. The presence of rater variability is pertinent to all except the first threat identified by Crooks et al. (1996). Xi (2010) mentioned an additional validity threat that is "rater bias against certain groups associated with the scoring of the Writing or Speaking sections" (p. 162).

Based on these common threats to evaluation inference, the examination of the literature on rater variability has shown that most of the language testing specialists examined rater effects in five main areas: (1) raters' rating experience, (2) amount and presence of rater training, (3) rater general background, (4) rater language background, and (5) rater types based on their approach to rating. To narrow down the focus, the current study is interested in uncovering rater variation that can happen because of differences in cognitive processes, rater's status (native or non-native speakers of English), and raters' level of accent familiarity. The following sub-sections provide further information about each source of rater variation starting with rater cognition, continuing with raters' language background and accent familiarity.

Rater cognition. Bejar (2012) and Wood (2014) stated that research into rater decision-making is a valuable source of information on potential threats to score validity. According to Bejar, it is important to understand "the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (p. 2). Research in rater cognition investigates the mental processes of decision-making that raters go through in order to better understand how raters apply a rating scale to arrive at a score. Research findings help reduce rater effects on test scores and inform rater training. Not many rater cognition studies address the topic of rater variability, and most of their foci are very specific to writing assessment. Thus, most of the findings cannot be applied to the assessment of speaking, but some overlap can be found.

The studies on rater cognition vary in their purposes. Some of the studies focus on the differences between experienced and novice raters, for example, Eckes (2008, 2012) tried to categorize raters into rater types based on their rating experience and amount of focus on a specific rating criterion. Other studies used rater decision-making differences in order to develop and validate writing rubrics (Cumming et al., 2002). This fact is important because the focus of the current study is the way raters use a well-developed rubric.

Rater cognition researchers claim that rater variability that lies in the decision-making process can persist even if raters receive training, necessary retraining, recalibrating, and monitoring, (e.g., Hoyt & Kerns, 1999; Weir, 2005; Wolfe, 2006). The cause of such unexplained variability that resists rater training can potentially be the individual differences in the decision-making processes and amount of focus on a particular feature outlined by the rubric (Wolfe, 2006).

Baker (2012) also supposed that individual differences in style of decision-making processes (e.g., avoidant, rational, intuitive) can explain some of the variability. Similar styles can occur in speaking performance, for example, avoidant raters may re-listen to test-takers' recordings multiple times in an attempt to avoid the final decision. Likewise, intuitive raters can make a rating decision momentarily after only generally listening to examinees' recordings; such raters can arrive at a decision intuitively, but still based on the rubric. The latter example matches the validity threat of "too holistic scoring" suggested by Kane and Cohen (1996). In addition, there can be rational raters who might be more focused on some specific criterion while listening such as vocabulary or pronunciation. Such raters may read and re-read the rubric using a strict rationale, which is an example of “too analytic scoring” also mentioned on the list of Kane and Cohen’s (1996) validity threats.

Researchers in writing have pinpointed that raters can be affected by their own essay reading style (e.g., Barkaoui, 2010; Crisp, 2008; Cumming, 1990; Cumming, Kantor, & Powers, 2002; Huot, 1993; Lumley, 2005). The essay reading style might appear to be a writing-specific difference, but it can probably occur in speaking assessment in the form of a listening style. Regardless of the similar rater training, raters in Vaughan’s study (1991) showed different reading styles: ‘first-impression-dominates,’ ‘single focus approach,’ ‘two-category,’ ‘laughing rater,’ and ‘grammar-oriented style.’ The author concluded that the individual approach by each rater contributed to score variation; differences in rater score showed that some raters failed and other raters passed the same essay. In addition to the ‘single focus’ and ‘two-category focus’ essay raters, research in speaking (Brown, 2000; May, 2006; Orr, 2002) described the same tendency of raters to attend to a set of salient response features, which were probably aligned with their perceived criteria importance. Similarly, speaking raters might make their decisions based on their overall first impression; they can also be criterion-oriented and make their decisions based on one dominant criterion such as grammar.

Eckes (2008, 2012) classified raters into groups based on criteria weight. He identified that some raters paid more attention to syntax, correctness, structure, and fluency whereas others paid significantly less attention (less criteria weight) to fluency and argumentation. Eckes classified raters into fluency and argumentation types when they had more criteria weight and into nonfluency and nonargumentation types when they had less criteria weight. Based on these results, Eckes (2009b) conducted another study where he analyzed self-reported criteria importance, criteria application easiness, and rater confidence by experienced scorers of L2 speaking performance. The results showed that the raters differed in weight attribution to scoring criteria showing “dimensional rater heterogeneity” (Eckes, 2009b, p. 51). In addition, the prevalent most important criterion was content of responses which was defined as meaningfulness of a response. Eckes argues that such findings provide grounds for his rater type hypothesis that raters can be classified into groups based on their rubric criteria interpretations. Additional rater variation can stem from raters’ personal preferences. For example, raters in Barkaoui (2010) were affected by the length of the written response and writer’s personal situation, which are non-rubric criteria. Likewise, speaking raters may have their own personal interactions with examinee responses.

The reviewed studies revealed that the raters tended to have different interpretations of the rubric because they assigned different importance to different rubric criteria, paid more or less attention to different aspects of examinees’ performance, or referenced non-rubric criteria. If raters score the same performance in different ways, it is necessary to look at the possible causes of such variation in order to inform rater training.

Raters might have different cognitive processes during rating: re-listening, taking notes, strategies of referring to the rubric, focusing on different parts of the performance, and dealing with unexpected responses or alternative ideas. For example, it can be possible that some raters would try to infer and interpret student’s response while others would not. These differences in raters’ approaches and decisions may affect raters’ final decision. For example, if one rater listens to a students’ recording more times, it is possible that they will understand the logic of the response better. In addition, differences in raters’ style of making decisions may affect the way raters use the rubric: more holistically or more analytically, which again leads to incomparability and threatens the validity.

Overall, research into cognitive rater types has been done more in writing assessment (e.g., Barkaoui, 2007, 2010; Cumming, 1990; Eckes, 2005, 2008, 2012; Vaughan, 1991) and less in speaking assessment (e.g., Ang-Aw & Goh, 2011; Orr, 2002; Chalhoub-Deville, 2005). Various rater differences were described through the lens of variation in rater cognitive processes, such as dissimilarities in rubric interpretation or different rating styles. Research on rater cognitive processes while rating speaking performance has been limited.

Native and non-native raters. Research has shown that some potential differences may occur between native and non-native rater groups due to the differences in construct interpretation. If the construct interpretation is not the same, then raters' foci during scoring would differ; therefore, raters' rubric utilization will not be comparable, which could result in a threat to validity. According to Johnson and Lim (2009), some researchers argue that native speakers should remain the ideal candidates to be raters and consider non-native speakers unacceptable. On the contrary, other researchers argue are non-native raters can be more suitable than native speakers (Hill, 1996).

We need to consider the inclusion of non-native speakers due to the process of globalization that has increased the use of English as an international language or lingua franca. Thus, language assessment has been affected as well because it has to match the use of English in real life. Academic English tests are becoming more international, for example, TOEFL iBT and IELTS results are accepted to fulfill English language proficiency requirements for English-medium universities in many different countries. If that is the case, then it is necessary to move towards assessing English in a more comprehensive way that would be more suitable for different educational contexts (Canagarajah, 2006). Including raters who are proficient non-native speakers of English is one way to do it. A study by Gu and So (2015) demonstrated that language teachers and language testers see this inclusion of non-native speakers as positive. Currently, non-native speakers can be employed by high-stakes testing companies if they reach near-native English proficiency. In addition, Xi and Mollaun (2011) stated that non-native speakers are often used for rating large-scale speaking tests. However, studies on rater variation argue that such inclusion may pose a threat to test validity because it is possible that native and non-native raters have different interpretations of testing constructs.

Language testing researchers have suggested that raters from mixed L1 backgrounds tend to use rubric criteria differently than native speakers (Brown, 1995; Shi, 2001). Additionally, non-native speakers can apply different standards that might not be comparable with those of native speakers (Zhang & Elder, 2011). Based on these claims, it is possible that the differences between native and non-native raters can add some construct-irrelevant variance into scoring. Studies addressing the differences of native versus non-native speakers showed mixed results in both writing and speaking assessment. Some studies reported no difference between native and non-native speakers rating L2 writing (e.g., Johnson & Lim, 2009; Shi, 2001) and L2 speaking (e.g., Kim, 2009; Xi & Mollaun, 2009); whereas other studies found differences (e.g., Chalhoub-Deville, 1995; Hill, 1997; Fayer & Krasinski, 1987; Zhang & Elder, 2011). The reasons for these differences remain obscure, but such inconsistent research results can come from the differences in scopes and designs; therefore, it is hard to compare the results that the studies yielded.

Studies in this area compared native and non-native speakers in a variety of ways, for example, teachers to naïve lay people (e.g., Chalhoub-Deville, 1995) and teachers to occupational professionals (e.g., tour guides in Brown, 1995). The studies discussed below include examinees and raters who shared and did not share an L1 and raters who were familiar or unfamiliar with test-takers’ L1 interlanguage phonology. It is important to note that it was not among the purposes of the studies to investigate how L1 match or familiarity factors relate to scoring. These studies focused only on whether raters are native speakers or non-native speakers of the tested language without exploring L1 match or familiarity effects.

One line of research has focused on whether there is a shared perception of speaking or writing proficiency among native and non-native raters. Zhang and Elder (2011) investigated the differences between native and non-native raters assessing speaking performance. The researchers concluded that the differences, analyzed in FACETS, on unguided holistic rubrics between the two groups of raters were marginal and insignificant. However, they found some differences in the decision-making process demonstrated by the written justifications of scores in specific categories (e.g., language use). The quantified qualitative differences between native and non-native groups were significant on five out of seven categories indicating the differences in interpreting the oral proficiency construct. These results align with Shi (2001) who analyzed rater behavior when grading written performance, where the differences lay not in holistic judgments but rankings of written score justifications. Other similarities between these two studies are that both of them used teachers as raters who did not have any training on the use of the rating scale. However, Zhang and Elder (2011) used MFRM, and Shi (2001) used MANOVA to compare scores assigned by the two groups.

Brown (1995) and Kim (2009) are two more studies where the differences in speaking scores assigned by native and non-native raters were analyzed using MFRM. These studies also used teachers as raters; however, instead of an unguided scale, a scale with descriptive bands was used. Similarly to the previous two studies, Kim’s raters were not explicitly trained since there was only a meeting to explain the research project and rating procedure, but raters in Brown’s study were provided with a one-day rater training. The studies also yielded similar results as raters did not differ in the overall scores assigned, but differed significantly in scores awarded to specific criteria. Unlike Zhang and Elder (2011), Shi (2001) and Kim (2009), in Brown's study, these specific criteria were not qualitatively generated by raters but were provided by the scale descriptors. Lastly, Brown (2005) and Kim (2009) did not find any differences in the consistency and severity between the two groups of raters.

Based on the results of these studies, the authors suggested that holistic rubrics do not distinguish the subtle variances in rating patterns, in other words, they mask the differences between native and non-native speakers in terms of specific traits. It can be seen that further qualitative analysis in these studies sometimes indicated that native and non-native speakers arrived at their holistic ratings by relying on different criteria. The differences described in these four studies are described below in more detail.

First, there were some differences in the way native and non-native raters paid attention to rating criteria. Brown (1995) stated that politeness and pronunciation were rated more harshly by non-native speakers than natives. Zhang and Elder (2011) demonstrated that both non-native and native speakers mostly focused on linguistic resources, content, and fluency. In terms of these three major criteria, there was only one difference – non-native speakers paid more attention to linguistic resources, whereas all other differences lay in less mentioned categories, namely interaction, demeanor, compensation strategy, and other general comments.

In addition, Shi (2001) reported that non-native speakers produced more negative qualitative comments. On the other hand, native and non-native speakers mostly focused on similar characteristics such as content ideas, content argument, general organization, and language intelligibility. The difference here was only in the assignment of different rankings of importance (number one, two, or three) for these major categories of concern. The study did not discuss the overall category focus differences between native and non-native teachers despite its ranking number. It can be inferred from the graphs that non-native speakers made more comments on the general organization than native raters, and native raters gave more comments on language intelligibility than non-native raters. The five most frequent categories for both groups included overall language use, pronunciation, and vocabulary that differ only in the rank-order. Furthermore, Kim (2009) stated that native speakers focused mostly on overall language use, pronunciation, vocabulary, fluency, and on specific grammar use, whereas non-native speakers drew most frequently on pronunciation, vocabulary, intelligibility, overall language use, and coherence. The other two most frequent characteristics for native speakers were fluency and specific grammar use, but intelligibility and coherence for non-native speakers (Kim, 2009).

Even though the analytic criteria or qualitative comments show differences between native and non-native speakers, the differences in major categories are not radically different. These trends demonstrate that the two groups may or may not have a common interpretation of the speaking proficiency construct when the rating is done without a well-established rubric with explicit criteria descriptors. It is also possible that raters' language background may not matter when native and non-native speakers are used as raters if holistic rubrics are used; however, differences might emerge when using analytic rubrics.

Overall, the studies comparing native and non-native raters have shown some differences in criteria focus between native and non-native raters; however, the studies on rater cognition also showed such differences within the native rater groups. In either case, it is important to investigate rater differences in order to adjust rater training and make raters more comparable to avoid validity threats. There has been a limited number of research studies that looked at the decision-making differences of native and non-native raters with ESL/EFL teaching experience. Most of the studies looking at such differences have been done in writing or speech perception. Therefore, there is a need for a study grounded in speaking assessment practices to address these issues.

Accent familiarity. Accent familiarity, which can be exhibited by either native or non-native raters, can also obscure the interpretation of test scores. Speakers who are more familiar with the accent of examinees might subconsciously give higher scores that can result in positive bias. On the other hand, raters whose accent familiarity level is low might assign lower scores that can lead to negative bias. Scores given by familiar/unfamiliar raters cannot be interpreted in the same way because they are contaminated by accent familiarity that is not part of the target domain description. In terms of positive bias, raters can assign more lenient scores to test-takers whose L1 is the same as the L1 or the L2 of the rater evaluating their speaking. Raters can be more lenient because they comprehend more than other raters due to their greater familiarity with the examinee's L1 and their own accent that they have when speaking English. At the same time, a negative bias can occur if raters are aware that they might be too lenient on such test-takers; therefore, they are consciously trying to be stricter to overcome the possible bias, which in its turn might result in unfairly lower scores. In addition, negative bias can occur when non-native raters penalize test-takers who have the same L1 as them when rating difficult but important aspects of language (e.g., politeness in Japanese described by Brown, 1995). Brown (1995) hypothesized that this could have happened because non-native raters have gone through the complex learning process of this feature themselves and are less tolerant of mistakes.

In this literature review, accent familiarity is explored only from the speaking assessment perspective even though accent familiarity has been vastly explored in the field of speech perception. Speech perception articles are not taken into account due to the fact that the results of those studies cannot be directly generalized to the field of speaking assessment since they did not have crucial parts of research performed in assessment setting, namely rater training, a detailed rubric, and sufficient time for decision-making (Winke et al., 2013).

Even though this dissertation is concerned only with the way accent familiarity was investigated in the field of speaking assessment, there are several approaches to accent familiarity operationalization. Specifically, accent familiarity was operationalized as (a) a broad concept of accent/interlanguage familiarity (Carey, Mannell & Dunn, 2011; Chalhoub-Deville, 1995; Huang, 2013), (b) L1 match (Xi & Mollaun; 2011; Wei & Llosa, 2015), or (c) L2 match (Winke, Gass, & Myford; 2013; Winke & Gass, 2013). Before discussing these studies in more details, it is important to keep these operationalization differences in mind.

First, the results of the two studies that explored familiarity of native speakers due to the L2 match (Winke et al., 2013) or general familiarity exposure (Carey et al., 2010) are reviewed. Winke et al. (2013) addressed the overall grades using a holistic TOEFL independent rubric. They used the MFRM approach to uncover potential biases in the rating process. The results revealed that a greater familiarity of raters (operationalized as the same L1 as speaker's L2) results in more lenient scores toward the familiar accent and harsher towards speakers with other accents. For example, raters who studied Spanish as their L2 were more lenient towards Spanish L1 test-takers and raters who studied Chinese as their L2 were more lenient towards Chinese L1 examinees (Winke et al., 2013). However, the magnitude of the effect was low and did not have a large impact on test-takers' scores.

Furthermore, Carey, Mannell, and Dunn (2010) is another research study that can be treated as investigating native speakers' familiarity since most of the raters were born in the countries of the inner circle (Kachru, 1985) except for the Indian raters who described English as their L2. The researchers looked at the pronunciation scores assigned by familiar and unfamiliar raters. Familiarity was operationalized as raters' L1, the experience of teaching students with a particular L1, residence in the country where people speak that language, or any other non-native accents they are familiar with. The variable was coded as dichotomous (familiar/unfamiliar) where unfamiliar meant no prolonged exposure. The results revealed that test-takers were awarded higher pronunciation grades when raters were familiar with their interlanguage pronunciation, and lower when unfamiliar. A higher score of 6 was more likely to be awarded by familiar raters, whereas a lower score of 4 was likely to be given by unfamiliar ones. The results of this study should be interpreted with caution because only pronunciation was rated and not general speaking ability. In addition, the number of speech samples was extremely limited (one OPI for each language group).

The results of Huang (2013) were not consistent with the previous studies. Her familiarity of native speakers was defined as taking Chinese classes and presence of interactions with non-native speakers. The results did not show any differences in raters' analytic rubric scores either due to accent familiarity or the presence of ESL/EFL teaching experience. However, the mean length of teaching was 3 years, and the study had an unguided rating context. In addition, the raters' native- or non-native-speaker status was not discussed. Thus, the results of the study should be interpreted with caution.

On the other hand, the qualitative data in Huang (2013) and Winke and Gass (2013) showed similar patterns in terms of showing that accent familiarity increased raters' comprehension that can be called a positive bias. In Huang (2013), raters provided qualitative data about their beliefs of the effects of familiarity on their ratings where most of the raters believed that accent familiarity enhanced their error detection and comprehension. Similarly, in Winke and Gass (2013), 15 out of 26 raters mentioned test-takers accents and overall made 29 comments on this topic; the comments were coded as positive (9 comments), negative (18 comments), and neutral (2 comments). The raters who made positive comments were aware of their bias and were afraid to be more lenient towards familiar accents. For example, one heritage speaker noted, "it's been my job for the past 18 years to fill in, to fill it in to make it sound more English-sounding so my mind already knows how to do that" (Winke & Gass, 2012, p.776). Moreover, the same trend of accent familiarity enhancing comprehension can be found in a study by Wei and Llosa (2015, p. 298):

> Yes, I did get what he said, but I am a native speaker of the accent he is. If I am to put myself in the shoes of somebody who is not, they would probably have had a hard time understanding.

The possibility of negative bias resulting from accent unfamiliarity was addressed in Winke and Gass (2012) that described raters who made negative comments showing that some accents were very strong and hard to understand. One rater noted that one speaker sounded as if she understood the topic and knew what to say, but the rater described that the test-taker's accent as "terrible". Such a strong reaction can be explained by the fact that most of the raters were undergraduate students without teaching experience. Such a strong description might be caused by the rater's unfamiliarity with the test-taker's accent. This type of negative reaction that results in lower comprehension of unfamiliar accents was also discussed in Scales, Wennerstrom, Richard, and Wu (2006), where English learners provided their opinions toward non-native accents. Even though this study shows a similar trend, Scales et al., (2006) focused on inclusion of foreign-accented speech to listening tests and was not about rating speaking.

Another study by Xi and Mollaun (2011) looked at accent familiarity also operationalized as L1 match and looked at the way familiarity can affect raters' grades. The researchers analyzed two groups of Indian raters: one with the usual TOEFL training with multilingual benchmark recordings and calibration samples, and the second with specific training with benchmark recordings and calibration samples from Indian speakers. The results showed no difference between the two groups; therefore, the researchers claimed that the usual ETS rater training procedures and rater certification offer enough practice to mitigate the possible rater L1-match effect. Even though both groups showed good interrater reliability, the researchers pointed out that the special training group showed slightly higher reliability. Moreover, Wei and Llosa (2015) also focused on Indian raters. Their results also yielded no quantitative group differences in the use of scoring criteria, attitudes, internal consistency, or severity of scores. However, as shown by qualitative data, Indian raters were better at identifying and understanding specificities of Indian language in vocabulary, syntactic structure, rhetorical organization, cohesive devices, and aspects of culture and pragmatics.

It can be seen from the literature review that accent familiarity or unfamiliarity can have mixed effects on raters' scores. Some studies (Winke et al., 2011; Xi & Mollaun, 2009) focused more on rater familiarity as a source of rater bias, other studies (Carey et al., 2011; Huang, 2013) attempted to address both familiarity and unfamiliarity effects at the same time. The results of the studies discussed are mixed, and no clear generalizability can be drawn from them.

Present Study

The reviewed literature has shown that performance assessment is susceptible to subjectivity that comes from rater variation. Rater differences can resist rater training and cause the variety of rater severity and add to the pool of unexplained error. Based on the argument-based approach to validity (Kane, 206), rater variation can affect the validity of a speaking performance test at the level of evaluation and, subsequently, the presence of rater variance can impact the generalization and extrapolation inferences resulting in weak validity argument. In other words, if raters are not comparable, then the results cannot be generalized beyond that set of raters to the universe of raters. Additionally, if the scores are contaminated by construct-irrelevant rater variation, then fair and objective extrapolations to the target domain situations cannot be drawn.

Rater variability can have a negative or positive effect on test-takers' scores. In order to make exam ratings fair and raters interchangeable, the decision-making processes and rater language background effects should be explored and adjusted. The reviewed literature outlined that, first, the effects of raters' decision-making processes have been studied in the field of writing assessment; however, there is a lack of such research in the area of speaking assessment. Second, the debate of native and non-native raters' suitability for scoring speaking assessment is still questionable. Third, the effects of rater accent familiarity have been mixed. Thus, the present study will aim at capturing possible rater effects in speaking performance assessment because of: (1) rater status (native/non-native), (2) rater cognitive processes, and (3) rater accent familiarity with examinee accents.

The current study focused on investigating the rating behavior of native and non-native raters in order to uncover and classify the differences in decision-making patterns when rating speaking performance by multilingual test-takers. A group of examinees with whom raters share the L1 was also included in order to examine another potential source of rater variability, which is L1 match of raters' and examinees. Rater's familiarity with other examinees' L1s was also taken into account and aimed to discern the potential presence of familiarity effects when raters are familiar with examinees' L1.

The results of the dissertation will help to reveal any flaws that are present in the evaluation inference of the validity argument. Since raters should be aware of their own specific patterns because these considerations of rater cognition should contribute to a better understanding of raters' various needs for rater training (Kim, 2015), the next step will be to utilize these results to propose a set of specific rater training guidelines to raise rater trainers' and raters' awareness of these potential sources of rater variability

Research Questions

Research questions for mixed methods studies commonly include three types: quantitative questions, qualitative questions, and a question that utilizes both quantitative and qualitative data (Creswell, 2013; Creswell, 2015; Creswell & Zhou, 2016; Plano Clark & Badiee, 2010; Guetterman & Salamoura, 2016). The present study had one research question using quantitative analyses, one research question drawing on qualitative analyses, and one research question that used both types of data. For the purposes of this study, the following research questions were posed:

RQ1: What are the differences between native and non-native rater groups in terms of their scoring patterns and comments that they provided on test-takers' performance?

a. To what extent do NS and NNS raters differ in terms of consistency and severity of their analytic scores?

b. Do NS and NNS raters show evidence of differential rater functioning related to rubric sub-criteria and examinee L1?

c. To what extent do NS and NNS raters differ in terms of scoring examinees by L1?

d. To what extent do NS and NNS raters differ in terms of the reported accent familiarity?

e. Is there a relationship between raters' familiarity, severity, and consistency?

f. To what extent do NS and NNS groups of raters differ in terms of the number and direction of their comments?

RQ2: What scoring strategies do NS and NNS raters use while grading L2 speaking performance?

RQ3: How do quantitative and qualitative findings complement each other?

Chapter 3: Method

Research design

The study used a mixed methods research design (Figure 2) that "combines qualitative and quantitative approaches into the research methodology of a single or multi-phased study" (Tashakkori & Teddlie, 1998, pp. 17-18). There is no single mixed methods design that can best describe the present study. Depending on what classification or approach is taken to classify this mixed methods design, it can be a concurrent, conversion, and sequential mixed methods design (Tashakkori & Teddlie, 2006), developmental and complimentary (Greene, Carcelli, Graham, 1989) or convergent parallel (Creswell, 2014).

The concurrent mixed methods design was utilized during the first part of the data collection when the raters provided quantitative scores for each recording simultaneously with qualitative comments. The term *concurrent* refers to the timing characteristic of such designs described as “one method implemented within the time frame spanned by the implementation of the other” (Greene et al., 1989, p. 263). In addition, due to the fact that the qualitative comments were subsequently re-coded into numeric categories, this mixed methods design was also a conversion design. Moreover, the sequential mixed methods design was utilized during the second stage of data collection, when the raters and the recordings for the qualitative think-aloud and interview part were selected based on the analyses of the quantitative data. Also, based on the classification by Greene et al. (1989), this mixed methods study can also be considered developmental because the analyses of the quantitative data determined data for the subsequent qualitative data collection. On the other hand, from the statistical analyses perspective and not from the data collection one, this mixed methods design can be called complementary. In *complimentary* mixed methods studies, "qualitative and quantitative methods are used to measure overlapping but also different facets of a phenomenon, yielding an enriched, elaborated understanding of that phenomenon" (Greene et al., 1989, p. 258). This type of mixed methods design is also called *convergent parallel* mixed methods design by Creswell (2014):

> Convergent parallel mixed methods is a form of mixed methods design in which the researcher converges or merges quantitative and qualitative data in order to provide a comprehensive analysis of the research problem. In this design, the investigator typically collects both forms of data at roughly the same time and then integrates the information in the interpretation of the overall results. Contradictions or incongruent findings are explained or further probed in this design. (p. 44)

Step 1: Raters provided quantitative scores and qualitative comments.
Step 2: Quantitative scores were analyzed to make statistically driven selections of examinee recordings and raters for further qualitative investigation.
Step 3: Qualitative think-aloud protocols and interviews with raters were held.
Step 4: Quantitative and qualitative findings were presented separately.
Step 5: Quantitative and qualitative findings were integrated.

Figure 2. Overview of the mixed methods research design.

To make the nature of the mixed methods design used in this study clear, it was decided to approach it from the data analyses and not from the data collection perspective. Thus, the current study calls its mixed methods design a complementary mixed methods design (Greene, Carcelli, Graham, 1989) since at the conceptualization stage, it was hypothesized that the qualitative part would confirm quantitative findings and enhance the breadth and depth of inferences proposed by the study as well as shed light on nuances that might have been masked by using only one type of data. This mixed methods design was also chosen because it allowed the researcher to address rater variability issues discussed in the literature based on the quantitative analysis, and subsequently perform a qualitative analysis to expand the knowledge on the possible reasons for rater variation. Quantitative and qualitative data were analyzed separately, then the results from both types of data were discussed together to increase interpretability and meaningfulness of the findings. According to Creswell (2013, pp. 51-52), a typical question for a complimentary mixed methods design can be "To what extent do the quantitative and qualitative results converge?" or "How do the qualitative findings provide an enhanced understanding of the quantitative results?" (Creswell, 2013, pp. 51-52). The final discussion was based on integration and corroboration of quantitative and qualitative parts.

Pilot Study

Before the current dissertation study was conducted, a pilot study was carried out to test out instruments, rater training procedures, and methods of analysis. The pilot study analyzed the scoring behavior of two groups of raters American (*n* = 5) and Russian (*n* = 5). All the raters filled out a background questionnaire asking about their teaching experiences, language learning history, the background of students in their classrooms, and their exposure to and familiarity with the non-native accents used in the study. Each rater received individual training and scored the same 12 recordings in response to an independent speaking task. The speech samples included Arabic (*n* = 4), Chinese (*n* = 4), and Russian (*n* = 4) speakers, which were pre-scored by two raters and the researcher to determine score variability. Since the TOEFL iBT independent speaking rubric was used analytically, each recording received four scores (i.e., Overall, Delivery, Language Use, and Topic Development) by 10 raters. Raters' scores were examined using the Multi-Faceted Rasch Measurement using FACETS (Linacre, 2014) software to test group differences in terms of consistency and severity. The qualitative analyses involved thematical coding of transcribed think-aloud sessions and interview sessions using content analysis (Strauss & Corbin, 1998) to investigate the cognitive processes of raters and their perceptions of their rating processes. The coding included such themes as decision-making and re-listening patterns, perceived severity, criteria importance, and non-rubric criteria (e.g., accent familiarity, L1 match). Afterward, the quantitative and qualitative results were analyzed together to describe potential sources of variability. This analysis was done employing side-by-side comparison of qualitative and quantitative data (Onwuegbuzie & Teddlie, 2003).

The pilot study results revealed that there were no significant differences between native and non-native speakers with non-native speakers exhibiting 0.30 logits more severe scores. In addition, all raters, regardless of the group, demonstrated several patterns of rating depending on their focus while listening to examinees' performance and interpretations during the decision-making process.

Lessons learned from this pilot study include: (a) it is not practical to classify raters into types based on think-aloud protocols since it is complicated to measure how many times the raters meaningfully mention the rubric categories; therefore, collection of rater qualitative comments was suggested; (b) it is not practical to perform non-selective coding of think-aloud protocols and interviews; therefore, selective coding based on the themes from the pilot study was proposed; and (c) self-report familiarity is not enough to subdivide the raters into familiar and unfamiliar groups; therefore, it was advisable to collect additional familiarity data when the raters report their familiarity by listening to short speech excerpts without L1 identification.

Participants

The participants in the study included examinees and raters divided into groups based on their L1. The two rater groups were native speakers (NS), who speak North American English as their L1 and non-native speakers (NNS), whose mother tongue is Russian. The three examinee groups were Arabic, Chinese, and Russian, and they speak Arabic, Mandarin Chinese, and Russian as their L1s. In addition, the background of two coders who were involved in coding qualitative comments is described at the end of the Participants section.

Raters. The raters in the study were comprised of 23 experienced Russian EFL/ESL teachers as NNS raters and 23 experienced North American EFL/ESL teachers as NS raters (Table 1). There were more males in the NS group (10 males) than in the NNS group (4 males). The average age of both rater groups was 32 years old (*SD* = 8). The age of the NNS ranged from 21 to 42 (*M* = 30, *SD* = 5); the age of the NS group was spread out from 24 to 71 (*M* = 34, *SD* = 10). If the oldest participant in the NS group (71 years old) is considered an outlier, the age range would be from 24 to 45 (*M* = 32, *SD* = 6). In terms of education, all raters had an MA degree in teaching English as a second/foreign language or equivalent; three NS raters and three NNS raters were at the beginning of their Ph.D. studies.

Table 1. *Rater Demographic Information*

	NS	NNS
Number	23	23
Age	*M* = 34, *SD* = 10*	*M* = 30, *SD* = 5
Gender	10 males and 13 females	4 males and 19 females
Teaching experience	*M* = 8.55, *SD* = 6.57	*M* = 7.78, *SD* = 5.41

Note. **M* = 32, *SD* = 6 without the oldest participant (71 years old).

In terms of language background, all NS raters identified North American English as their L1 (with only one participant who described themselves as an English-Spanish bilingual naming English as their dominant language). The participants in the NNS rater group spoke Russian as their L1 with four participants who identified themselves as bilingual speakers (i.e., Russian-Ukrainian, Russian-Belorussian, Russian- Circassian, and Russian-Armenian). Concerning L2 studies, all NS raters reported at least one second language and there were only three NNSs who did not state that they had a ranging command of a third language. Most of the

NS raters put Spanish or French as their L2 as well as German, Italian, Portuguese, Vietnamese, Korean, Japanese, Chinese, Kyrgyz, Russian, Uzbek, Mongolian, Cape Verdean Creole, and Jamaican Creole. In addition to English, NNS raters reported various proficiency in German, French, Italian, Spanish, Romanian, Czech, Ukrainian, Belorussian, Turkish, Thai, Japanese, and Chinese.

Regarding teaching experience, on average, all raters together had eight years of teaching English as a second/foreign language (*SD* = 5). The minimum number of years teaching was two and the maximum was 30 for NS (*M* = 8.55, *SD* = 6.57) and 23 for NNS (*M* = 7.78, *SD* = 5.41). Six NS raters taught only in the US; others had experience teaching in Mexico, Costa Rica, South Korea, Japan, Bhutan, Puerto Rico, Chile, Jamaica, Cabo Verde, Mozambique, Colombia, China, Peru, Vietnam, Myanmar, Kyrgyzstan, Georgia, Mongolia, Uzbekistan, and Spain. All the teachers in this group had students from four to more than 30 different countries all over the world. On the other hand, NNS teachers taught primarily in their home country, Russia, with only six who taught in China, USA, Malta, Ukraine, Belarus, Thailand, Germany, or Moldova. This group of teachers taught mostly Russian students; only 7 mentioned having had non-Russian-speaking students in their classrooms.

The target population of raters was not highly experienced trained raters, but rather highly educated classroom practitioners who assess their students in the classroom on a daily basis and who are skilled enough to be potentially hired as high-stakes exam raters. The raters were recruited in the US and Russia. The recruitment advertisement was posted on various Applied Linguistics and TESOL Facebook communities, myTESOL Lounge, and forwarded to various university and IEP listservs in Russian and the US. The recruitment email outlined the requirements that raters had to meet in order to participate, namely an MA degree in TESOL or an equivalent area and at least 2 years of EFL/ESL teaching experience. This amount of teaching experience ensured that the teachers were skilled and would have been performing formal and informal language assessment as part of their daily teaching routine. Raters' background information was obtained using the background questionnaire (see Instruments).

The rationale for having two rater groups was to investigate the differences in scoring behavior between native and non-native raters. If either rater group had scored *only* those test-takers with whose L1 accents in English they were more familiar than the other rater group, the group differences might have been attributed to the differences in accent familiarity. Thus, it was decided to include Arabic and Chinese examinees (usually highly familiar to raters from North America and unfamiliar to Russian raters) and Russian examinees (highly familiar to Russian raters and more or less unfamiliar to North American raters). In addition, the inclusion of Russian L1 test-takers allowed the consideration of another factor – the effects of the L1 match between raters and examinees.

Examinees. This study used 99 speech recordings in response to two independent speaking prompts (see Instruments) from a semi-direct speaking test. The recordings included Chinese (*n* = 33), Arabic (*n* = 33) and Russian (*n* = 33) speakers (Table 2). In terms of gender, there were more male than female recordings (50 males and 25 females). There were 5 female and 20 male recordings for Arabic speakers, 9 females and 16 males for the Chinese group, and 11 female and 14 male speech samples in the Russian group.

The recordings from Arabic and Chinese L1 backgrounds were obtained from an archived database from an administration of a placement test at an IEP in the United States, while Russian L1 recordings were collected from an IEP located in Russia using the same administration process of the speaking task. The recordings were balanced in terms of test-taker proficiency (see Procedures). Out of 99 recordings, 75 were used for rating and 24 for rater training purposes. Each recording lasted between 43 and 60 seconds (*M* = 54.41, *SD* = 4.18). Table 1 describes the number and L1 background of the recordings that were used for training, calibration, and rating. It is important to note that due to the primary focus on raters, the examinees were used more as an instrument to trigger raters' responses rather than to investigate test-takers' language ability.

Table 2. *Total Number of Recordings by Each L1*

Procedure	Arabic	Chinese	Russian	Total
Training and Calibration	8	8	8	24
Individual Rating	25	25	25	75
Total per L1	33	33	33	99

Coders. This study used three coders for coding raters' qualitative comments. The coders consisted of a male undergraduate student majoring in English and two Ph.D. students specializing in assessment, a male and a female. The undergraduate student performed initial coding of the all the comments to see how the rough pilot draft of the coding scheme worked. The Ph.D. students coded the data after the coding scheme was revised and more reliable. They coded 30% of the data (15% each).

Instruments

The instruments for the study included two background questionnaires (one for raters and one for Russian examinees), an accent familiarity scale, two independent speaking prompts, and the TOEFL iBT independent speaking rubric.

Speaker background questionnaire. Due to the fact that the only background data that could be obtained for Arabic and Chinese recordings from the archive were their gender and native language, the same data were collected from the Russian test-takers. A questionnaire (see Appendix A) was used to collect general background data from Russian respondents in order to describe the sample. This questionnaire was presented to Russian students in a Russian translation.

Rater background questionnaire. A questionnaire (see Appendix B) was used to collect background information from all raters. The questionnaire was developed by the researcher for the purpose of the study utilizing parts of the Language Experience Questionnaire (Harding, 2012) and Rater Language Background Questionnaire (Wei & Llosa, 2015).

The questionnaire informed the researcher about the participants' general, academic, language learning, language teaching, previous rating background together with raters' level of familiarity with test-takers' L1s. The questionnaire consisted of five parts: (a) general demographic and academic information, (b) language background, (c) teaching experience, (d) rating experience, and (e) accent familiarity. This information helped to choose qualified participants and describe the rater pool.

The questionnaire was divided into two parts, one collected before raters' participation in the study and one after. The part enquiring about raters' experiences teaching students of Chinese, Arabic, and Russian L1 background and raters' familiarity with those accents was collected after the study in order not to cue raters that only these three L1 backgrounds were used in the study. The second part of the questionnaire used the familiarity scale (described below).

Accent familiarity scale. An accent familiarity scale was used to determine raters' perceived familiarity with English spoken by the L1 groups in the study (i.e., Arabic, Chinese, Russian) (see end of Appendix B). In the present dissertation study, raters' accent familiarity was elicited with and without L1 identification (see Procedures). The word *accent* was not used in the directions for using the scale in order not to focus raters' attention on the purely phonological meaning of the word accent. For example, the raters were prompted to provide their judgments by the following sentence: “To what extent are you familiar with non-native English speech similar to this one (e.g., peculiarities of grammar, vocabulary, pronunciation)?” (see Appendix B). The familiarity scale had 6-points with the following descriptors: No, Very Little, Little, Some, A Lot, Extensive. The present 6-point scale was tested during the pilot study with 10 raters and showed Cronbach's α reliability of .70.

In the dissertation study, the reliability coefficients for the scale with L1 identification (3 items per L1) was .67 with all L1s together (Arabic = .93, Chinese = .94, Russian = .93). The reliability coefficients for the scale without L1 identification (8 items per L1) was .93 with all L1s together (Arabic = .94, Chinese = .96, Russian = .88). In addition, a composite accent familiarity score was calculated for each rater (see Results). The reliability coefficients for the composite accent familiarity scale (11 items per L1) was .93 with all L1s together (Arabic = .95, Chinese = .97, Russian = .91).

Speaking prompts. Two speaking prompts were used to obtain speakers’ speech samples (see Appendix C). Task 1 was an opinion task asking an alternative question about how a person prefers to study for an exam (i.e., alone or in a group). Task 2 was another opinion task asking an alternative question about what size of classes is better for students (i.e., big or small). Both prompts were selected because independent opinion tasks are commonly utilized by testing companies for language proficiency tests. The current prompts were selected from the research database of the U.S.-based IEP from a pool of eight other independent prompts. The topics of the selected prompts seemed accessible as they did not require any specific knowledge from the respondents and relevant since they were related to academic life. Another reason to select the prompts was the practically issue because these tasks had the most examinee responses in the research database.

Rating rubric. The raters used the TOEFL iBT independent rubric in this study (see Appendix D). The rubric was chosen because it represents a common speaking rubric with four rating sections including General Description, Delivery, Language Use, and Topic Development. In addition, the validity of the test using this rubric was established by research (e.g., Chapelle, Enright, & Jamieson, 2008). The use of the established and validated rubric helped to solely concentrate on rater differences without the possibility of results being contaminated with irrelevant variance brought in by the questionable validity of the rubric. TOEFL iBT rubric was used as an analytic tool (scores for Delivery, Language Use, and Topic Development) in order to obtain more information about raters' focus on each sub-rating criteria; however, the raters also provided their holistic scores (scores for General Description) in the same manner as TOEFL iBT holistic scores are given. Both analytic and holistic scores were used for quantitative analyses.

Procedures

Data collection. This section provides a detailed description of the data collection process. First, this section talks about how examinee data were obtained, screened and organized. Next, the rater recruitment and screening is described followed by the details about (a) individual rater training; (b) individual ratings (quantitative); (c) individual think-aloud protocols and interviews (qualitative); and (d) accent familiarity background questionnaire.

Examinee data preparation. As described in the Participants section, the study used examinee recordings from three L1 backgrounds: Arabic, Chinese, and Russian. The following sub-sections describe examinee data preparation in more details.

Chinese and Arabic. Chinese and Arabic speakers' recordings were retrieved from a research database at a U.S.-based IEP where students learn English for academic purposes.

These samples were collected during two administrations of the IEP's placement test for incoming students of various proficiency. The administration from Fall 2012 provided students' responses to Task 2 (see Speaking prompts), the administration from Fall 2013 provided student's responses to Task 1. According to the grades assigned by IEP's raters, the students' proficiency varied from 0 (no attempt) to 4 (the best score) as evaluated using the TOEFL iBT independent speaking rubric (see Rating rubric) with most of the scores concentrated in the middle (scores 2 and 3).

Variability of Arabic and Chinese accents. It is possible that test-takers who identify their L1 as either Arabic or Chinese can come from different countries or different parts of the same country. In the current study, Chinese refers to the Modern Standard Chinese language, which is also known as Mandarin, Guoyu, or Putonghua, which is the official governmental and educational language of the People's Republic of China and Taiwan. In addition, Arabic refers to the official governmental and educational language of Saudi Arabia, Kuwait, and other countries. In the archive database, there was no available information that could help accurately describe the specific regional accents of the participants. Thus, to provide the readers with the information about regional accent variability, an attempt was made to describe the sample using the impressionistic judgments of native speakers of these languages. One native speaker of Mandarin from central China and one native speaker of Arabic from Saudi Arabia were asked to listen to the Chinese and Arabic recordings and describe the test-takers' accents. According to the information provided by the Arabic native speaker, most of the Arabic recordings were described as produced by speakers from the Gulf countries and two as potentially from Egypt. The native speaker of Mandarin described most of the recordings as Mandarin speakers mentioning two people who might speak Cantonese.

Russian. Recordings from Russian speakers were collected (using the same speaking tasks) from students recruited from a similar IEP where students learn English for academic purposes located in Russia. Speech samples from Russian students were collected under the same conditions as samples obtained from the U.S.-based IEP. The students had one minute to prepare and one minute to record their answer. In order to align scores from Russian speakers with the scores from Arabic and Chinese speakers, the researcher collected more speech samples than needed and subsequently filtered them in order to choose the same number of scores 1, 2, 3, and 4 as the other two L1 backgrounds had. General information that was available for the Arabic and Chinese students (gender and L1) was also collected for the Russian students using a student background questionnaire (see Instruments).

Selection process. The recordings from all L1 backgrounds varied in test-takers' language ability from low to high proficiency based on the holistic scores 1 to 4, which they were assigned by raters at the U.S.-based IEP. There were no examples for the score of 0 as 0 means the task was not attempted or the response was unrelated to the topic. To determine the compatibility of language proficiency for students across language groups, the holistic ratings were retrieved from the U.S.-based IEP research database for Chinese and Arabic recordings. These holistic scores were assigned by trained U.S.-based IEP's raters. To obtain similar holistic scores for Russian recordings, four experienced assessment specialists trained and employed at the same U.S.-based IEP were recruited. The researchers and the raters scored all the recordings; the mode of their scores was used to estimate the proficiency of the Russian test-takers. By the end of the selection process, the researcher had three groups of recordings: Arabic, Chinese, and Russian, which were collected at similar facilities, answered the same two prompts, holistically scored by the U.S.-based IEP's trained raters, and represented a range of language proficiency.

Noise, length, and content screening. All of the obtained recordings contained some background noise because of the fact that several speakers were recorded in the same room. This is not uncommon for language proficiency speaking tests, for example, TOEFL iBT is taken by multiple test-takers in the same room. All the speech samples were analyzed in terms of the amount of noise and only recordings with a level of background noise that did not severely impede comprehension were used. To determine which recordings had too much background noise that prevented speech comprehension, the recordings were listened to by the researcher and a layperson. The recordings marked as having excessive background noise were excluded.

The next step involved length screening. It was decided to exclude speech samples that were less than 43 seconds in order to have a more balanced sample in terms of recording length. The last step involved screening for content. In order to not reveal students' country of origin, any recordings that had such information were not included.

Benchmarking. After the recordings were collected and screened for noise, length and content, they underwent the process of benchmarking (see Appendix E) performed by the researcher in order to select 24 recordings (12 from each task) that have the best fit for each score band and come from each L1 group (Table 3). These selected recordings were randomly allocated to appear in rater training or calibration practice.

Table 3. *Recordings for Rater Training and Calibration Practice per Rater*

L1	Score of 1	Score of 2	Score of 3	Score of 4	Total per L1
			Task # 1		
Arabic	1	1	1	1	4
Chinese	1	1	1	1	4
Russian	1	1	1	1	4
			Task # 2		
Arabic	1	1	1	1	4
Chinese	1	1	1	1	4
Russian	1	1	1	1	4
Total per score	6	6	6	6	24

Recordings for individual ratings. As it was described in Table 2, 75 speech files were utilized for the individual ratings. The spread of the selected recordings resembled a normal distribution (Table 4) with most of the samples being from band 2 (*n* = 24) and 3 (*n* = 27), and fewer speech samples coming from band 1 (*n* = 12) and 4 (*n* = 12). Due to the combination of fully and partially crossed recordings for each rater (described in Procedure), there were 35 responses to Task 1 and 40 to Task 2.

Table 4. *Recordings for Individual Rating*

	Score of 1	Score of 2	Score of 3	Score of 4	Total per L1
Arabic	4	8	9	4	25
Chinese	4	8	9	4	25
Russian	4	8	9	4	25
Total per score	12	24	27	12	75

Note: Out of 75 recordings, 35 were for Task 1 and 40 for Task 2.

Recordings for reflective think-aloud. Another set of recordings was used for reflective think-aloud protocols. These recordings were chosen among those speech samples that were used for the individual rating part. After the raters had completed scoring, Facets analyses was conducted. The recordings to appear in this part of data collection were either selected because they were given a mixture 1, 2, 3, 4 scores in the quantitative part of the study (e.g., 1 for Delivery, 2 for Language Use, and 4 for Topic Development) or based on the Facets examinee report (highlighted by Facets as poorly predicted by the model). Four such recordings from each L1 background were selected (12 recordings in total). Table 5 shows the gender of the speakers and the initial holistic grade assigned before the study.

Rating Procedures. As described in the Participants section, the study had two groups of raters, NS (North American raters) and NNS (Russian raters). The following sub-sections provide further details about the rater recruitment process and training coupled with more information about quantitative and qualitative data collection.

Table 5. *Recordings for Reflective Rating with Think-Aloud*

L1	Score of 1	Score of 2	Score of 3	Score of 4	Total
			Task # 1		
Arabic		F	F	M	3
Chinese		F, M			2
Russian	M				1
			Task # 2		
Arabic		M			1
Chinese		M, M			2
Russian		M, M		F	3
Total per score	1	8	1	2	12

Note. M stands for male and F stands for female. Each letter denotes one speaker.

Recruitment and screening. The researcher recruited Russian and American raters. The IRB-approved recruitment email was posted on various Applied Linguistics and TESOL Facebook communities, myTESOL Lounge, and forwarded it to various university and IEP listservs in the US and Russia. Raters who responded were asked to fill out the first part of the background questionnaire (10 minutes at their convenience) to ensure that they had an MA degree and at least 2 years of ESL/EFL experience. Overall, there were approximately 80 participants who responded, 70 who filled out the questionnaire, and 23 NS and 23 NNS raters who completed all the steps of quantitative data collection. The raters were remunerated for their participation in the study.

Individual rater training Skype sessions. After recruitment, the raters took part in individual rater training Skype sessions that lasted 1.5 hours for each rater. The researcher held individual online rater training session following the same script (see Appendix F). Rater training was operationalized as explaining and discussing the rubric and the tasks, listening to benchmarked samples, practicing using the rubric by scoring benchmarked samples, discussing rationales for the assigned grades, and passing calibration.

For rater training, the raters were provided with 24 benchmarked recordings; 12 in response to Task 1 and 12 in response to Task 2. All the benchmarked recordings included eight recordings from each L1 background representing each band score as was described in Table 3. Six responses to Task 1 and six responses to Task 2 were randomly assigned to appear in the norming part, and the rest appeared in the calibration part.

Rubric and task familiarization. First, each rater read and studied the tasks and the rubric to become familiar with them. Then the researcher explained the rubric following the same script (see Appendix F) and discussed any questions with the raters. After that the researcher informed the raters that the recordings were from a proficiency test, so the students did not study this topic in the classroom, and their ideas were on-the-spot ideas. Also, the researcher indicated that the students who took this test came from various language backgrounds, so the raters can expect to hear students from different L1 backgrounds (but no L1s were specified). If a rater inquired about what examinee L1s they would grade, the researcher politely responded that this information could not be provided.

Norming. Rubric familiarization was followed by the norming session. Each rater listened to benchmark speech samples in order to be normed and adopt the rubric. Each rater listened to 12 recordings with 6 of the recordings in response to Task 1 and 6 in response to Task 2. The raters listened to the recordings in the same order. This order was randomly assigned to the recordings by the researcher by putting the numbers of the recordings into a hat and drawing one at a time. Each rater listened to a recording once; then, if needed, the rater was allowed to listen to the recording again and pause if needed. For the first 6 recordings, the researcher indicated what holistic score each recording was given and discussed the rationale for giving the score using the rubric descriptions. For the next 6 recordings, based on the rubric, the raters assigned their own sub-category scores (Overall, Delivery, Language Use, and Topic Development) providing their own justifications for each score. Next, the researcher provided the predetermined score given to the recording, and if the rater's holistic score was different, the rater tried to adjust their grading.

Calibration practice. After rubric familiarization and norming, each rater had a calibration scoring session. Each rater listened to another set of 12 recordings with six of the recordings in response to Task # 1 and six in response to Task # 2. Raters listened to the recordings in the same order that was randomly assigned to the recordings by the researcher by putting the numbers of recordings into a hat and drawing one at a time. Each rater was allowed to listen to a recording once; then, if needed, the rater was allowed to listen to the recording again and pause. Then, based on the rubric, the rater assigned their own sub-category scores (overall, delivery, language use, and topic development) to the recordings.

Individual rating. After the training, the raters scored 39 recordings at their convenience, which took about 2 hours per rater. The raters were provided with a Qualtrics link where they listened to the recordings and assigned their grades. Overall, the raters graded 75 recordings and left comments for each speech sample. As Facets does not require each test-taker to be rated by each rater (Linacre, 2012; Linacre & Wright, 2002; Myford & Wolf, 2000), the data were partially crossed, so each rater graded 39 recordings, where 30 recordings were fully crossed, and 9 recordings were partially crossed. The researcher decided to use partially linked data because of practicality issues: This design helped to save time and score more student responses, therefore, allow the researcher to look at a larger amount of quantitative data.

The outlined rating plan (see Appendix G) defined how groups of 10 raters were connected. All the raters were randomly divided into groups of 10. Each rater out of 46 was randomly assigned a number from 1 to 10 by using a random number generator, which determined what recordings they graded. The current rating plan provided the strongest linking network by connecting all the raters directly together when not all the raters rate all examinees. Each rater rated 39 examinees: 30 common and 9 partially crossed recordings.

Overall, there were 45 partially crossed recordings. The linking pattern enabled connection of all the raters directly together when each of 45 test-takers was rated by at least 8 out 43 raters. The 30 fully-crossed recordings served as anchored data, and additional linking data were ensured when the raters scored common partially crossed recordings connecting them to other raters through commonly rated students. The 30 fully-crossed speech samples included speakers from different L1s and proficiency levels.

The level of proficiency of test-takers for each L1 background varied from 1 to 4 with most of the samples falling into 2-3 proficiency band as described in Table 4. There were 25 Arabic recordings, 25 Chinese, and 25 Russian. Each rater scored on average 13 recordings produced by test-takers from each L1 background. Each rater scored 39 recordings with 19 or 20 in response to Task 1 and 19 or 20 in response to Task 2. The way the examinees' proficiency level and L1 background were distributed is described in Table 4. The raters scored their recordings task by task with Task 1 scored first.

Raters' comments. After the raters provided their qualitative comments, they were asked to provide one or more comments about each rated recording in a text box. The stimulus for this was as follows, "What positive or negative feature(s) of this speaking stood out for you?". The raters were not required to leave comments, and it was possible to leave this box empty and proceed to the next recording. The raters were not limited in the number of words they could write. The raters were told that they could type any kind of comments (i.e., positive or negative) about anything that stood out for them about students' performance (e.g., grammar, pronunciation, ideas) after they had assigned their grades.

Individual think-aloud and interview Skype sessions. After all the quantitative data were collected and analyzed, 16 raters (7 native and 9 non-native speakers) who participated in the quantitative part also took part in the qualitative data collection, both think-aloud protocols and interviews. These raters were chosen based on the quantitative analyses (consistent and inconsistent raters). This selection process is described later in more detail. The raters who agreed to participate in the qualitative part reviewed the benchmarks verbalizing justifications, performed reflective rating with a think-aloud protocol, and answered interview questions. All these steps took about 2.5 hours per rater and were done in one meeting. All the raters received monetary compensation for their participation.

In order to be consistent during the think-aloud protocols and interviews, the researcher followed the same script (Appendices H and I). In addition, throughout the qualitative data collection, the researcher kept a neutral stance and avoided any references to researcher's hypotheses about accents or differences between NS and NNS raters in order not to influence raters' responses. In addition, the researcher avoided phrases that show that the researcher agreed or disagreed with raters' opinions, including "Okay" and "Uhum". If a rater did not provide enough information, the researcher used phrases such as "Because?," "Could you elaborate?" and other neutral expressions. The raters usually did not need to be prompted more than once and were responsive to such elaboration elicitations.

Rater selection for qualitative inquiry. This section provides further details about the rater selection procedure mentioned before. To select raters to participate in think-aloud protocols and interviews, Facets statistics were used. The rater measurement reports were obtained using the data when all the scores from all the rubric criteria were analyzed together, and when the scores from each rubric criterion (i.e., Overall, Delivery, Language Use, and Topic

Development) were analyzed separately. The researcher chose internally consistent (with stable infit statistics) and inconsistent raters (with unstable infit statistics) to participate in order to compare their decision-making patterns. Initially, 26 fitting and misfitting raters were selected to participate, but 10 of them either did not reply or rejected the invitation for various reasons, for example, they could not commit within the suggested timeframe.

Table 6. *Rater Severity and Infit Measures*

	Severity All	Infit All	Severity O	Infit O	Severity D	Infit D	Severity LU	Infit LU	Severity TD	Infit TD
NS7	-.41	1.28	-1.50	1.38	-1.89	1.63	-1.18	1.37	-1.24	1.58
NS14	-.37	.97	-1.38	.75	-1.13	.98	-1.47	1.09	-2.04	.91
NS17	.92	1.20	-.07	1.50	-.14	1.36	-.06	1.21	-.23	1.13
NS30*	.60	.87	-.40	.99	-.03	.97	-.62	.90	-.71	.70
NS34*	.21	.90	-.95	.91	-.63	1.01	-.99	.83	-.96	.84
NS37	-.15	1.40	-.71	1.46	-.58	1.14	-1.29	1.59	-1.91	1.55
NS40#	-.23	1.01	-1.17	1.29	-.99	1.01	-1.59	.87	-1.72	.95
NNS2#	.95	1.06	-.07	1.04	-.08	1.24	-.12	.85	-.17	.99
NNS6#	-1.17	1.06	-2.24	1.00	-2.61	.98	-2.14	1.34	-2.39	.98
NNS10	-2.19	1.07	-3.45	1.25	-2.68	1.00	-3.40	.85	-3.78	1.73
NNS13	.14	1.24	-.51	1.21	-1.30	1.05	-.53	1.45	-1.29	1.32
NNS24*	.50	.87	-.78	.99	.07	.73	-.73	.83	-.90	.88
NNS28	-.14	1.34	-.94	1.57	-1.01	1.26	-1.05	1.47	-1.66	1.16
NNS35*	-.44	1.13	-1.47	1.18	-2.11	.98	-1.73	1.03	-.97	1.06
NNS45#	.20	1.09	-.94	1.05	-.38	1.47	-.95	.88	-1.38	.82
NNS46*	.04	1.03	-1.05	1.15	-.71	1.04	-1.03	1.17	-1.47	.69

Note. D – Delivery, LU – Language Use, TD – Topic Development, G – General. * indicates a consistent rater; # indicates a semi-consistent rater.

Based on the Facets infit indices, the raters who participated in the qualitative part of the study represented stable, semi-stable, and unstable raters. Table 6 provides severity and infit measures for 16 raters who participated in the think-aloud. Based on the values, NS14, NS30, NS34, NNS24, NNS35, and NNS46 were the raters whose values never exceeded the rigorous 1.20 infit cut-off. This means that they maintained their overall consistency and were not prone to awarding haphazard or unexpected scores. Nevertheless, it is important to mention that sometimes their values went below the .80 limit, namely NS14 (Overall), NS30 (Topic

Development), NNS24 (Delivery), and NNS46 (Topic Development), which means that the raters exhibited over-consistency (overfit). Furthermore, some raters, NS40, NNS2, NNS6, and NNS45, exceeded the 1.20 limit only once, so they can also be considered almost stable or semi-consistent raters; these raters did not have any values below .80. Finally, raters NS7, NS17, NS37, NNS10, NNS13, and NNS28 exceeded the limit more than once and, therefore, were considered misfitting or not consistent raters. In sum, the pool of raters for the qualitative part included six stable, four semi-stable, and six unstable raters, whose decision-making patterns were examined.

Benchmark review. Before the think-aloud session started, the researcher provided a rater training refresher. First, the researcher discussed the prompts and the rubric once again. Next, guided by the researcher, each rater reviewed 6 benchmarks (3 for Task 1 and 3 for Task 2), which were used in the rater training before. Having listened to a recording, the raters provided their scores and verbalized their sub-category score justifications in the manner of a think-aloud. The researcher was following the same script with all raters in order to be consistent (Appendix H).

Reflective rating with think-aloud protocols. After the benchmark review, the raters performed reflective rating with think-aloud protocols. This think-aloud protocol was called reflective since the raters had an opportunity to reflect on their scores by comparing them to a score one band lower and one band higher. It was decided to give the raters this opportunity because the pilot study showed that such comparisons provided more insights into rater's own beliefs about their scoring. In addition, the raters in the pilot and the dissertation study found it engaging; such reflective comparisons made rating process more interesting and less tiring for the raters.

At the beginning of the session, the researcher explained what think-aloud protocols entail. Each reflective think-aloud rating was audio reordered. No samples of think-aloud protocols were provided in order not to bias raters' own verbal reporting styles (Gass & Mackey, 2000). The process of benchmark review with verbalizing justifications gave the raters the needed practice to adopt the process of verbalizing their thoughts and rationales while arriving at scores; therefore, no additional practice with think-aloud protocols was given.

The raters were prompted to follow their usual rating processes while verbalizing their thoughts. The raters were given 12 recordings selected for the think-aloud session based on the quantitative analysis (Table 5). These recordings were selected because they either were given a variety of criteria scores, or received grades, which were farther away from statistically predicted grades by Facets (misfit).

Due to the nature of rating speaking, the think-aloud protocols were retrospective. The raters were able to pause, rewind and listen again to each recording as many times as needed; however, just as in the pilot study, the raters chose not to pause a recording in order to verbalize their thoughts, even though they were given this opportunity. The raters either (a) listened from the beginning to the end and then verbalized their thoughts, (b) listened twice and then verbalized, or (c) listened once, verbalized, re-listened to the whole or half of the recording to confirm their decision, and then verbalized again. Each rater listened to the recordings in the same order that was randomly assigned to the recordings by the researcher by putting the numbers of the recordings into a hat and drawing one at a time.

To provide more details about the process, all the steps are described in the sequential order. First, each rater thought-aloud while arriving at analytic scores and explained their thinking processes and rationale for the specific criteria score. Many raters noted that this activity was natural for them as they sometimes think-aloud while grading their students' work.

Second, after the raters arrived at the scores, the researcher asked the raters to describe what they were doing while listening (e.g., thoughts and actions). This question was not asked right after the rater completed listening in order not to disturb the rating process and not to make a pause between listening to the speech sample and assigning scores. This was done because it is cognitively demanding to keep student's mental response in the working memory while talking about the actions during listening. The researcher decided to use this question since the pilot study showed that this question effectively elicited information relevant to the research questions.

Third, after the raters reported what they did while listening, the researcher elicited information about the level of perceived difficulty of scoring the recording and possible reasons for that. It was decided to enquire this because the pilot study highlighted that this question elicited more insights into raters' decision-making processes.

Finally, the researcher stated that some other raters gave a higher/lower score to this same recording and asked the raters to reflect on their own rating in order to uncover the potential reasons from the rater's perspective that allowed other raters to decide on a higher/lower holistic score. The raters were informed that their grades are not wrong or incorrect, but the researcher is trying to get the participants' expert insights about what could possibly have prompted other raters to give a different score.

Interview questions. Additional qualitative data were collected by asking interview questions to further elaborate some noticeable patterns from the think-aloud and ask additional questions. The interview was held right after the reflective grading with think-aloud protocols.

This session allowed the raters to reflect on the rating experience and express their concerns or difficulties during the rating procedure. The researcher prompted the raters to share their scoring patterns and strategies, perceived level of leniency or severity, levels of importance for each rating category, potential use of non-rubric criteria, and level of accent familiarity. The interview questions were compiled for the purpose of the study in order to ask raters about the topics of research interest (Appendix I).

During the interviews, the researcher avoided any references to researcher’s hypotheses about accents and differences between NS and NNS raters in order not to influence raters’ responses. Moreover, the researcher avoided phrases that showed that the researcher agreed or disagreed with a rater’s opinions by using neutral phrases, including “Okay” and “Uhum.” If a rater did not provide enough information, the researcher used phrases such as, “Because?,” “Could you elaborate?” and other neutral expressions. In order to be consistent, the researcher followed the same question order and the same script with all raters (Appendix I). The raters could also see the questions on their computer screen.

Accent familiarity. As the final step of the data collection, all 46 raters filled out the second part of the rater background questionnaire to describe their accent familiarity. The raters reported their familiarity using two methods (a) with examinee L1 identification and (b) without examinee L1 identification. For both methods, the raters provided their familiarity scores using the 6-point familiarity scale: 1 - No, 2 - Very Little, 3 - Little, 4 - Some, 5 - A Lot, 6 - Extensive (end of Appendix B). The first method (with examinee L1 identification) asked the raters to report their familiarity with Arabic, Chinese, and Russian L1 speakers based on (a) general familiarity with L1 speakers, (b) communication in English with speakers of these L1s, and (c) experience teaching speakers from these L1s. The second method (without examinee L1 identification) entailed reporting general familiarity with three L1s used in the study after listening to short recordings. The raters provided their familiarity scores for 24 recordings that were previously used for rater training and calibration (Arabic = 8, Chinese = 8, Russian = 8). Only the first 12 seconds of each recording were used because the raters were asked for their impressionistic decisions.

Data Analyses

According to the research questions and data collected, there were quantitative, qualitative, and mixed methods analyses to investigate rater variation. The first quantitative research question used data from individual ratings and coded comments to explore rater variation statistically. The second qualitative research question used data from think-aloud protocols and interviews to investigate the cognitive processes of raters and their perceptions of their rating processes. The last research question synthesized both quantitative and qualitative data to see the relationship between quantitative and qualitative findings. This section provides information about the analyses employed to answer each research question. For readers' convenience the research questions are repeated below:

RQ1: What are the differences between native and non-native rater groups in terms of their scoring patterns and comments that they provided on test-takers' performance?

a. To what extent do NS and NNS raters differ in terms of consistency and severity of their analytic scores?

b. Do NS and NNS raters show evidence of differential rater functioning related to rubric sub-criteria and examinee L1?

c. To what extent do NS and NNS raters differ in terms of scoring examinees by L1?

d. To what extent do NS and NNS raters differ in terms of the reported accent familiarity?

e. Is there a relationship between raters' familiarity, severity, and consistency?

f. To what extent do NS and NNS groups of raters differ in terms of the number and direction of their comments?

RQ2: What scoring strategies do NS and NNS raters use while grading L2 speaking performance?

RQ3: How do quantitative and qualitative findings complement each other?

Quantitative analyses. The first research question, "What are the differences between native and non-native rater groups in terms of their scoring patterns and comments that they provided on test-takers' performance?" was answered utilizing Facets analyses of raters' quantitative scores and content analysis of raters' comments. First, this sub-section overviews Facets analyses that were used to answer sub-questions *a* through *e*. After that another sub-section describes content analysis used for coding raters' comments for sub-question *f*; this sub-section describes the coding scheme and procedures. Finally, the sub-question *f* is introduced.

MFRM analyses. Since interrater reliability does not allow examining raters' ratings at the individual level, the Many-Facet Rasch Measurement (MFRM) model was used to provide more precise information to answer the first research question. Such analysis was chosen because it offers fine-grained information to better understand the individual scoring patterns of raters. MFRM allows detection of potential rater characteristics from a statistical perspective. Rater characteristics that can be statistically revealed are rater leniency/severity, centrality, inaccuracy, and differential dimensionality (also called differential rater functioning (Wolf & McVay, 2004) or rater bias (Lumley & McNamara, 1995)). MFRM analysis has been applied substantively to model rater effects in the field of language testing and assessment for writing and speaking (Bachman et al., 1995; Eckes, 2005, 2008; Engelhard & Myford, 2003; Lumley, 2002; Lumley & McNamara, 1995; Myford & Wolfe, 2000, 2003, 2004).

The computer program Facets, version 3.71.4 (Linacre, 2014) was used for the analyses. The analyses were performed using the 7176 scores from 75 recordings (30 fully crossed and 45 partially crossed) awarded by 46 raters for 4 rubric criteria during the individual rating. To match the variables in the study, 3 main facets were included in the model: Examinee (N = 75); Rater (N=46); and Criteria (N = 4). Examinees included Arabic (*n* = 25), Chinese (*n* = 25), and Russian (*n* = 25). The raters were NS (*n* = 23) and NNS (*n* = 23). The criteria contained Overall, Delivery, Language Use, and Topic Development. the Examinee facet was non-centered due to the established convention of centering the frame of reference (e.g., raters, tasks) and allowing the objects of measurements (e.g., test-takers) to float to be placed based on the frame of reference (Linacre, 2017; Winke et al., 2011). In addition, there were two dummy facets anchored at zero, which were used for bias (interaction) analyses: Examinee Group (N = 3), Arabic, Chinese, and Russian; and Rater Group (N = 2), native and non-native. Data were analyzed using general models and two bias models.

Sub-question a. This sub-question asked, “To what extent do NS and NNS raters differ in terms of consistency and severity of their analytic scores?” This question was answered based on the Facets output for rater consistency (infit values) and rater severity (measure logit values). Data for the NS and NNS rater groups were compared using independent *t*-tests. In addition, data for NS and NNS raters were analyzed in Facets separately and, based on the Facets output for criteria difficulty, two groups of raters were descriptively compared.

Sub-questions b. This sub-question asked, “Do NS and NNS raters show evidence of differential rater functioning related to rubric sub-criteria and examinee L1?” The question was answered using differential rater functioning analysis or bias analyses that checks for interactions between facets. Bias analyses reveal any deviations from what is expected by the model, in other words, it uncovers any unexpected tendencies of raters who exercise more severe or lenient judgments. In general, if raters exhibit any bias, Facets reports a table with bias size (in logits) for each rater and information on how significant the bias is (in *t*-scores).

Two bias models were specified in Facets for each question. The first model checked interactions between the Rater L1 Group and the Rating Criteria facet. To this end, the analytic scores for each rating category (i.e., Overall, Delivery, Language Use, and Topic Development) given by NS and NNS rater group were analyzed to explore if any raters showed any bias towards any rating criteria. The second model analyzed interactions between the Fater L1 Group and the Examinee L1 Group facet. To achieve this, the dummy facet Examinee L1 Group was used to explore the potential bias of the raters towards examinees’ L1.

Sub-question c. This sub-question asked, “To what extent do NS and NNS raters differ in terms of scoring examinees by L1?” To answer this research sub-question, three separate Facets analyses were run by rater group (NS and NNS) for each examinee L1.

Sub-question d. This sub-question asked, “To what extent do NS and NNS raters differ in terms of the reported accent familiarity?” To answer this sub-question, accent familiarity ratings for NS and NNS were calculated based on two measures (a) familiarity with identification meaning that the raters knew which examinee L1s they were reporting their familiarity for; and (b) familiarity without identification, when examinee L1s were not identified when the raters reported their familiarity.

For the first measure, familiarity with identification, the raters self-reported their familiarity with the examinee L1s used in the study (i.e., Arabic, Chinese, and Russian) in terms of general familiarity, communication, and teaching experience. The 6-point scale was used to measure raters' perceived accent familiarity (No, Very Little, Little, Some, A Lot, Extensive). For the second measure, familiarity without identification, the raters listened to twenty-four 12-second recordings (eight from each L1) and reported their accent familiarity on the same 6-point scale. To compare the reported overall accent familiarity obtained by two measures, the percentages were calculated and descriptively compared. Two types of familiarity were used in order to provide a better picture about raters' accent familiarity.

Sub-question e. This sub-question asked, "Is there a relationship between raters' familiarity, severity, and consistency?" To answer this sub-question, a composite accent familiarity score was calculated based on two accent familiarity measures (see Procedures). A composite familiarity score was used in order to include both sides of familiarity – the way the raters perceived their accent familiarity when the raters new and did not know the examinees' L1s.

First, the raters were grouped into six groups (i.e, No, Little, Very Little, Some, A Lot, and Extensive familiarity) to see any observable patterns of a relationship between raters' severity and accent familiarity. Three Facets analyses were run, namely for Arabic L1, Chinese L1, and Russian L1 examinees.

Second, three correlation analyses (one for each L1) were run to determine if there is a relationship between raters' average severity and accent familiarity. Then, three more correlation analyses (one for each L1) were run to determine if there is a relationship between raters' average consistency and accent familiarity. Normality checks were performed for each variable resulting in deviations in normality for severity variable for Arabic L1 and consitency variables for Chinese and Russian L1s. Thus, Spearman, not Pearson correlations were used for these variables.

Finally, the raters were subdivided into familiar and unfamiliar groups based on their scores. The raters who had A Lot and Extensive familiarity were considered highly familiar and the raters who reported No, Very Little, Little, and Some familiarity were grouped as relatively unfamiliar. These new rater groups, familiar and unfamiliar, were analyzed in Facets. Three different Facets analyses were conducted, one for each examinee L1.

Content Analysis. Raters' comments were retrieved from Qualtrics, separated from the quantitative scores, and saved as separate Excel files. Each file had 39 lines of comments by the same rater where each line contained comments typed for one examinee (Appendix J). Not all the raters provided comments for all examinees since some raters skipped several. To answer the third research question, all the comments provided by raters (*N* = 3292) were coded and counted. Next, percentages for the comments by rater group and by individual rater were calculated and descriptively analyzed. The following sub-sections describe the coding scheme development and the coding process.

Coding scheme for raters' comments. To code the comments, a coding scheme (Appendix K) was developed by the researcher using content analysis (Strauss & Corbin, 1998). First, the TOEFL iBT independent rubric was scrutinized and all the descriptors from all the bands were compiled into four categories mentioned on the rubric: Delivery, Language Use, Topic Development, and General. Then the descriptors for each rubric category were added to the coding scheme and coded as numbers; Delivery was 1, Language Use was 2, Topic Development was 3, and General was 4. For example, comments such as "unintelligible speech" and “the speech is clear” were coded as 1; “poor vocabulary” and “bad grammar” were coded as 2; “well-developed ideas” and “only basic ideas” were coded as 3; “good response” and “well-presented” were coded as 4 (Appendix K).

Due to the fact that the raters were asked to provide either negative or positive comments, the raters’ comments were marked as negative, positive, or neutral based on their nature. For example, “unintelligible speech” was coded as a negative comment for Delivery, whereas “good development of ideas” was classified as a positive comment for Topic Development. A comment was considered neutral if it did not have any descriptors allowing attribution to either negative or positive side, for instance, “topic development,” and “vocabulary” were coded as neutral comments.

To code the comments, the researcher specified the unit of analysis as a word or a phrase that describes a negative, positive, or neutral feature that can be classified as relating to the rubric sub-categories (i.e., Delivery, Language Use, Topic Development, and General). For example, a comment “pauses and hesitations” was coded as two negative comments about Delivery; “poor grammar and lack of vocabulary” as two negative comments about Language Use; “well-developed reasons and good examples” as two positive comments about Topic Development. Moreover, the longer elaborated sentences were also subdivided into codes, for example, “That's a nice response, the speaker doesn't come across as having any trouble expressing her ideas” was coded as a positive General comment and a positive Topic Development comment (see Appendix K).

An undergraduate student assisting with the data analysis performed initial coding of all the comments, which was checked by the researcher afterward (the exact agreement rate was approximately 70%). After that, the researcher revised the coding scheme to clarify how to classify comments that can be attributed to more than one category, for example, “easy to follow,” “unable to understand,” and “not clear”. If no additional attributive language was present, such comments were labeled as General. However, if these comments were followed by words that could identify the direction, for example, “easy to follow the ideas,” “unable to understand reasoning,” “not clear ideas,” they were placed into the Topic Development category. If such comments were followed by other descriptors, for instance, “not clear speech” or “unable to understand pronunciation”, the comments were counted towards the Delivery criteria. After that, the researcher checked codes for all the comments.

Coding of raters’ comments. After the coding scheme was revised and comments checked, the researcher recruited and trained two additional coders who tagged 31% of the data (1035 out of 3292 comments). Each coder worked on comments by six raters, where two sets of comments were the same (one set by a NS and one by a NNS rater) in order to calculate the inter-coder reliability between the coders. The coders were two Ph.D. students specializing in assessment: a female non-native speaker (Coder 1) and a male native speaker (Coder 2) who received compensation for their assistance. The inter-coder reliability between Coder 1 and Coder 2 was 95% of exact agreement with Cohen’s kappa of .92 (based on 160 co-coded comments). The inter-coder reliability between Coder 1 and the researcher was 98% of exact agreement with .98 Cohen’s kappa (based on 456 co-coded comments). The inter-coder reliability between Coder 2 and the researcher was 98% of exact agreement with .97 Cohen’s kappa (based on 579 co-coded comments).

Sub-question f. This sub-question asked, “To what extent do NS and NSN raters differ in terms of the number and direction of their comments?” This sub-question was answered using raters’ comments coded using the described above content analysis (Strauss & Corbin, 1998).

Analyses of think-aloud protocols and interviews. The second research question asked, “What scoring strategies do NS and NNS raters use while grading L2 speaking performance?” To answer this research question, qualitative observations of phenomena in the form of themes from think-aloud protocols and interviews were analyzed. The think-aloud protocols and interview answers were transcribed and then thematically coded using content analysis (Strauss & Corbin, 1998), which is also called grounded theory. Grounded theory is defined as "theory that was derived from data, systematically gathered and analyzed through the research process" (Strauss & Corbin, 1998, p. 12). In this study, the new themes were generated from both think-aloud protocols and interviews. The qualitative data were coded using deductive reasoning based on the themes from the pilot study. This coding was selective, and the unit of analysis was a thought-group operationalized as a sequence of thematically connected utterances used to describe an action, a thought-process, or a belief.

In the pilot study, inductive reasoning was used and, according to the procedure described in Lumley (2005), the pilot study utilized a set of questions that was created to guide content analysis according to the topics of research interest: (a) What did the rater do while listening (e.g., take notes, look at the rubric, just listened, thinking about a holistic score)?, (b) What did the rater do while scoring (e.g., re-read the rubric silently, skim the rubric and read-aloud part of it out loud)?, (c) What other comments did the raters make (e.g., describing their beliefs and thoughts, concerns, non-rubric factors or non-linguistic factors)?, (d) How did the rater perceive their severity level (e.g., severe, lenient)?, and (e) How did the rater perceive the criteria importance (e.g., Delivery is more important than Language Use)? Additionally, application of inductive reasoning entailed reading and re-reading transcripts for several times in order to identify new emerging patterns that can form new themes.

The content analysis in this dissertation singled out patterns to answer the aforementioned questions using a coding scheme (Appendix L). In addition, raters' ability to distinguish examinee L1 groups was noted. In general, content analysis was centered on the patterns of cognitive differences in raters or rater groups during grading, differences in criteria importance, perceived importance of each rating criteria, differences in respect to examinees' L1 background, non-rubric references, references to raters' prior experiences or own beliefs, comments about examinees' individual differences such as confidence of speaking or softness of voice. Themes from the think-aloud protocols and interviews were analyzed together to identify patterns for each rater. Then, the patterns were examined to determine similarities among raters and rater groups.

The 16 raters who participated in the qualitative part produced approximately 60,8906 words total during the think-aloud protocols and approximately 47,470 words during the interview sessions. The average number of words per rater was 6,622 – the minimum words spoken was 4121 (NS17), and the maximum was 13,628 (NNS6).

Mixed methods analyses. The third research question asked, "How do quantitative and qualitative findings complement each other?" To answer this question, quantitative and qualitative results were synthesized to integrate the information in order to see how the findings from each strand complement each other. To this end, side-by-side comparison analysis (Onwuegbuzie & Teddlie, 2003) was employed.

Chapter 4: Results

This chapter presents results for three research questions, which were answered using quantitative, qualitative, and mixed methods, respectively. The research questions aimed to investigate the rating behavior of native (NS) and non-native (NNS) raters in order to uncover and classify differences in decision-making patterns when rating speaking performance of multilingual test-takers. Rater's accent familiarity with test-takers' L1s was also examined to discern the potential presence of familiarity effects when raters are familiar with examinees' L1. In addition, the study looked at another potential source of rater variability, which is the match between raters' and examinees' L1 background.

First, this section presents the quantitative results from Facets and correlation analyses of raters' scores and content analyses of raters' comments provided while scoring. Next, the qualitative results from content analysis of think-aloud protocols and interviews are provided. And then the results of synthesizing quantitative and qualitative findings are described.

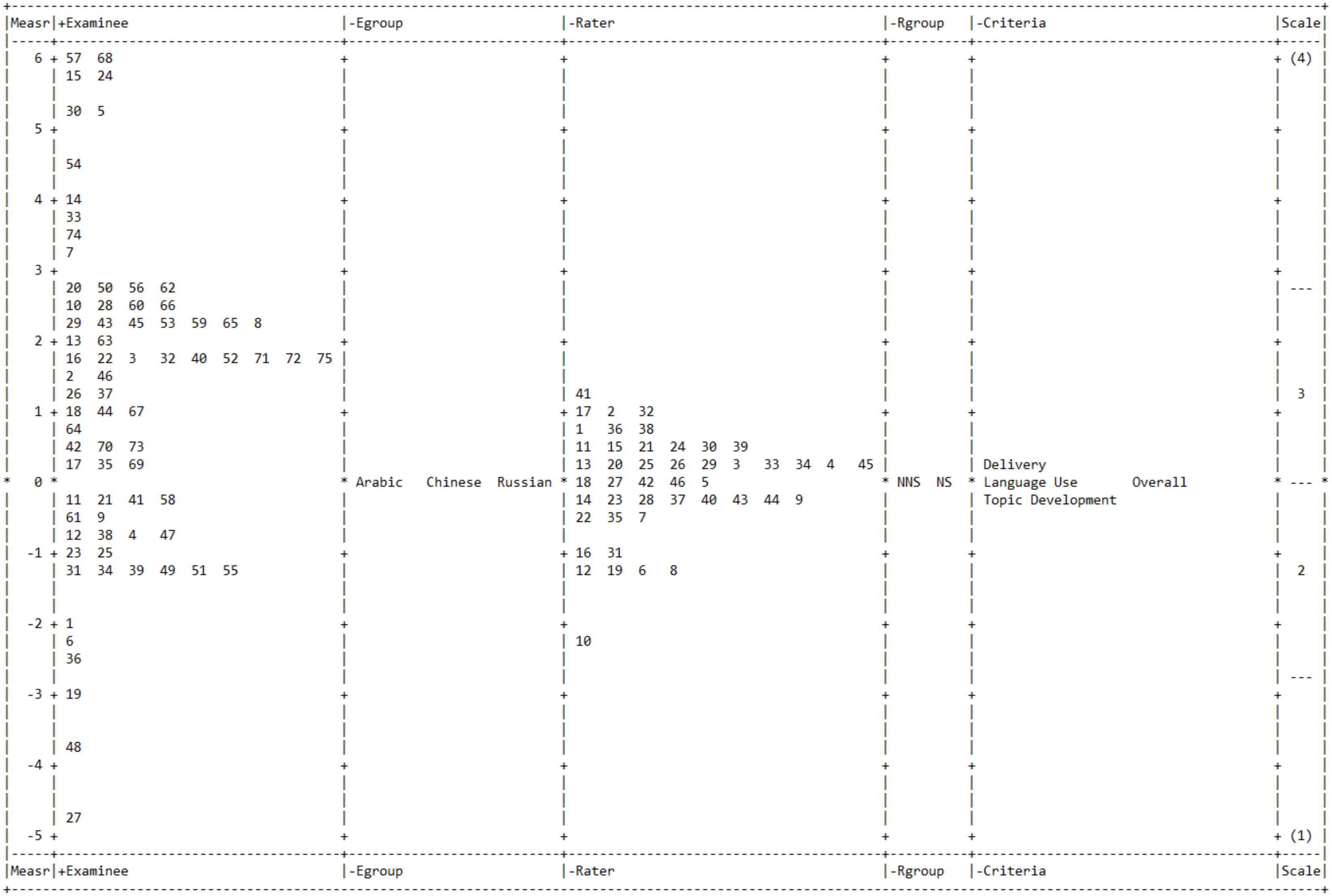
The quantitative results start with an overall Facets section that describes each facet and the rating criteria functioning, and the subsequent sections answer the research sub-questions. Then the quantitative analyses of rater accent familiarity and L1 match between raters and test-takers are presented. These results are followed by descriptions of rater types stemming from raters' comments. The qualitative results provide information about raters' decision-making patterns. This section describes raters' listening and grading strategies, non-rubric criteria for Delivery and Topic Development criteria, and raters' perceived severity and category importance. The chapter ends with a synthesis of quantitative and qualitative results where raters' statistical and perceived severity are compared first and followed by an overview of raters' infit values through the lens of their decision-making patterns.

Overall Facets Results

Facets summary. The computer program Facets calibrates examinees, raters, and the rating scale to position all facets on the same scale. Thus, the results from all facets can be interpreted based on a single frame of reference. The model scale is in log-odds units, or logits, that constitute an equal-interval scale by transforming the probability of receiving a specific response to show true differences among facets (see Eckes, 2011; McNamara, 1996). Such a single frame of reference for all the facets facilitates comparisons within and between them.

Figure 3 depicts the Facets summary or variable map where the five facets (i.e., Examinee, Examinee L1 Group, Rater, Rater L1 Group, and Rating Criteria) are compared by being put on the same logit scale. The first column of the variable map, measure, represents the difficulty expressed in logits, where the average difficulty is conventionally set at zero logits for all facets except for the non-centered examinee facet, which was allowed to float. Thus, the Examinee facet created a single framework of reference and enabled interpreting the measures for all other facets based on this comparable logit scale (Linacre, 2017; Winke et al., 2011). For examinees, the measure in logits shows their ability; for raters, it displays their severity; and, for criteria, it represents difficulty. Based on this logit measure, the elements in all facets are positioned higher or lower, which indicates differences.

The second column of the variable map represents the examinee facet comparing 75 of them in terms of their speaking ability. Lower ability speakers are placed at the bottom of the column and higher ability speakers at the top. An examinee whose ability is expressed as 0 logits has a 50 percent chance of getting an item of average difficulty right. The third column of the variable map is the dummy facet indicating Examinee L1 Group, which was based on the test-takers' L1s (i.e., Arabic, Chinese, and Russian). The Examinee L1 Group facet was not used



```
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee                           |-Egroup                    |-Rater                                  |-Rgroup   |-Criteria                             |Scale|
|-----+------------------------------------+---------------------------+----------------------------------------+----------+--------------------------------------+-----|
|   6 + 57  68                             +                           +                                        +          +                                      + (4) |
|     | 15  24                             |                           |                                        |          |                                      |     |
|     |                                    |                           |                                        |          |                                      |     |
|     | 30  5                              |                           |                                        |          |                                      |     |
|   5 +                                    +                           +                                        +          +                                      +     |
|     |                                    |                           |                                        |          |                                      |     |
|     | 54                                 |                           |                                        |          |                                      |     |
|     |                                    |                           |                                        |          |                                      |     |
|   4 + 14                                 +                           +                                        +          +                                      +     |
|     | 33                                 |                           |                                        |          |                                      |     |
|     | 74                                 |                           |                                        |          |                                      |     |
|     | 7                                  |                           |                                        |          |                                      |     |
|   3 +                                    +                           +                                        +          +                                      +     |
|     | 20  50  56  62                     |                           |                                        |          |                                      | --- |
|     | 10  28  60  66                     |                           |                                        |          |                                      |     |
|     | 29  43  45  53  59  65  8          |                           |                                        |          |                                      |     |
|   2 + 13  63                             +                           +                                        +          +                                      +     |
|     | 16  22  3   32  40  52  71  72  75 |                           |                                        |          |                                      |     |
|     | 2   46                             |                           |                                        |          |                                      |     |
|     | 26  37                             |                           | 41                                     |          |                                      |  3  |
|   1 + 18  44  67                         +                           + 17  2   32                             +          +                                      +     |
|     | 64                                 |                           | 1   36  38                             |          |                                      |     |
|     | 42  70  73                         |                           | 11  15  21  24  30  39                 |          |                                      |     |
|     | 17  35  69                         |                           | 13  20  25  26  29  3   33  34  4   45 |          | Delivery                             |     |
*   0 *                                    * Arabic   Chinese  Russian * 18  27  42  46  5                      * NNS  NS  * Language Use       Overall           * --- *
|     | 11  21  41  58                     |                           | 14  23  28  37  40  43  44  9          |          | Topic Development                    |     |
|     | 61  9                              |                           | 22  35  7                              |          |                                      |     |
|     | 12  38  4   47                     |                           |                                        |          |                                      |     |
|  -1 + 23  25                             +                           + 16  31                                 +          +                                      +     |
|     | 31  34  39  49  51  55             |                           | 12  19  6   8                          |          |                                      |  2  |
|     |                                    |                           |                                        |          |                                      |     |
|     |                                    |                           |                                        |          |                                      |     |
|  -2 + 1                                  +                           +                                        +          +                                      +     |
|     | 6                                  |                           | 10                                     |          |                                      |     |
|     | 36                                 |                           |                                        |          |                                      |     |
|     |                                    |                           |                                        |          |                                      | --- |
|  -3 + 19                                 +                           +                                        +          +                                      +     |
|     |                                    |                           |                                        |          |                                      |     |
|     |                                    |                           |                                        |          |                                      |     |
|     | 48                                 |                           |                                        |          |                                      |     |
|  -4 +                                    +                           +                                        +          +                                      +     |
|     |                                    |                           |                                        |          |                                      |     |
|     |                                    |                           |                                        |          |                                      |     |
|     | 27                                 |                           |                                        |          |                                      |     |
|  -5 +                                    +                           +                                        +          +                                      + (1) |
|-----+------------------------------------+---------------------------+----------------------------------------+----------+--------------------------------------+-----|
|Measr|+Examinee                           |-Egroup                    |-Rater                                  |-Rgroup   |-Criteria                             |Scale|
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Figure 3. Variable map

for estimating the measures, therefore, this facet was anchored at zero and used only for bias analyses to indicate any unexpected rating patterns. Due to this anchoring, the elements within this facet are at the same logit measure, 0 logits. Unlike other facets, the position of this facet does not provide any information about the examinee groups.

The fourth column shows 46 raters where more severe ones are at the top and more lenient ones are at the bottom. The fifth column is the dummy facet of Rater L1 Group according to rater's L1 where NS stands for native speaking North American English L1 raters and NNS for non-native speaking Russian L1 raters. Just like the Examinee L1 Group facet, the Rater L1 Group facet was not used for estimating the measures but only for bias analyses to investigate any interactions caused by unexpected rating patterns of raters in either group. Rater L1 Group facet was anchored at zero and that is why the elements within this facet are positioned at the same measure of 0 logits; this logit measure does not provide any information about the rater groups.

The seventh column of the variable map represents the Rating Criteria facet where the names indicate the rubric categories: Overall, Delivery, Language Use, and Topic Development. This column shows more severely rated criteria at the top and more leniently rated criteria at the bottom. The last column in Figure 4 represents the rating scale. The horizontal lines across the column show when the likelihood of getting a higher rating starts to exceed. For example, the examinees with logits between -5 and -3 were more likely to receive a score of 1, whereas the examinees in-between -3 and 0 were more likely to receive a rating of 2. The length of each scale point on the variable map tells us about rating point utilization – the longer the distance between the lines separating the numbers, the more often that score was used.

Explanation of measure and fit statistics. Facets measurement reports provide measure and fit statistics for each facet. For examinees, the measure in logits stands for their ability, for raters, it displays their severity, and, for criteria, it means difficulty. Regarding fit statistics, Facets reports infit and outfit mean squares, which measure if anything diverges from the expected pattern predicted by the model (Weigle, 1998). The infit is weighted and sensitive to unexpected responses, whereas the outfit is not weighted and sensitive to extreme scores. Since the outfit statistic is more affected by outliers, the infit statistic is preferred by researchers. The infit statistic shows how predictable examinees' scores are, how self-consistently the raters awarded the scores, and how appropriate the rating criteria were. Later, more details will be provided on how to interpret fit statistics for each facet.

For the fit statistic, the value of 1 is considered to be ideal, and variation between 1.2 and .80 is considered to be the more conservative, strict criteria (McNamara, 1996). However, Myford and Wolfe (2004a) noted that there are no clear-cut rules for setting these upper and lower bounds, and that decisions can depend on the context of assessment. If the assessment context is high-stakes, then the conventional 1.2 and .8 (McNamara, 1996) or 1.4 and .6 (Bond & Fox, 2007) should be utilized whereas low-stakes contexts can adopt looser limits. Since rater performance is important in this study, the strict criteria of 1.2 and .8 was used as the acceptable infit statistic range. Additionally, to provide more information, all the tables demonstrate not only infit but also outfit values beyond 1.2 and 0.8.

Examinee measurement report. The second column of Figure 4 shows that the examinees were widely spread out in terms of their ability levels, ranging from -4.83 to 7.48 (*M* = 1.21, *SD* = 2.33) with a total spread of 12.31 logits. Lower ability examinees were placed at the bottom of the column and higher ability speakers at the top. According to the ability logit values, 51 examinees were placed above the average ability of zero logits and 24 test-takers were located below this value. The positive mean indicates that the test was not difficult for this group of students. The separation index 5.96 with strata 8.29 and reliability of .97, (χ^2 = 7399.0, *df* = 74, *p* < .01) indicates that the speaking tasks reliably separated 75 speakers into eight distinguishable ability levels.

For the Examinee facet, the fit statistic shows whether the scores that the examinees received approximate the model-predicted scores, and the ideal value of 1 exemplifies such close approximation. Infit values higher than 1.2 (misfit) show that these ratings are farther from what the model expected, and infit values lower than 0.8 (outfit) mean that these ratings are closer to what the model predicted. In other words, misfit flags noisiness and unpredictability in scores, for example, due to inter-rater disagreement; and outfit shows that test-takers received the same scores regardless of the differences in ability, or they were given the same scores on different criteria (Barkaoui, 2014). Both misfit and outfit highlight that those examinees were not appropriately measured by the test, but misfit is usually considered a bigger problem (McNamara, 1996; Bond & Fox, 2007). Table 7 shows only misfitting and overfitting examinees, while 48 examinees who had in-between values are omitted and indicated by "--". Table 7 illustrates that examinee fit statistic indices ranged between 2.48 and 0.36 demonstrating that 10 test-takers exceeded the upper-control limit (misfit) of 1.2 and 17 the lower-control limit (overfit) of 0.8. The ability of 10 examinees who showed misfit was not measured appropriately and they can be highlighted as exhibiting unpredictable, erratic scores. The ability of 17 examinees who displayed overfit was measured too predictably or overly consistently meaning that there was a lack of score variation (Eckes, 2011). Lastly, there were two test-takers who received identical scores from all the raters for all the categories; therefore, their infit values were at maximum. Due to the context of the study, it was important to see which examinees were rated unpredictably to use these recordings for further qualitative inquiry.

Table 7. *Measurement Report for Speakers (Arranged by Infit Mean Square)*

Examinees	Ability Logit	Model Error	Infit Mean Square	ZStd	Outfit Mean Square	ZStd
46	1.53	.29	2.48	4.7	2.45	4.7
47	-0.65	.26	1.69	2.6	1.68	2.6
5	5.20	.28	1.03	0.2	1.59	1.6
63	1.97	.30	1.54	2.1	1.46	1.8
35	0.29	.27	1.50	2.0	1.51	2.0
6	-2.16	.13	1.48	4.4	1.47	4.5
37	1.31	.27	1.37	1.6	1.37	1.5
9	-0.60	.12	1.35	3.0	1.34	2.9
10	2.50	.13	1.28	2.7	1.34	3.1
31	-1.28	.28	1.30	1.2	1.31	1.2
13	1.92	.12	1.28	2.7	1.30	2.9
55	-1.25	.30	1.20	0.8	1.23	0.9
--	--	--	--		--	--
41	-0.30	.29	0.77	-0.9	0.76	-0.9
65	2.30	.27	0.75	-1.3	0.80	-1.0
36	-2.38	.31	0.75	-1.4	0.75	-1.4
73	0.58	.28	0.75	-1.1	0.75	-1.0
58	-0.30	.29	0.75	-0.9	0.75	-1.0
17	0.17	.12	0.74	-2.8	0.74	-2.8
60	2.38	.27	0.72	-1.5	0.74	-1.3
40	1.66	.26	0.68	-1.7	0.83	-0.8
22	1.83	.12	0.68	-3.7	0.69	-3.5
42	0.49	.28	0.64	-1.7	0.64	-1.7
49	-1.32	.25	0.64	-1.8	0.64	-1.8
52	1.66	.25	0.65	-1.9	0.63	-2.0
66	2.39	.29	0.61	-2.0	0.61	-2.0
56	2.83	.31	0.59	-2.1	0.58	-2.1
62	2.77	.32	0.59	-2.1	0.57	-2.1
50	2.87	.29	0.61	-2.2	0.56	-2.0
44	1.04	.28	0.36	-3.6	0.37	-3.6
57	7.48	1.83	Maximum			
68	7.47	1.83	Maximum			
M	1.21	.28	0.99	-0.1	0.97	-0.2
SD	2.33	.27	0.30	1.6	0.31	1.6

Note. Reliability = .97; Separation: 5.96; Strata: 8.29; Fixed chi-square: 7399.0 (*df* = 74; *p* < .01).

Rater measurement report. According to the fourth column in Figure 3, the raters were spread out based on their severity levels from the most severe rater 1.23 at the top to the most lenient rater -2.19 at the bottom with a total spread of 3.42 logits (*M* = 0.00, *SD* = .69). To clarify, raters with positive logit measures were more severe, and the raters with negative logit measures were more lenient. Generally speaking, there were 28 raters above the average severity level of 0 and 18 raters below, which means that there were more severe than lenient judges. The separation index of 4.59 and strata 6.45 with reliability of .95 illustrate that the raters reliably exercised approximately six levels of severity, which was confirmed by a significant fixed chi-square statistic (χ^2 = 935.2, *df* = 45, *p* < .01).

Table 8. *Measurement Report for Raters (Arranged by Infit Values)*

Raters	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
NS25	0.21	.14	1.26	2.1	2.15	5.8	.65
NS20	0.15	.14	1.43	3.3	1.80	4.3	.74
NS37	-.15	.14	1.40	3.1	1.61	3.1	.70
NNS10	-2.19	.17	1.07	.5	1.57	1.3	.73
NNS13	0.14	.15	1.24	1.9	1.51	2.9	.71
NS7	-0.41	.14	1.51	3.9	1.28	1.5	.71
NS41	1.23	.14	1.25	2.0	1.42	2.8	.71
NS17	0.92	.15	1.20	1.6	1.41	2.6	.81
NNS43	-0.29	.15	1.35	2.8	1.16	0.9	.79
NNS28	-0.14	.15	1.34	2.7	1.27	1.5	.76
NNS46	0.04	.15	1.03	.3	1.24	1.4	.76
NNS15	0.51	.15	1.21	1.6	1.09	0.6	.81
--	--	--	--	--	--	--	--
NS44	-0.33	.15	0.79	-1.9	0.71	-1.7	.82
NS26	0.26	.14	0.70	-2.8	0.87	-0.8	.81
NNS27	0.12	.14	0.76	-2.2	0.69	-2.2	.83
NNS31	-1.01	.15	0.76	-2.2	0.68	-1.5	.80
NNS19	-1.15	.15	0.74	-2.3	0.65	-1.6	.81
NS16	-0.93	.15	0.73	-2.4	0.62	-1.9	.84
NS21	0.49	.14	0.59	-4.2	0.62	-3.0	.86
NS33	0.13	.14	0.62	-3.8	0.56	-3.4	.87
NS9	-0.17	.14	0.64	-3.6	0.56	-3.2	.85
NS29	0.37	.14	0.55	-4.5	0.51	-4.0	.87
NNS18	.04	.15	.049	-5.4	.44	-4.4	.90
M	0.00	.15	1.00	-.1	1.00	-.2	.79
SD	0.69	.01	.24	2.2	.35	2.1	.05

Note. Reliability = .95; Separation: 4.59; Strata: 6.45; Fixed chi-square: 935.2 (*df* = 45; *p* < .01). Inter-Rater agreement opportunities: 131156; Exact agreements: 68597 = 52.3 %, Expected: 6726 = 51.6%.

The fit statistic shows to what degree each rater exhibited self-consistency or whether they awarded scores predictably or erratically. The raters who have values higher than the upper-control limit of 1.2 show misfit, and the raters who have values lower than the lower-control limit of 0.8 show overfit. Misfitting and overfitting raters can be interpreted as follows. The misfitting raters show unexpected rating behavior, in other words, they tend to score speakers' performance in an erratic, unpredictable way and they are not self-consistent in assigning scores. The overfitting raters are overly consistent and show too little variation, in other words, these raters tend to assign similar scores to speakers of different ability, and they do not utilize the whole rating scale appropriately to measure speakers' ability. Table 8 shows only misfitting or overfitting raters, while 26 raters who had good infit values are omitted and indicated by "--". Table 8 shows that nine raters exceeded the upper-control limit of 1.2, and 11 raters exceeded the lower-control limit of 0.8. Since raters' behavior was the interest of the study, it is informative to see the spread in raters' severity and infit values.

Rating criteria measurement report. Column number six (Figure 3) illustrates four rating criteria (i.e., Overall, Delivery, Language Use, and Topic Development) where more difficult criteria are at the top, and less difficult are at the bottom. Logit measures for the rating criteria can be interpreted from the raters' and examinees' perspectives. From the raters' perspective, the criteria can be scored more severely or more leniently, and, from the examinees' perspective, the rating criteria can be more difficult or easier. The positive logits indicate more difficult or more harshly scored criteria, whereas the criteria with negative logits are easier or scored less severely.

According to Table 9, the more harshly scored criteria were Delivery and Overall with .13 and .11 logits, and the more leniently scored categories were Language Use and Topic

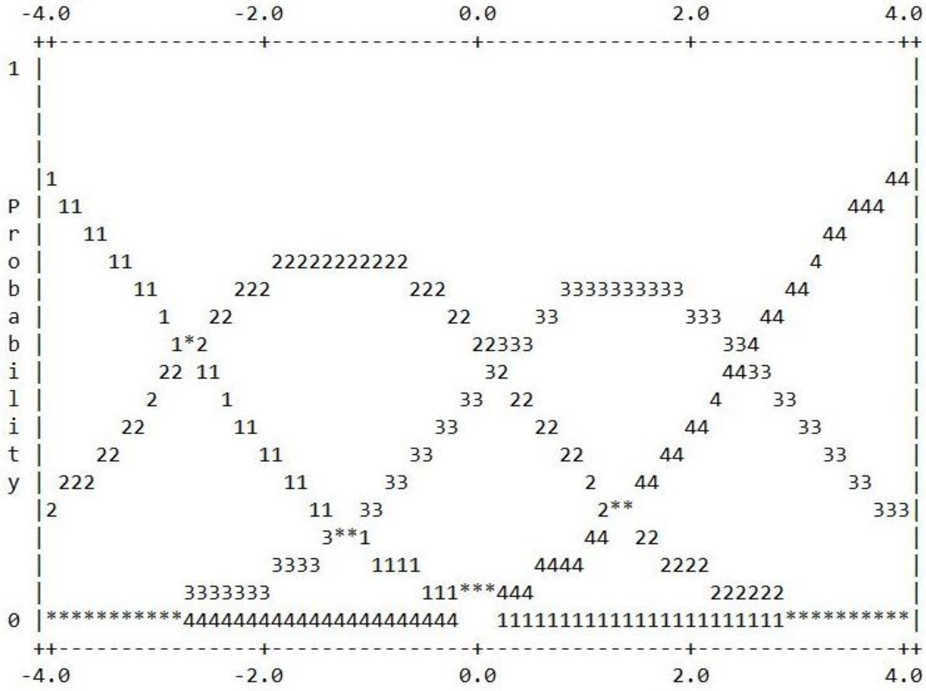
Development whose difficulty measures were -.06 and -.18. As it can be seen from significant chi-square statistic (χ^2 = 37.0, *df* = 3, p < .01), high reliability of .89, separation index 2.88 and strata 4.17, the four rating categories represented four levels of difficulty and were not interchangeable. Furthermore, the fit statistics ranged between 1.09 and .82 indicating a good fit that is very close to the ideal value of 1, which indicates that all four criteria were stably rated. Additionally, this serves the evidence of unidimensionality (Bond & Fox, 2007), which is an important assumption for Facets analysis. To clarify, unidimensionality means that the measured sub-constructs represent the same larger construct. In conclusion, all rating categories had different levels of difficulty and stable infit statistics; therefore, each rubric sub-category tapped into a distinct aspect of students' speaking ability and the raters were able to differentiate and use the rubric criteria consistently.

Table 9. *Measurement Report for Rating Criteria (Arranged by Difficulty Measure)*

Criteria	Difficulty logit	Model error	Infit mean square
Delivery	.13	.04	1.07
Overall	.11	.04	.82
Language Use	-.06	.04	1.02
Topic Development	-.18	.04	1.09
M	.00	.04	1.00
SD	.13	.00	.11

Note. Reliability = .89; Separation: 2.88; Strata 4.17; Fixed chi-square: 37.0 (*df* = 3; *p* < .01).

Scale functioning. The last column in Figure 3 represents how each point on the 4-point scale was utilized. The longer the scale for a scalar number, the more often this score was used. In addition to it, Figure 4 shows the probability curves for the 4-point scale, which are a visual representation of the probability of a certain score being given to a speaker based on that speaker's ability (Wright & Masters, 1982; Linacre, 1999). The horizontal axis represents the range of speakers' ability expressed in logits, and the vertical axis shows the probability of receiving a specific score. When it is most probable for a test-taker at a certain ability level to receive a specific score, the numbers form clearly defined curves. In this respect, Figure 4 depicts a good functioning scale where the probability curves peak successively and distinctly. Based on this evidence, we can see that the raters used the 4-point scale appropriately and were able to distinguish among the examinees' varying L2 speaking ability. In other words, the lower-level examinees were more likely to receive lower scores, and the higher-level examinees were awarded higher scores. In conclusion, the 4-point scale for speaking assessment was a well-functioning scale.



```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                     |
  |                                                                     |
  |                                                                     |
  |                                                                     |
  |1                                                                  44|
P | 11                                                             444  |
r |   11                                                         44     |
o |     11           22222222222                                4       |
b |       11      222           222         3333333333        44        |
a |         1   22                 22     33          333   44          |
b |          1*2                     22333               334            |
i |         22 11                     32                 4433           |
l |        2     1                  33  22              4    33         |
i |      22       11              33      22          44       33       |
t |    22           11          33          22      44           33     |
y | 222               11      33              2   44               33   |
  |2                    11  33                 2**                   333|
  |                      3**1                 44  22                    |
  |                  3333    1111         4444      2222                |
  |           3333333            111***444              222222          |
0 |***********4444444444444444444444   11111111111111111111111**********|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 4. Probability curves for the 4-point scale.

Comparison of NS and NNS Rater Groups Based on the Scores

The first research question calls for a comparison of the NS and NNS groups of raters in terms of their internal consistency, severity, scoring biases, scoring specific examinee L1, and accent familiarity. Facets analyses reports are presented to answer the research sub-questions *a* through *e*.

Raters' internal consistency and severity. First, the NS and NNS raters are compared in terms of their ability to maintain their internal consistency. Table 10 displays the descriptive statistics based on Facets rater measurement report for NS raters, and Table 11 shows the same report for the NNS raters. Regarding rater self-consistency, there were 5 misfitting NS raters and 4 NNS raters as well as 7 overfitting NS raters and 4 overfitting NNS raters. In other words, there were almost the same number of NS and NNS raters who exhibited erratic rating patterns, and there were more NS raters who showed overly-consistent rating patterns. In terms of self-consistent raters, there were 11 NS and 15 NNS whose infit statistics were within the targeted 1.2

Table 10. *Measurement Report for NS Raters (Arranged by Infit Values)*

Raters	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
NS25	0.21	.14	1.26	2.1	2.15	5.8	.65
NS20	0.15	.14	1.43	3.3	1.80	4.3	.74
NS37	-0.15	.14	1.40	3.1	1.61	3.1	.70
NS7	-0.41	.14	1.51	3.9	1.28	1.5	.71
NS41	1.23	.14	1.25	2.0	1.42	2.8	.71
NS17	0.92	.15	1.20	1.6	1.41	2.6	.81
NS3	0.26	.14	1.20	1.6	1.07	0.5	.80
NS5	0.01	.14	1.15	1.2	1.01	0.1	.79
NS40	-0.23	.15	1.01	0.1	1.15	0.8	.80
NS36	0.84	.14	1.04	0.3	0.96	-0.2	.78
NS4	0.28	.14	0.94	-0.5	1.04	0.3	.84
NS11	0.54	.14	0.97	-0.1	0.84	-1.1	.85
NS14	-0.37	.15	0.97	-0.2	0.81	-1.1	.82
NS30	0.61	.14	0.87	-1.1	0.81	-1.4	.84
NS34	0.21	.14	0.90	-0.8	0.80	-1.4	.80
NS12	-1.22	.16	1.00	0.0	0.79	-0.8	.80
NS44	-0.33	.15	0.79	-1.9	0.71	-1.7	.82
NS26	0.26	.14	0.70	-2.8	0.87	-0.8	.81
NS16	-0.93	.15	0.73	-2.4	0.62	-1.9	.84
NS21	0.49	.14	0.59	-4.2	0.62	-3.0	.86
NS33	0.13	.14	0.62	-3.8	0.56	-3.4	.87
NS9	-0.17	.14	0.64	-3.6	0.56	-3.2	.85
NS29	0.37	.14	0.55	-4.5	0.51	-4.0	.87
M	0.12	.14	0.99	-.3	1.02	-0.1	.80
SD	0.55	.00	0.27	2.4	.42	2.5	.06

Note. Reliability = .93; Separation: 3.67; Strata: 5.23; Fixed chi-square: 318.7 (*df* = 22; *p* < .01).

and 0.8 range. Overall, the mean infit square for both groups were close to each other *M* = .99, *SD* = .27 for the NS group and *M* = 1.01, *SD* = .20. Thus, there was no significant group difference in self-consistency: *t* = -0.31, *df* = 44, *p* = 0.758025 coupled with minimal Cohen's *d* = 0.09. Overall, the NS and the NNS rater groups exhibited similar internal consistency patterns.

Table 11. *Measurement Report for NNS Raters (Arranged by Infit Values)*

Raters	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
NNS10	-2.19	.17	1.07	0.5	1.57	1.3	.73
NNS13	0.14	.15	1.24	1.9	1.51	2.9	.71
NNS43	-0.29	.15	1.35	2.8	1.16	0.9	.79
NNS28	-0.14	.15	1.34	2.7	1.27	1.5	.76
NNS46	0.04	.15	1.03	0.3	1.24	1.4	.76
NNS15	0.51	.15	1.21	1.6	1.09	0.6	.81
NNS39	0.40	.14	1.18	1.5	1.04	0.2	.79
NNS35	-0.44	.15	1.13	1.1	1.04	0.2	.76
NNS1	0.79	.14	1.12	1.0	0.98	-0.1	.83
NNS45	0.20	.14	1.09	0.7	0.94	-0.3	.82
NNS2	0.95	.14	1.06	0.5	0.98	-0.1	.80
NNS8	-1.17	.15	1.05	0.4	1.01	0.1	.72
NNS38	0.72	.14	0.94	-0.4	1.03	0.2	.77
NNS6	-1.17	.15	1.06	0.5	0.89	-0.4	.77
NNS22	-0.41	.15	0.99	0.0	0.87	-0.6	.80
NNS24	0.50	.14	0.87	-1.1	0.94	-0.3	.79
NNS32	1.04	.14	0.96	-0.3	0.85	-1.1	.80
NNS42	0.05	.14	0.93	-0.5	0.84	-1.0	.82
NNS23	-0.21	.15	0.86	-1.2	0.79	-1.2	.80
NNS27	0.12	.14	0.76	-2.2	0.69	-2.2	.83
NNS31	-1.01	.15	0.76	-2.2	0.68	-1.5	.80
NNS19	-1.15	.15	0.74	-2.3	0.65	-1.6	.81
NNS18	0.04	.15	0.49	-5.4	0.44	-4.4	.90
M	-0.12	.15	1.01	0.0	0.98	-0.2	.79
SD	0.78	.01	0.20	1.8	.026	1.5	.04

Note. Reliability = .96; Separation: 5.20; Strata: 7.27; Fixed chi-square: 594.6 (*df* = 22; *p* < .01).

Second, the NS and NNS raters were compared in terms of their statistical severity measures based on the Facets measurement report. The raters with positive logits are positioned above the average severity level of 0, thus, are considered more severe, and raters with negative logits are positioned below the average severity level, therefore, they are considered more lenient. There were 15 NS and 13 NNS raters placed above the average severity level of 0, while 8 NS raters and 10 NNS raters exercised more lenient rating patterns. The mean severity logits for both groups were close to each other with *M* = .12, *SD* = .55 for the NS group and *M* = -.12, *SD* = .78 for the NNS group. Therefore, there were no significant group differences in severity: *t* = 1.15, *df* = 44, *p* = 0.256356; however, Cohen's *d* = 0.34 showed a small effect size. In other words, although the NS raters showed a tendency to provide more severe ratings, there were no statistically significant differences regarding the overall severity of the NS and NNS groups of raters.

Raters severity and consistency by examinee L1. To compare whether the NS and NNS groups rated each examinee L1 group differently, three separate Facets analyses were conducted, one for each examinee L1. The statistical information from Facets output files about rater performance is presented in Table 12 for Arabic L1, Table 13 for Chinese L1, and Table 14 for Russian L1. Rater groups' consistency and severity measures were compared across examinee L1s.

First, looking at infit statistics of rater groups for each examinee L1, it can be seen that neither NS nor NNS raters exceeded the targeted 1.2 - 0.8 values. Infit measures for the NS raters were *M* = .95, *SD* = .26 for Arabic L1, *M* = .96, *SD* = .33 for Chinese L1, and *M* = 1.04, *SD* = .62 for Russian L1. For the NNS raters, the infit measures were *M* = 1.04, *SD* = .33 for Arabic L1, *M* = 1.03, *SD* = .38 for Chinese L1, and *M* = .99, *SD* = .51. It can be concluded that, on average, both rater groups exhibited similar internal consistency across examinee L1 groups viz. both rater groups scored each examinee L1 group consistently.

Table 12. *Measurement Report for NS and NNS Groups (Arabic L1)*

Group	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
			NS				
M	.07	.25	.95	-.4	.96	-.3	.77
SD	.71	.02	.26	1.4	.33	1.4	.07
			NNS				
M	-.07	.25	1.04	.0	1.14	.2	.73
SD	.77	.01	.33	1.8	.62	1.8	.09
			Both Groups				
M	.00	.25	.99	-.2	1.05	-.1	.75
SD	.75	.02	.30	1.7	.51	1.6	.09

Table 13. *Measurement Report for NS and NNS Groups (Chinese L1)*

Group	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
			NS				
M	.06	.25	.96	-.3	.96	-.3	.75
SD	.84	.01	.33	1.7	.31	1.6	.06
			NNS				
M	-.06	.25	1.03	.0	1.01	.0	.76
SD	.96	.01	.38	1.9	.39	1.8	.09
			Both Groups				
M	.00	.25	1.00	-.2	.98	-.1	.76
SD	.90	.01	.36	1.8	.35	1.7	.08

Table 14. *Measurement Report for NS and NNS Groups (Russian L1)*

Group	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
			NS				
M	.27	.28	1.04	-.2	1.27	.1	.87
SD	.58	.01	.62	2.3	1.21	2.1	.06
			NNS				
M	-.27	.29	.94	-.4	.81	-.4	.87
SD	1.02	.01	.35	1.7	.41	1.1	.06
			Both Groups				
M	.00	.28	.99	-.3	1.04	-.2	.87
SD	.88	.01	.51	2.0	.93	1.7	.06

Second, severity of NS and NNS raters was compared per examinee L1. Based on the severity mean for the NS group (*M* = .07, *SD* = .71) and the NNS group (*M* = -.07, *SD* = .77), there were no radical differences between the NS and NNS raters scoring Arabic L1 students.

Moreover, there were no differences between these groups rating Chinese L1 students (NS: *M* = .06, *SD* = .84 and NNS: *M* = -.06, *SD* = .96). Although there were no differences, in both cases, the NNS group can be described as a more lenient one. Furthermore, more difference can be seen between the rater groups when they rated Russian L1 students. The NNS raters exhibited a more lenient scoring pattern (*M* = -.27, *SD* = 1.02) than the NS group (*M* = .27, *SD* = .58), which was statistically significant (*t* = 2.16, *df* = 44, *p* = .036308, Cohen's *d* = 0.65). Overall, based on the severity measures, the NS rater group was more severe across L1s. There were no significant differences between the NS and NNS rater groups for Arabic and Chinese L1s, but the NNS rater group was significantly more lenient when rating Russian L1 examinees who share the same L1.

Criteria difficulty. To further compare NS and NNS raters in terms of their severity, but on a more specific level, groups' criteria difficulty measures were investigated. To make such comparisons, data for NS and NNS raters were analyzed separately in Facets to look at raters' severity differences based on the rubric criteria difficulty. Table 15 and Table 16 illustrate how NS and NNS raters scored each rubric category. The rubric categories are sorted regarding their difficulty logits showing which criteria were rated more severely or leniently. The more severely scored criteria have positive difficulty logits, and the more leniently scored criteria appear with negative logits. It can be seen that both NS and NNS rater groups showed a similar pattern of criteria severity where the Overall and Delivery categories were assigned more severe ratings, while the Language Use and Topic Development were assigned more lenient ratings. The order of the rating criteria difficulty for NS and NNS speakers slightly differed as the NS raters gave harsher ratings for Delivery than Overall, and NNS raters provided more severe ratings for Overall than Delivery.

Table 15. *Measurement Report for Criteria for NS (Sorted by Difficulty)*

Criteria	Difficulty logit	Model error	Infit mean square
Delivery	.17	.06	1.07
Overall	.12	.06	.80
Language Use	-.07	.06	1.03
Topic Development	-.22	.06	1.11
M	.00	.06	1.00
SD	.16	.00	.12

Note. Reliability = .84; Separation: 2.28; Strata: 3.37; Fixed chi-square: 24.7 (*df* = 3; *p* = .00).

Table 16. *Measurement Report for Criteria for NNS (Sorted by Difficulty)*

Criteria	Difficulty logit	Model error	Infit mean square
Overall	.12	.06	.82
Delivery	.11	.06	1.06
Language Use	-.06	.06	1.02
Topic Development	-.17	.06	1.08
M	.00	.06	1.00
SD	.12	.00	.10

Note. Reliability = .74; Separation: 1.67; Strata: 2.56; Fixed chi-square: 15.1 (*df* = 3; *p* = .00).

Based on the chi-square statistic and strata for both rater groups, these four rating criteria were not interchangeable regarding their difficulty; however, in the NNS group, the difficulty level for Delivery (.12) and Overall (.11) criteria were almost the same. In addition, Delivery received slightly harsher scores from NS raters (.17) than from NNS raters (.11), Language Use was scored similarly (-.07 and -.06), and Topic Development was rated slightly more leniently by NS raters (-.22) than NNS raters (-.17). In terms of internal consistency, both rater groups had no deviations based on rating criteria infit values, which means that the criteria were rated stably. In conclusion, there were no radical differences in the way NS and NNS groups of raters scored rubric sub-categories in terms of criteria severity and consistency.

Bias analyses. To investigate if any rater groups displayed unexpected rating patterns regarding the rating criteria or examinee L1, bias analyses were added to the Facets model. A bias analysis reveals any deviations from what was expected by the model, in other words, it uncovers any unexpected tendencies of raters exercising more severe or lenient judgments. Two bias models were conducted: (a) Rater Group x Rating Criteria and (b) Rater Group x Examinee L1 Group.

The first bias analysis examined interactions between the Rater Group facet and the Rating Criteria facet. According to Facets bias calibration report, there were 8 bias terms. A bias term indicates that there was some interaction, and each bias term has a bias size to measure that interaction. A bias size that is greater than zero shows that the observed scores were higher than the expected scores, namely more lenient. Accordingly, a bias size that is smaller than zero indicates that the observed scores were lower than the expected scores, namely harsher ratings (Eckes, 2011). For each bias term the bias size is also measured in standardized *t*-scores. When the *t*-score values range between -2 and +2, there is no significant bias, whereas higher *t*-scores mean a significant bias. Based on the *t*-scores for each of the eight bias terms revealed by Facets for Rater Group x Rating Criteria interactions, none of the biases showed significance and explained 0.00% of the variance.

The second bias analysis looked at interactions between the Rater Group facet and the Examinee L1 Group facet. Facets bias calibration report returned 6 bias terms, two of which were significant and explained 0.10% of the variance. The two significant interactions are shown in Table 17. We can see that the NNS group showed some bias towards Examinee6 (bias size = .16, *t* = 2.67) and the NS group exhibited bias towards Examinee3 (bias size = -.15, *t* = -2.64). Both examinees belonged to the Russian L1 examinee group. To clarify the bias direction, the positive *t*-values mean more lenient scores than expected, while the negative *t*-values indicate more severe scores than expected. Thus, Examinee6 was scored more leniently by the NNS rater group and Examinee3 was scored more strictly by the NS rater group.

Table 17. *Measurement Report for Rating Criteria (Arranged by Difficulty Measure)*

Rater Group	Examinee	Examinee group	Observed score	Expected score	Bias size	Model S.E.	*t*	*p*
NNS	6	Russian	3403	3357	.16	.06	2.67	.0077
NS	3	Russian	3239	3284	-.15	.06	-2.64	.0085

In conclusion, even though some insignificant interactions regarding Rater Group x Rating Criteria and two significant interactions for Rater Group x Examinee L1 Group were found, neither NS nor NNS groups showed consistent positive or negative bias patterns.

Accent familiarity of NS and NNS raters. The raters in the study reported their familiarity using two methods (a) with examinee L1 identification and (b) without examinee L1 identification. When the raters reported their familiarity with L1 identification, they marked how familiar they are with Arabic, Chinese, and Russian L1 speakers on a 6-point scale. The raters reported their accent familiarity in terms of (a) general familiarity with L1 speakers, (b) communication in English with speakers of these L1s, and (c) experience teaching speakers from these L1s. When the raters reported their accent familiarity without L1 identification, they listened to 24 short excerpts (eight from each L1) and reported their general familiarity on the same 6-point scale. Two types of accent familiarity were used in order to receive more detailed information about raters' accent familiarity.

First, the results are described for accent familiarity reported by raters with examinee L1 identification. When the raters were asked to report their familiarity with English spoken by people for whom Arabic, Chinese, and Russian are L1s, the NS and the NNS groups showed differences. Table 18 describes raters' familiarity levels when they reported it with identification, which included three sub-parts, namely general familiarity, familiarity due to communication, and familiarity due to teaching. The participants provided their answers on a scale from 1 to 6 (No, Very Little, Little, Some, A Lot, Extensive), and there were 23 raters in each rater group. Therefore, the minimum score for each L1 could have been 23 (if everyone responded 1) and the maximum score could have been 138 (if everyone responded 6). Evidenced by the numbers, the NS group had a lot of familiarity with Arabic and Chinese speakers, while the NNS group did not. On the contrary, the NS group did not have as much familiarity with Russian speakers, while the NNS group had extensive familiarity.

Table 18. *Rater Familiarity with Examinees' L1s (With Identification)*

	General	Communication	Teaching	Total	*M*	*SD*
		NS				
Arabic	116	115	116	347	15.09	2.92
Chinese	121	119	119	359	15.61	2.18
Russian	90	89	72	251	10.91	3.41
		NNS				
Arabic	71	69	36	176	7.65	3.14
Chinese	72	72	41	185	8.04	3.78
Russian	130	135	129	394	17.13	1.49

Note. The scale ranged from 23 (minimum possible) to 138 (maximum possible).

Another measure of raters' accent familiarity was obtained without revealing target L1s. The participants listened to 24 12-second recordings (eight recordings from each L1 group) without any L1 identification and reported their familiarity using the same 6-point scale. There were 23 raters in each rater group. Thus, the minimum score for each L1 could have been 184 (if everyone responded 1 for all eight speakers) and the maximum score could have been 1104 (if everyone responded 6 for all eight speakers). Based on the totals and averages in Table 19, a similar trend can be seen. The NNS raters had less familiarity with Arabic and Chinese speakers and more familiarity with students who speak Russian as their L1. The NS raters were more familiar with the students who speak Arabic and Chinese L1s and had less familiarity with Russian L1 speakers.

Table 19. *Rater Familiarity with Examinees' L1s (Without Identification)*

	Arabic	*M*	*SD*	Chinese	*M*	*SD*	Russian	*M*	*SD*
NS	850	32.26	5.78	913	44.39	5.44	763	33.17	7.44
NNS	659	25.35	7.43	620	30.26	9.82	901	39.17	6.11

Note. The scale ranged from 184 (minimum possible) to 1104 (maximum possible).

To compare both accent familiarity scales (i.e., with and without L1 identification), the totals from Table 18 and Table 19 were converted into percentages. Conversion to percentages was done due to the fact that the data differed in the number of items, namely three items per L1 for were used for accent familiarity with L1 identification and eight items per L1 for were used for accent familiarity without L1 identification. The first graph in Figure 5 illustrates the percentages for accent familiarity reported with L1 identification and the second bar graph displays the percentages for accent familiarity reported without L1 identification.

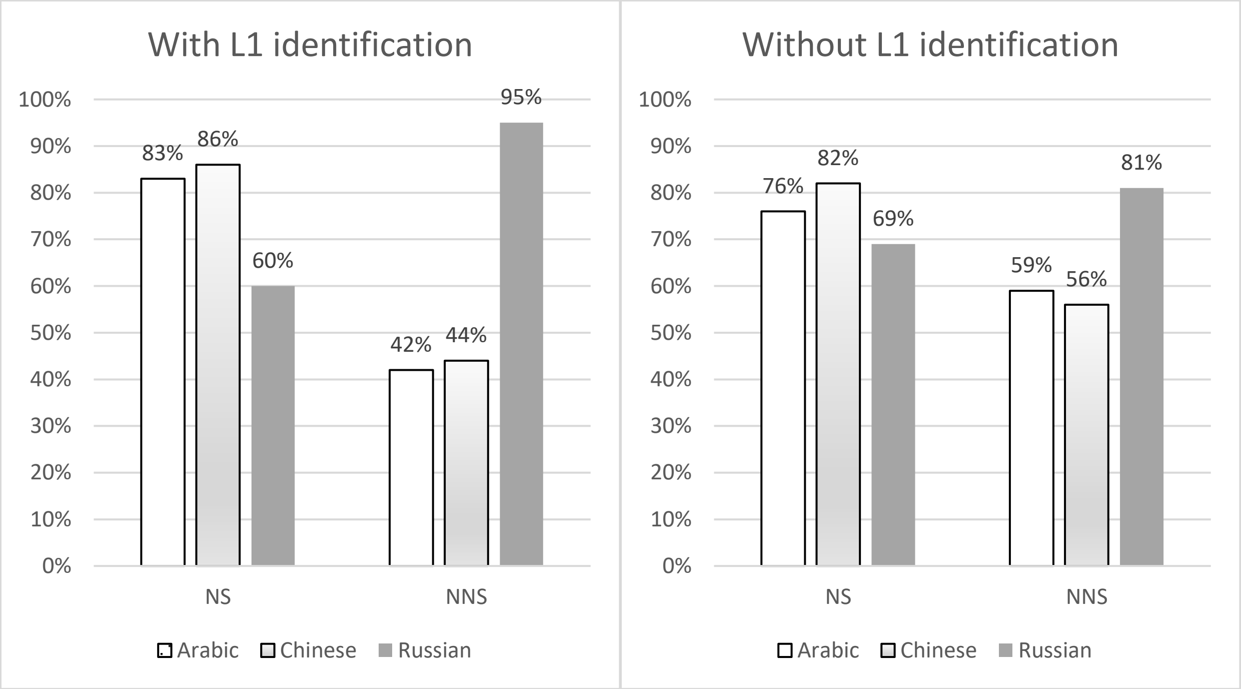


Figure 5. Rater group familiarity with examinees' L1s (in percentages).

The results obtained from the two measures, with and without identification, can be compared using the graphs in Figure 5 since they are on the same scale (percentages). Both figures demonstrate a similar picture, nevertheless, the respondents' familiarity levels shifted.

When accent familiarity was reported without L1 identification, the NS raters showed a lower familiarity with the Arabic L1 (changed from 83% to 76%) and a greater familiarity with the Russian L1 (from 60% to 69%). The NNS participants revealed a much higher familiarity with the Arabic speakers (from 42% to 59%) and Chinese (from 44% to 56%), while a lower familiarity with the Russian L1 (from 95% to 81%). Overall, it can be seen that raters' familiarity ratings were not the same when the raters reported their accent familiarity with and without examinee L1 identification. Thus, it can be hypothesized that some raters could not always identify the examinees' L1s from listening correctly.

Relationship between familiarity, severity, and consitency. This section answers sub-question e, "Is there a relationship between raters' familiarity, severity, and consistency?". Before this research question could be answered, a composite accent familiarity score (Figure 6) was calculated for each rater by summing all the scores on both accent familiarity measures for each examinee L1 group. Each rater received a composite accent familiarity score on a scale from 11 to 66 for each examinee L1.

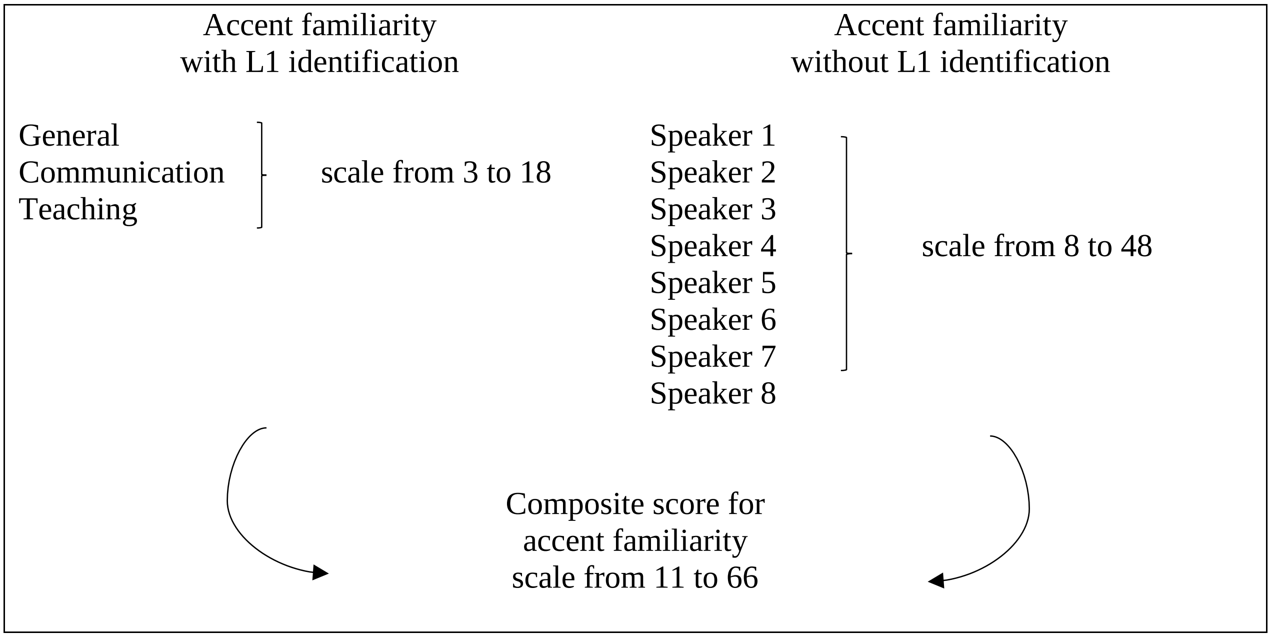


Figure 6. Accent familiarity scales that formed composite accent familiarity score.

First, raters' familiarity scores were added to Facets (Appendix M). Matching the initial 6-point scales, the six groups were raters with No, Very Little, Little, Some, A Lot, and Extensive accent familiarity. The variable maps did not show any clear dispersion of more familiar or less familiar raters. Next, to see if there is any statistical relationship between raters' familiarity and severity as well as raters' familiarity and consistency, correlation analyses were conducted. Due to deviations in normality for severity variable for Arabic L1 and consitency variables for Chinese and Russian L1s, Spearman correlations were used, while Pearson correlations were conducted for all other comparisons. For Arabic L1, there was no statistically significant relationship between familiarity and severity (r_s(44) = 0.006, p = .968) and between familiarity and consistency (r(44) = -0.216, p = .150). For Chinese L1, there was also no statistically significant relationship between familiarity and severity (r(44) = 0.033, p = .826) and between familiarity and consistency (r_s(44) = -0.279, p = .061). For Russian L1, there was no statistically significant relationship between familiarity and severity (r(44) = -0.092, p = .545) and between familiarity and consistency (r_s(44) = -0.29, p = .849). Overall, there was no statistically significant relationships neither between raters' accent familiarity and severity nor between raters' accent familiarity and consistency.

Table 20. *Measurement Report for Familiar and Unfamiliar Groups (Arabic L1)*

Group	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
			Familiar				
M	.14	.25	.93	-.5	.94	-.4	.78
SD	.53	.01	.29	1.6	.36	1.5	.07
			Unfamiliar				
M	-.14	.26	1.05	.1	1.17	.3	.72
SD	.89	.02	.30	1.6	.60	1.6	.09
			Overall				
M	.00	.25	.99	-.2	1.05	-.1	.75
SD	.75	.02	.30	1.7	.51	1.6	.09

Table 21. *Measurement Report for Familiar and Unfamiliar Groups (Chinese L1)*

Group	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
			Familiar				
M	-.03	.25	.95	-.4	.94	-.4	.76
SD	.91	.01	.32	1.7	.31	1.6	.06
			Unfamiliar				
M	.03	.25	1.06	.1	1.04	.1	.76
SD	.89	.01	.39	1.9	.39	1.8	.10
			Overall				
M	.00	.25	1.00	-.2	.98	-.1	.76
SD	.90	.01	.36	1.8	.35	1.7	.08

Table 22. *Measurement Report for Familiar and Unfamiliar Groups (Russian L1)*

Group	Severity logit	Model error	Infit mean square	ZStd	Outfit mean square	ZStd	Correlation
			Familiar				
M	-.01	.28	.98	-.3	1.04	-.2	.87
SD	.82	.01	.47	1.9	.86	1.6	.06
			Unfamiliar				
M	.02	.28	1.00	-.4	1.02	-.2	.87
SD	1.01	.02	.60	2.3	1.09	2.0	.07
			Overall				
M	.00	.28	.99	-.3	1.04	-.2	.87
SD	.88	.01	.51	2.0	.93	1.7	.06

Due to the fact that the correlation analyses did not show any relationship, it was hypothesized that not all levels of familiarity can have an effect on raters' severity and consistency. It was decided to separate the raters into familiar and unfamiliar rater groups. To achieve this, all the raters whose composite familiarity scores ranged between 11 to 44 (i.e., who reported No, Very Little, Little, or Some accent familiarity) were marked as unfamiliar raters; whereas all the raters whose composite familiarity scores ranged from 45 to 66 (i.e., who reported A Lot or Extensive familiarity) were tagged as familiar raters. Based on these cut-off points, the raters were classified into highly familiar and relatively unfamiliar groups per examinee L1 (Appendix N). Overall, there were 23 familiar and unfamiliar raters with Arabic L1, 25 familiar and 21 unfamiliar raters with Chinese L1, and 33 familiar and 13 unfamiliar raters with Russian L1. Based on this classification, the raters were added as familiar or unfamiliar raters to Facets.

Based on the severity means, there were no radical differences between the familiar and unfamiliar groups raters. To provide more details, for the Arabic L1, the familiar group was rather more severe (.14), for the Chinese L1, the familiar group was slightly more lenient (-.03), and for the Russian L1, the familiar group was slightly more lenient (-.01). Also, there were no differences in terms of rater consistency. In conclusion, the rater sub-groups, familiar and unfamiliar, did not show any observable group differences regarding severity and consistency.

Comparison between NS and NNS Rater Groups Based on the Comments

This section compares NS and NNS rater groups in terms of their comments, which raters typed after awarding their grades for each student. Analyzing raters' comments for each student can shed light on what was salient for raters' while scoring. The NS and NNS rater groups were compared in terms of (a) the overall number of comments; (b) percentages of positive, negative, and neutral comments; (c) percentages of comments attributed to Delivery, Language Use, Topic Development, and General categories; and (d) rater types based on the direction of their qualitative comments.

Number of overall comments. To answer the research question, raters' comments provided by 23 NS and 23 NNS raters were counted. As described in the Analyses section, the researcher specified the unit of analyses as a word or a phrase that describes a negative, positive, or neutral feature that can be classified as relating to the rubric criteria (i.e., Delivery, Language Use, Topic Development, and General). For example, a comment "pauses and hesitations" was coded as two negative comments about Delivery; "poor grammar and lack of vocabulary" as two negative comments about Language Use; "well-developed reasons and good examples" as two positive comments about Topic Development. Moreover, the longer elaborated sentences were also subdivided into codes, for example “That's a nice response, the speaker doesn't come across as having any trouble expressing her ideas” was coded as a positive General comment and a positive Topic Development comment (see Appendix K).

The data in Table 23 demonstrate that NS and NNS raters provided almost the same number of comments: 1674 by NSs and 1648 NNSs. To be precise, the NS group provided 26 more comments than the NNS. Thus, a conclusion can be made that there was no difference between the NS and NNS groups of raters regarding the overall number of their comments.

Comment direction: negative, positive, and neutral. Comparing NS and NNS raters based on their comments can show if either group was more focusing on the drawbacks of students’ performance or on the positive features. To compare the overall number of positive, negative, and neutral comments provided by NS and NNS raters, the coded comments were counted and their percentages were calculated. Table 23 illustrates the number and percentage of negative, positive, and neutral comments by rater group. Overall, NS and NNS rater groups produced more negative than positive comments. It can be seen that NS and NNS raters typed a similar number of negative (NS - 59% and NNS - 57%), positive (NS - 40% and NNS - 42%) and neutral (NS - 1% and NNS - 0.36%) comments. Again, a conclusion can be made that there was no difference between the NS and NNS groups of raters in terms of the percentages of negative, positive, and neutral comments.

Table 23. *Number and Percentage of Negative, Positive and Neutral Comments by Rater Group*

	# of comments by NS	% of comments by NS	# of comments by NNS	% of comments by NNS
Negative	992	59%	944	57%
Positive	665	40%	698	42%
Neutral	17	1%	6	0.36%
Total	1674	100%	1648	100%

Comment direction: Delivery, Language Use, Topic Development, and General. To compare the overall number of comments that NS and NNS raters made about students' Delivery, Language Use, Topic Development and General proficiency, the coded comments were counted in Excel and their percentages were calculated. Table 24 demonstrates the number and percentage of comments by rater group. In general, both rater groups made more comments that targeted Delivery (NS - 38% and NNS - 39%) and Topic Development (NS - 36% and NNS - 31%) and fewer comments for the Language Use (NS - 19% and NNS - 21%) and General (NS - 7% and NNS - 10%) categories. This suggests that Delivery and Topic Development could be more salient or important for both groups of raters. NS and NNS raters provided a similar number of comments for all the categories; since the numbers are very similar, it can be concluded that there was no difference between the NS and NNS groups of raters.

Table 24. *Number and Percentage of Comments by NSs and NNSs (by Criteria)*

	# of comments by NS	% of comments by NS	# of comments by NNS	% of comments by NNS
Delivery	643	38%	637	39%
Language Use	312	19%	350	21%
Topic Development	597	36%	503	31%
General	122	7%	158	10%
Total	1674	100%	1648	100%

Rater types: negative and positive. Raters' comments were used to classify raters into types based on their comments. The percentages of negative, positive, and neutral comments each rater in both rater groups wrote were compared to see if the raters can be classified into negatively-, positively-, or neutrally-oriented types. Table 25 shows number and percentage of comments for NS raters and Table 26 for NNS raters.

Table 25. *Number and Percentage of Negative, Positive and Neutral Comments Provided by NS Raters*

Rater	Negative	% Negative	Positive	% Positive	Neutral	% Neutral	Total
NS3	51	65%	27	35%	0	0%	78
NS4	51	57%	38	43%	0	0%	89
NS5	29	58%	20	40%	1	2%	50
NS7	38	50%	38	50%	0	0%	76
NS9	53	53%	35	35%	12	12%	100
NS11	35	73%	13	27%	0	0%	48
NS12	24	44%	30	56%	0	0%	54
NS14	58	60%	39	40%	0	0%	97
NS16	37	88%	5	12%	0	0%	42
NS17	19	49%	20	51%	0	0%	39
NS20	12	31%	27	69%	0	0%	39
NS21	43	90%	5	10%	0	0%	48
NS25	51	62%	28	34%	3	4%	82
NS26	30	73%	11	27%	0	0%	41
NS29	33	69%	15	31%	0	0%	48
NS30	51	61%	33	39%	0	0%	84
NS33	59	60%	39	40%	0	0%	98
NS34	52	62%	31	37%	1	1%	84
NS36	69	57%	52	43%	0	0%	121
NS37	29	56%	23	44%	0	0%	52
NS40	23	49%	24	51%	0	0%	47
NS41	86	59%	59	41%	0	0%	145
NS44	59	53%	53	47%	0	0%	112
Total	992	59%	665	40%	17	1%	1674

According to the information in both tables, there were no raters whose overall majority of comments was neutral. Three raters provided more positive than negative comments: two raters (NS12 and NS20) in the NS group and one rater (NNS6) in the NNS group. That being said, the rest of the raters can roughly be allocated to the negatively-oriented group. However, it is also possible to single out one more group among the three; if a rater gives an approximately similar number of positive and negative comments, they can be classified as a balanced rater. Thus, the raters who provided around 50% of both positive and negative comments (e.g., 45% and 55%, 53%, and 47%) were considered balanced. There were three balanced raters in the NS group (NS7, NS40, NS44). The NNS group had eight balanced raters (NNS2, NNS10, NNS13, NNS18, NNS24, NNS43, NNS45, and NNS 46). To conclude, first, there are no radical differences between rater groups regarding the number of positive and negative comments produced if the raters are subdivided only to negative or positive. Second, there are some differences between the NS and NNS groups in terms of the number of balanced raters who provided similar numbers of positive and negative comments. The NNS group had more balanced raters; therefore, it can be suggested that the NS group, in general, had slightly more negatively-oriented raters, which, potentially, could have affected their severity measures.

Table 26. *Number and Percentage of Negative, Positive and Neutral Comments Provided by NNS Raters*

Rater	Negative	% Negative	Positive	% Positive	Neutral	% Neutral	Total
NNS1	81	77%	24	23%	0	0%	105
NNS2	58	55%	47	45%	0	0%	105
NNS6	60	38%	97	62%	0	0%	157
NNS8	44	67%	22	33%	0	0%	66
NNS10	52	53%	47	47%	0	0%	99
NNS13	17	45%	18	47%	3	8%	38
NNS15	75	56%	59	44%	0	0%	134
NNS18	49	51%	48	49%	0	0%	97
NNS19	42	58%	31	42%	0	0%	73
NNS22	28	67%	14	33%	0	0%	42
NNS23	34	62%	20	36%	1	2%	55
NNS24	23	52%	21	48%	0	0%	44
NNS27	27	63%	16	37%	0	0%	43
NNS28	40	75%	13	25%	0	0%	53
NNS31	39	64%	22	36%	0	0%	61
NNS32	14	70%	6	30%	0	0%	20
NNS35	18	56%	14	44%	0	0%	32
NNS38	30	64%	17	36%	0	0%	47
NNS39	35	64%	18	33%	2	4%	55
NNS42	41	75%	14	25%	0	0%	55
NNS43	50	50%	51	50%	0	0%	101
NNS45	42	52%	39	48%	0	0%	81
NNS46	45	53%	40	47%	0	0%	85
Total	944	57%	698	42%	6	0.36%	1648

Rater types by rubric criteria. The second classification of raters was made based on the comments the raters attributed to the following rubric criteria: Delivery, Language Use, Topic Development, and General (see pie-charts in Appendix O). Table 27 presents percentages for NS and Table 28 for NNS.

Table 27. *Number and Percentage of Comments Provided by NS Raters (by Criteria)*

Rater	D	% D	LU	% LU	TD	% TD	G	%G	Total
NS3	28	36%	15	19%	25	32%	10	13%	78
NS4	29	33%	8	9%	45	51%	7	8%	89
NS5	9	18%	8	16%	24	48%	9	18%	50
NS7	24	32%	18	24%	29	38%	5	7%	76
NS9	33	33%	32	32%	34	34%	1	1%	100
NS11	8	17%	19	40%	19	40%	2	4%	48
NS12	16	30%	12	22%	26	48%	0	0%	54
NS14	42	43%	19	20%	29	30%	7	7%	97
NS16	32	76%	8	19%	1	2%	1	2%	42
NS17	14	36%	2	5%	15	38%	8	21%	39
NS20	18	46%	0	0%	17	44%	4	10%	39
NS21	21	44%	7	15%	16	33%	4	8%	48
NS25	35	43%	27	33%	10	12%	10	12%	82
NS26	17	41%	1	2%	18	44%	5	12%	41
NS29	30	63%	5	10%	13	27%	0	0%	48
NS30	42	50%	25	30%	13	15%	4	5%	84
NS33	43	44%	10	10%	40	41%	5	5%	98
NS34	30	36%	15	18%	30	36%	9	11%	84
NS36	49	40%	30	25%	41	34%	1	1%	121
NS37	23	44%	2	4%	18	35%	9	17%	52
NS40	17	36%	7	15%	22	47%	1	2%	47
NS41	58	40%	33	23%	53	37%	1	1%	145
NS44	25	22%	9	8%	59	53%	19	17%	112
Total	643	38%	312	19%	597	36%	122	7%	1674

Note. D – Delivery, LU – Language Use, TD – Topic Development, G – General.

First, the raters were classified into groups when the majority of their comments (around 50%) were attributed to one particular category and there was no other category with more than 35%. There were no Language Use-oriented or General-oriented rater types among the NS raters and only two (NNS43 and NNS13) among the NNS raters with 58% of comments devoted to Language Use and General proficiency accordingly. In terms of Delivery, there were three NS raters who focused mostly on this category with NS16 (76%), NS29 (63%), and NS30 (50%). In the NNS group, there were 10 raters who concentrated on Delivery: NNS1 (49%), NNS15

(41%), NNS22 (52%), NNS23 (64%), NNS28 (60%), NNS31 (61%), NNS32 (60%), NNS38 (62%), NNS39 (58%), and NNS42 (64%). For Topic Development, the same number of NS and NNS gave preference to this category, namely four NS raters: NS4 (51%), NS5 (48%), NS12 (48%), NS44 (53%) and four NNS raters: NNS2 (49%), NNS6 (62%), NNS35 (47%), and NNS46 (53%).

Table 28. *Number and Percentage of Comments Provided by NNS Raters (by Criteria)*

Rater	D	% D	LU	% LU	TD	% TD	G	%G	Total
NNS1	51	49%	13	12%	30	29%	11	10%	105
NNS2	24	23%	26	25%	51	49%	4	4%	105
NNS6	39	25%	16	10%	97	62%	5	3%	157
NNS8	29	44%	4	6%	32	48%	1	2%	66
NNS10	48	48%	41	41%	9	9%	1	1%	99
NNS13	6	16%	1	3%	9	24%	22	58%	38
NNS15	55	41%	33	25%	35	26%	11	8%	134
NNS18	29	30%	19	20%	31	32%	18	19%	97
NNS19	16	22%	20	27%	27	37%	10	14%	73
NNS22	22	52%	6	14%	4	10%	10	24%	42
NNS23	35	64%	1	2%	17	31%	2	4%	55
NNS24	14	32%	7	16%	16	36%	7	16%	44
NNS27	16	37%	9	21%	14	33%	4	9%	43
NNS28	32	60%	9	17%	9	17%	3	6%	53
NNS31	37	61%	15	25%	7	11%	2	3%	61
NNS32	12	60%	3	15%	5	25%	0	0%	20
NNS35	7	22%	6	19%	15	47%	4	13%	32
NNS38	29	62%	12	26%	6	13%	0	0%	47
NNS39	32	58%	13	24%	8	15%	2	4%	55
NNS42	35	64%	6	11%	9	16%	5	9%	55
NNS43	29	29%	59	58%	1	1%	12	12%	101
NNS45	17	21%	15	19%	26	32%	23	28%	81
NNS46	23	27%	16	19%	45	53%	1	1%	85
Total	637	39%	350	21%	503	31%	158	10%	1648

Note. D – Delivery, LU – Language Use, TD – Topic Development, G – General.

Second, the raters who did not favor one specific category were categorized into two sub-categories: a) raters who had two salient categories and b) raters who were more or less balanced. The majority of the raters who prioritized two categories were NS with 13 of them: NS3, NS14, NS17, NS20, NS21, NS26, NS33, NS34, NS36, NS37, NS40, NS41 who provided more comments for Delivery and Topic Development; NS11 who gave more comments for Language Use and Topic Development, and NS25 who mentioned Delivery and Language Use more often. In the NNS group, there were three raters: NNS8 and NNS27 who commented more about Delivery and Topic Development and NNS10 who commented more about Delivery and Language Use. The last type comprised of more or less balanced raters with two NS raters (NS7 and NS9) and four NNS raters (NNS18, NNS19, NNS24, NNS45). These raters either mentioned only three major categories of Delivery, Language Use, and Topic Development almost the same number of times (e.g., NS9) or talked about all four criteria more or less equally (e.g., NNS18), so that they could not be placed into other rater types. In conclusion, based on the descriptive comparisons, there were some observed differences between NS and NNS groups of raters. Overall, more NNS raters tended to pay more attention to Delivery, whereas more NS raters treated both Delivery and Topic Development as more important criteria.

Summary

The first research question compared NS and NNS raters in several ways: (a) statistical comparisons of raters' consistency and severity based on the analytic scores they awarded to the students, (b) statistical comparisons of raters in terms of grading specific examinee L1 groups, (c) the effect of raters' accent familiarity with L1 student groups on raters' scoring patterns, and (d) number and direction of raters' comments.

The NS and NNS groups of raters exhibited no significant differences in their internal consistency (Cohen's *d* = 0.09) and no significant differences regarding their severity measures (Cohen's *d* = 0.34). Also, both groups utilized the rating criteria in a similar way showing comparable criteria difficulty and consistency of application. In addition, bias analyses that aimed at identifying interactions for the Rater Group facet and the Rating Criteria facet and between Rater Group facet and Examinee L1 Group facet did not show any specific patterns of bias.

Accent familiarity of NS and NNS raters was comparable but not absolutely the same when it was collected with and without L1 identification. When the NS and NNS raters' grading patterns were compared by examinee L1, NNS raters were significantly more lenient (Cohen's *d* = 0.65). Furthermore, when all the raters were subdivided into familiar and unfamiliar types, no differences were found in their severity and consistency when the comparisons were made by student L1.

When comparisons between NS and NNS were made based on the comments, there were no differences between NS and NNS raters in terms of the overall number of comments, average percentages of negative and positive comments as well as percentages attributed to the rating criteria. When the raters were classified into types based on their positive and negative comments, there were more negatively-oriented raters among NS and more balanced NNS raters. When the raters were classified into groups based on their focus on rating criteria, the NNS group tended to pay more attention to Delivery, and the NS raters concentrated on Delivery and Topic Development; both groups showed much less focus on Language Use criteria.

Qualitative Results

Qualitative analyses of think-aloud protocols and interviews facilitated answering the second research question about the scoring strategies that NS and NNS raters used while grading L2 speaking performance. The raters' patterns of decision-making were described in terms of their listening strategies, grading strategies, non-rubric criteria references, perceived severity, and perceived category importance. The listening strategies are defined as what the raters did during the time a student's recording was playing, and the grading strategies refer to raters' decision-making processes after listening to a student's recording. Non-rubric criteria references include raters' references to additional criteria that the rubric does not describe while providing their score justifications during the think-aloud protocol or reflecting on their grading processes during the interview. Perceived rater severity refers to raters' opinions about themselves as more lenient or severe raters, and criteria importance means what weight the raters ascribed to a scoring rubric criterion.

NS and NNS raters

The results for the second research question describe what raters did while listening and grading, what non-rubric criteria they employed, and their thoughts about perceived severity and category importance. The raters did not form any particular groups based on their NS or NNS status as comparable number of NS and NNS raters followed various strategies. However, it is important to acknowledge the fact that the nature of this qualitative investigation was exploratory and the limited numbers of NS (*n* = 7) and NNS (*n* = 9) raters did not allow the researcher to make clear distinctions and generalize what patterns were more or less common for NS or NNS raters.

Listening Strategies

The raters employed various listening strategies, for example, some raters did or did not do the following: took notes, looked at the rubric while listening, drafted scores, looked at the length of the recording, or re-listened. Based on their strategies, the raters were classified into note-takers, hard listeners, multitaskers, and re-listeners.

Note-takers and hard listeners. Out of 16 raters who participated in the think-aloud protocols and interviews, 9 raters took notes (NS7, NS30, NS34, NS37, NNS2, NNS6, NNS24, NNS28, and NNS46) and 7 raters did not take notes (NS14, NS17, NS40, NNS10, NNS13, NNS35, and NNS45). All the raters had justifications for why they took or did not take notes. The raters who took notes felt that it helped them to be an involved listener, not to rely on their working memory, and to decide on the grades. Excerpts 1 and 2 illustrate raters' reasons for taking notes while listening to examinees' responses.

> *Excerpt 1, NNS6: While I was listening, I was taking notes and I tried to give myself some structure of the response and still I looked over my notes and I said that same ideas were repeated yes and no clear structure*

> *Excerpt 2, NS2: I always take notes. I never rely on my memory… I feel that I need to take lots of notes because it helps me to assess the response.*

The raters who took notes, also exhibited individual patterns based on what they made their notes about, the number of notes they took, and the regularity of their notes. Most of the note-takers reported taking notes about students' pronunciation, reasons, and grammar. Some raters took extensive notes writing keywords and phrases (Excerpts 3, 4, and 5), and others took shorter notes using symbols (Excerpt 6). Additionally, some raters stopped making notes if it was evident for them that a student was producing little content and was not going to score higher than a 1 or a 2 (Excerpt 6).

> *Excerpt 3, NS7: Just writing, taking notes. I just wrote keywords. Transitions, fluid, easy to listen to, sustained… Taking notes. I noted that it was good right off the bat. Pacing was slow. He does use conjunctions and some transitions. Very basic grammar and repetitive. Those are the words I wrote.*

> *Excerpt 4, NNS24: I wrote "long pause, unclear, and overall development is limited."…*

> *Excerpt 5, NS34: I was taking notes. I have 5 lines of notes. I noted that I couldn't understand the second phrase. I noted serial lists. I noted 25 seconds in I didn't understand the sentence.*

> *Excerpt 6, NNS28: I wrote words. I wrote real words that included mistakes, I guess. There were check marks, there were dots, something like that…While listening I made some notes, so I tried to count out arguments, then I tried to write down main words or phrases connected with these arguments. That's it. ... I was staring at the screen and then I tried to make notes maybe to count his points, but I failed, there was only one, so I tried to make notes but I didn't.*

Two raters, NNS2 and NNS6, explained their note-taking strategy in detail. They were sure that due to the fact that this is a life-changing exam, they had to be as thorough as possible. NNS2 (Excerpt 7) tried to make an outline of each students' answer, whereas NNS6 (Excerpt 8) had separate parts of the notebook devoted to notes about vocabulary, reasons, and grammar errors.

> *Excerpt 7, NNS2: I don't have a specific pattern of taking notes, but I try to. First of all, I'm expecting some sort of introduction, and I'm marking whether I get this introduction, whether I hear it or not. Then I'm looking for the opinion, yeah, of course, if it's this choice question, then I see whether the student actually makes his or her choice. I always try to write meaningful words that form the skeleton of their answer. If I notice some really bad grammar mistakes I always write them down to be able to back-up my explanation. So, yeah, for some reason I write good vocabulary and bad grammar, and I write or try to follow the structure. If I hear some good linking words and phrases I always try to mark them down.*

> *Excerpt 8, NNS6: Each recording is separated and tried to put down the general idea of the speaking by noting down some core collocations let's say and brief sentences or phrases or collocations so from these notes I get the idea of the volume of vocabulary of the speaker and what he or she mentioned, right, the reasons. And the right side of the paper is divided, on the right side I put down some grammar notes or vocabulary notes that show me the mistakes that the person makes but I remember that you don't need to avoid or not avoid maybe but how to put it, it doesn't have to be excellent right to be 4 it's appropriate to make some minor mistakes so when I make notes in this column I really take a look at these notes and I think are they really major or are they really so umm, so what, sorry I got sidetracked. Are they acceptable, yeah, do they really influence the overall idea of the speaker.*

On the contrary, the raters who did not take notes reported that any actions while listening to students' recordings are distracting and can result in less focus and worse comprehension, which is illustrated in Excerpt 9 and 10. Many raters noted that it helped them to close their eyes while listening to students' responses because it allowed being particularly focused, which is also mentioned by NS14 in Excerpt 9.

> *Excerpt 9, NS14: I was closing my eyes and listening. Yeah, I feel like if I'm looking at something that I'll get distracted and I'll miss what she says…I don't take any notes, I try to be as focused as possible.*

> *Excerpt 10, NS40: I want to try to give the same amount of attention, time, and consideration as I possibly can. I am not really a note-taker; I don't write things whenever I'm rating at least. I don't write things down, I just try to give full attention. I don't look at anything else but the laptop screen and I listen in… Just stare blankly at the laptop with my ear close to it so I could really hear what she was trying to say. Taking mental notes, I wasn't writing anything down.*

Not all the raters who did not take notes had such a strong opinion that taking notes can adversely affect their rating decisions, for example, NNS13 in Excerpt 11 mentioned that they never take notes because it is not their style.

> *Excerpt 11, NNS13: Though I have a pen and a paper here on my table, I never used it today. That's because it's easier for me to just to grade listening material just listening and concentrating on audio material than take notes. And as far as I know some teachers they while checking listening activities and even speaking, yeah, so they take notes and I never do it. For me, it's better to just to have general idea and then I can go into details.*

In addition, it was interesting to see one rater (NS30 in Excerpt 12), who can potentially be described from both perspectives – they valued the notes, but at the same time realized that writing could distract and lead to being less attentive. Thus, this rater always took very short notes trying to listen more and write less.

> *Excerpt 12, NS30: I try to focus, listen really hard, focus really hard, sometimes I take notes on content. When I work hard to take quick notes, so that I'm not writing more and listening less. If I hear something particularly strong that's related to the rubric, if I hear really good pronunciation or a lot of pauses, something like that, I'll write this down really quickly in shorthand and then go back to taking my content notes. Or if I hear clearly in the beginning that delivery is a 3, I'll write D3, so I have a shorthand. So I think my biggest strategy is I just try to not write a lot, and just listen very carefully.*

Overall, it is debatable which one is a better listening strategy – to take note or not to take notes. While taking notes can certainly be seen as a distracting factor, it can also be viewed as a benefit or a strategy of people who do not like to rely on their working memory. In addition, taking notes can also be seen as a teacher strategy since teachers have to take notes about students' performance to provide more detailed feedback on students' spoken performance. An important concern about the note-taking approach to grading is rater fatigue. This opinion was verbalized by NNS6 (Excerpt 13) who always took extensive notes. NNS6 questioned their strategy of combining two actions, writing and listening as they thought it could be too cognitively demanding and tiring resulting in decreased quality of rating. NNS6 thought about changing the note-taking strategy and write down less, but they did not change it.

> *Excerpt 13, NNS6: So for now I didn't lose my focus, my concentration but already I start to feel that maybe I should change my strategy, we'll see in the following tasks, but maybe I should put down less.*

However, a combination of listening and writing is common in real life since people take notes from listening often, for example, during academic lectures or professional development sessions. The fact that taking notes from listening is a well-practiced type of efficient multitasking, it might not contribute to excessive rater fatigue.

Multitaskers. Some raters not only were taking notes while listening to the recordings, but also looking at the rubric, drafting preliminary scores on the screen, checking the length of the recording, drinking water, or snacking. The following five raters engaged in multitasking more often: NNS45, NS7, NS17, NS37, and NS10. Excerpts 14 to 17 illustrate that the raters were scanning the rubric while listening to students' responses to locate keywords in the rubric that describe examinees' level of response. It can be hypothesized that these raters did not internalize the rubric and needed some guidance from the keywords to decide where on the scale they can position a student.

> *Excerpt 14, NS7: I was trying to pay attention to what she was doing but I was also mindful of the rubric and looking for keywords there so I was double tasking.*

> *Excerpt 15, NS17: I was looking at the rubric a lot. I scrolled to where in the rubric I think she's at. I was hovering in that 3 to 2 range.*

> *Excerpt 16, NS37: I took a few notes. Just two words of notes. I glanced at the rubric for about 10 seconds during the recording. I was looking for keywords on pronunciation.*

> *Excerpt 17, NNS24: I was looking at the rubric and trying to choose an appropriate for her answer phrases, that would suit her answer.*

Three more raters (NS34, NNS35, NNS24) were not classified as multitaskers because they did not multitask throughout the whole think-aloud protocol. They glanced at the rubric several times in the beginning but then never did that again. In particular, NNS35 (Excerpt 18) specified that they think that looking at the rubric was distracting and that was the reason for the change in their listening strategy.

> *Excerpt 18, NNS35: Yes, at first, I was trying to rate them as the recording goes. I was looking at the rubric and looking at descriptions while I was listening to the recording. But then I realized it distracted me from the speech of the person, so I stopped looking at the rubric and started listening to the person at first and referring to the rubric when the recording was over.*

While for some raters looking at the rubric was a preferred listening strategy, others noted that they never look at the rubric whenever they grade. The following eight raters never looked at the rubric while listening either because they were busy taking notes (NNS2, NNS6, NNS46) or did that on purpose (NNS13, NS30, NS40, NS14, NNS28). Some raters who did not look at the rubric on purpose explained that reading the rubric distracts from listening (Excerpt 19), while other raters stated that they did not do it because they were confident in their knowledge of the rubric (Excerpt 20).

> *Excerpt 19, NS30: I don't think people should read the rubric but focus intensely on listening.*

> *Excerpt 20, NS14: I didn't look at the rubric because I feel fairly confident with my knowledge of the rubric.*

Overall, multitasking in the form of combining lsitening and reading (skimming the rubric while listening) is a controversial issue that can be seen mostly from a negative perspective. First, looking at the rubric can be considered a distracting factor that overloads raters' working memory while listening to a students' response and acts as a catalyst for rater fatigue and rater inconsistency. Second, scanning the rubric for keywords to describe a students' performance can be a strategy employed by the raters who did not internalize the rubric and need more practice. Third, skimming the rubric while listening can also indicate that the raters are trying to save some time and have the final grades confirmed with the rubric right by the time a test-taker is finished talking. Such strategy can save some time spent on decision-making so that they can move on to grading the next recording right away. If this is the case, such behavior should be modified since raters should pay their full attention to students' responses.

Re-listeners. Another strategy that the raters employed or refrained from was re-listening to examinees' recordings. Some raters believed that re-listening is not fair because it means more time allotted to some students and less to other students. Other raters stated that it is unfair not to re-listen to students' samples if raters need more time to adjust to students' pronunciation. The following five raters never listened twice: NS7, NS37, NS13, NNS24, NNS45, but they did not have any particular explanation to why they did that. Three more raters listened only once because they had a strong stance on re-listening being unfair and having a negative impact on their consistency – NNS46, NS30, and NS40 (Excerpt 21).

> *Excerpt 21, NS40: I listen to once, to best of my abilities, you know, I can't listen to certain speakers two or three times because that would, perhaps, influence my opinion more, so I just try to be as consistent as possible, in short.*

One rater, NS14, tried to listen only once, but re-listened if the quality of the audio or pronunciation needed more attention. They explained that listening more than once makes them liable to overthink and give untrustworthy grades (Excerpt 22).

> *Excerpt 22, NS14: I listen once, usually with my eyes closed… And for some of them, it's fine, that one listening is good. I give the rating, I feel good about it. If some of the parts are hard to understand, if the audio, the pronunciation was difficult, I listen to it the second time because I like to listen for content and LU. That's sometimes is more difficult to catch during the first listening. Generally speaking, I'll listen to it once and I take my first initial reactions because I don't want to overanalyze and get too confused about*

> *what I want to do… And if I start trying to think about it too hard, if I take too long or listen to it too many times I start to neat-pick on the background noise here, or there was this many pauses or that many pauses. I think that then I get too much into the specifics and I end up giving a grade that I don't actually trust.*

Seven other raters (NS17, NS34, NNS2, NNS6, NNS10, NNS28, NNS35) strongly believed that they need a second listen to overcome students' difficulties with pronunciation (Excerpts 23 and 24). In addition, in Excerpt 25, NNS6 was also worried about the responsibility that raters have. NNS6 believed that it is necessary to do everything possible to adjust and understand students.

> *Excerpt 23, NNS6: I started to make notes and I figured out that I can't really properly formulate what I hear and I really started my work from second half of the recording when I adjusted myself to this accent so I had to listen till the end and get back to the beginning and I listened to the first part again and then I heard some really good expressions that I didn't understand from the first time.*

> *Excerpt 24, NS34: Listening twice helped me to adapt to the accent a little bit.*

> *Excerpt 25, NNS6: It was hard because I really struggle and I feel more, I feel that I have to be more attentive when I listen to the accent that I don't really understand and it gives me the additional sense of responsibility. That's why when I don't understand the accent I tend to pause the audio and listen again…*

Re-listening to students' responses is a contentious issue in the present study. Just as the raters mentioned, it can affect raters' grades in various ways: (a) raters can assign different scores because they try or do not try to tune to speakers' pronunciation, (b) raters can assign more lenient scores due to better comprehension because of re-listening, and (c) raters can render inconsistent scores due to overthinking. It is important to note that while rating writing, raters always have a possibility to re-read a part of an essay to make sure they are following a writer's logic or understood a reason correctly, therefore, it probably should not be forbidden in speaking. On the other hand, others would argue that when people speak in real life, they do not have an opportunity to listen again. However, in real-life situations, interlocutors always have an opportunity to ask a clarification question, ask to repeat, or negotiate the meaning in some other ways, for example, using gestures. To this end, NS34 (Excerpt 26) noted that the way semi-direct exams are structured is not natural.

> Excerpt 26, *NS34: I'm a strategic communicator. If I do not understand someone, I will try to understand someone, I will try to negotiate that communication… I am a strategic communicator. If I don't understand something, I'm willing to work it out with them, which you can't do in this context.*

Overall, we can see that re-listening to examinees' speech samples is a debatable issue. Drawing a straightforward conclusion about whether raters should or should not re-listen is not possible at this point.

In sum, the information from the think-aloud protocols and interviews not only allowed classifying raters into note-takers, hard listeners, multitaskers, and re-listeners but also helped to describe raters' various approaches to taking notes, beliefs about multitasking while rating examinees' L2 speaking performance, and rationales behind re-listening to student samples.

Grading Strategies

The raters in this study employed two grading strategies – top-down and bottom-up. The top-down raters relied on their overall impression about an examinee's answer before making a decision, and the bottom-up raters relied on analytic rubric criteria first. Regardless of the approach, the second step that the raters most frequently followed was to briefly or more thoroughly consult the rubric in order to find some keywords that can help them become sure about their decisions. Furthermore, all the raters showed some recurring patterns regarding what factors made them change their initial score decisions. The two most common reasons were the differences in students' performance throughout the recording and a controversy between the raters' general impression and the final score that the raters decided on. The raters did not show any specific patterns depending on the NS or NNS rater group.

Before the raters' grading strategies are described, it is important to explain how the TOEFL iBT independent speaking rubric works. The rubric has four bands (i.e., 0, 1, 2, 3, 4) and four criteria (i.e., General Description, Delivery, Language Use, and Topic Development). To give a General Description score or, in other words, an overall holistic score, raters should consider all other rubric criteria because, after a brief holistic description, the rubric states, "A response at this level is characterized by at *least two* of the following". This means that two rubric categories should be awarded, for example, a 2 to arrive at a holistic score of 2. The exception is band 4 where all criteria should be given a 4 since the rubric says, "A response at this level is characterized by *all* of the following". Ultimately, the task of a rater is to assign a rank order number to each response on a holistic scale from 0 to 4; however, raters can arrive at the same holistic score guided by different criteria score judgments. For example, to get a score of 3 on a TOEFL independent speaking task, two out of three judgments must fall into this band, but also a person cannot be given a 4 if even one partial score is at a 3 level. This means that the combinations for getting a score of 3 may vary according to these patterns: 3-3-3, 2-3-3, 3-2-3, 3-3-2, 4-3-3, 3-4-3, 3-3-4, 4-4-3, 3-4-4, 4-3-4. A smaller amount of variance in the analytic scores exists for a score of 2: 2-2-2, 3-2-2, 2-3-2, 2-2-3, 1-2-2, 2-1-2, 2-2-1. The number of combinations for a score of 1 is even smaller: 1-1-1, 1-1-2, 1-2-1, 2-1-1. There can be only one combination for a score of 4 and 0. A score of 4 can be given only if all sub-ratings are given a 4, and 0 means that there is no attempt or the response is not on the topic of the prompt.

Based on the organization of the TOEFL iBT independent speaking rubric, we can draw an inference that raters should use this rubric in both holistic and analytic ways in order to arrive at a final holistic decision. Based on the rubric statements discussed above, a holistic score has to be supported by partial criteria scores. However, the rubric does not provide any guidelines about the order of the decisions. Thus, the raters' grading strategy can either be to consider the analytic criteria first and subsequently assign a holistic score or to assign a holistic score and subsequently confirm it by looking at the analytic criteria.

Top-down and bottom-up. The raters in this study followed two grading approaches – top-down and bottom-up. The top-down raters used their overall impression about an examinee's answer to make a decision, wherease the bottom-up raters judged each analytic rubric criterion first. Out of 16, 10 raters (NS7, NS14, NS17, NS30, NS40, NNS2, NNS6, NNS13, NNS24, NNS35, NNS46) used the top-down approached. While listening, the top-down raters formed a preliminary holistic impression about an overall grade that they were going to give to a test-taker and then, after the recording was finished, skimmed the rubric, thought more, and assigned partial criteria grades. Excerpts 27 and 28 demonstrate the top-down approach.

> *Excerpt 27, NS40: Whenever I hear the recording I automatically try to determine what I would give as the O rating and I deconstruct it from there band by band… After the recording finished I knew that I was going to give a 1 so I started on the website, scoring each band individually.*

> *Excerpt 28, NNS35: So I got an overall idea of the response in the end, so I started looking at the rubric and using the descriptions after I heard the whole response, and then I just went one by one and looked at each of the sections of the rubric.*

Usually, the top-down raters checked their first impressions with the rubric. Excerpts 29 and 30 illustrate that the raters briefly skimmed the rubric to find some statements that they can agree or disagree with in order to finalize their decision.

> *Excerpt 29, NS7: I went to consult the rubric. I found statements on the rubric that justified my feelings.*

> *Excerpt 30, NNS13: Well, that means that actually I was ready with a grade after listening period, you see. So, after I have just listened to the student I already notice some failures in speech and some maybe some advantages in using some grammar or speech patterns that I liked and so maybe that is this way I just then while going through the*

> *rubrics again I could just say OK yeah that was here and I can agree or I can disagree with this rubric or that one so that I could just compare looking through the rubrics to my first impression because in my memory there are some details that could be corresponded to the features in rubrics. That's mostly taken from my memory.*

Additionally, the top-down raters reported that they formed their opinion about a holistic grade or a range of grades (e.g., between 2 and 3 or between 3 and 4) approximately half-way through a recording. This strategy was mostly used when the raters were sure about their grades and most commonly happened when a rater positioned a recording at extremes, namely 1 or 4. Moreover, in such cases, the raters tended not to scan the rubric to confirm their decision. Excerpt 31 demonstrates a rater who decided that a students' response is worth the top grade in the middle of listening to it. This excerpt also shows that the rater did not read the rubric to check the accuracy of their decision.

> *Excerpt 31, NS40: I applied a similar technique. First, I tried to give my full attention for as long as I can and once I got an idea of how he was gonna organize this but it was like, ok he is about to make his first point, let's see what information he says. I was trying to listen to see if I could get 2 or 3 distinct reasons to follow his logic. I was trying to follow the speaker's logic to make sure that he was going to fully deliver a good argument which he did. I was comfortable based on his pronunciation, intonation and his vocabulary and his grammatical structure. O, this was easy for me to determine this was a 4. I did not refer to the rubric immediately after listening to this file…I would say very easy. I would say about halfway into it, barring any major setbacks, I was comfortable giving him a 4.*

Three raters (NS37, NNS10, NNS28) used the bottom-up approach (Excerpt 32). They thought about partial criteria scores while listening and then, after the recording was finished, confirmed the scores, and used them to make a decision about the overall grade that they were going to give. Furthermore, there were some similarities between the bottom-up and the top-down raters – occasionally, the bottom-up raters also made a decision on the sub-categories while listening, and they referred to the rubric after listening to confirm their scores (Excerpt 32).

> *Excerpt 32, NNS28: Sometimes I remember that I put - right away, the grade, maybe I put notes and the grade, next to it, and if I remember it correctly I didn't put the overall grade, but I put the grade by category...and then I made the calculation, in the end. But*

> *after my notes, I came back to the rubric, and in places where I was doubtful, I tried to look through and to find something to be based on, I guess. And after this double check I decided on the grade, I decided between two or three or three or four, and so on.*

Two raters, NS34 and NNS45, noted that it is easier to use both the top-down and bottom-up approaches. These raters used the top-down approach when it was easy to decide on a grade and the bottom-up approach when a recording posed some challenges. Excerpt 33 describes a rater who used both approaches.

> *Excerpt 33, NS34: For the easier clear-cut ones, I was able to come up with an overall rating and then justify it. Though if there are contrasts, I like to go piece-by-piece... In some cases I got the upper idea and then I had to adjust the categories, my score of the categories to this overall idea. And then some cases, I just felt more comfortable to go one by one over the categories and then, calculate the overall score just mathematically. So it depends on the recordings. I'm trying right now to figure out in which cases it does. You know maybe when the speech is a bit more advanced, so maybe when it's three or four, you tend to get the overall impression and then try to adjust the categories. And maybe when it's lower, it's a bit easier to assess the categories in isolation and you try to first do this job and then to move on to the overall calculations.*

Changing a scoring decision. Regardless of the raters' approach to grading, top-down or bottom-up, several factors made the raters hesitate and change their initial score decisions. Two most common factors were the unstable quality of students' performance and lack of agreement between raters' initial impression and the score they decided on.

Raters' opinion could undergo changes if the quality of a student's performance fluctuated during the recording. Excerpts 34 and 35 illustrate raters' reactions to inconsistent answer quality throughout the recording. Since the first half of the student's answer was very good, these raters expected to hear more were confused when it did not happen.

> *Excerpt 34, NNS2: the speaker was fairly assured in the beginning, and I got an impression that more substantial information will follow. And then he stops and says some basic words, which confused me.*

> *Excerpt 35, NS30: I don't know what happened to the speaker in the second half. He started out so well. he started mumbling a lot, it was very odd.*

Not all the raters were confused by such inconsistent performance. Several raters noted that the students have some time to prepare and write down ideas and, therefore, have a good start; but then their answer quality can decrease due to the fact that they run out of prepared language and use a more natural spontaneous language (Excerpt 36).

> *Excerpt 36, NNS6: Well, there was a good start and it gave me the initial impression that that might be 3 and higher … because that person might prepare some kind of an answer in advance … that gave a good start but after that you need to use your natural language to advance to craft your response and to yeah to give some other reasons, and that person doesn't really have the ability to do it with the level of the language that he has… I would still give it 2, right because some good well-prepared phrases maybe in this response because maybe before the recording but still I can hear that the person struggles when he needs to use the natural language. So it's not 3.*

The other reason due to which raters’ impression tended to change was the exact opposite – when students had a low start and then began producing better language later in the recording. Excerpts 37 and 38 demonstrate raters’ reaction to this type of unstable performance.

> *Excerpt 37, NS34: I think particularly in the beginning of this, there's a bit of a slow start to unraveling thoughts.*

> *Excerpt 38, R14: because of the time it took him to warm up to the answer…Yeah because when he started out it definitely started out as a 1 but by the end he definitely started to pick this up.*

Both types of fluctuations in the students’ performance (better at the beginning and worse at the end or worse at the beginning and better at the end) caused raters to hesitate, think more, and change their decisions.

Furthermore, the raters did not exhibit congruence in grading performance by students whose response quality varied. The raters divided into two categories – some wanted to average viz. to take into account the good and the bad parts of the performance, and other raters stated that they do not want to overestimate students’ performance and usually decided on a lower grade. The raters in the first group had more hesitations about how to grade answers of inconsistent quality. Excerpt 39 shows how NNS24 still decided to give an overall grade of 3 even though they were mindful of the differences in quality throughout the response.

> *Excerpt 39, NNS24: For TD, it was very good for... actually, it's 3 for TD as he gave some relevant ideas for not studying in big classes. The overall development was very good in the very beginning, but in the end he had lack of ideas for small classes. SO, that's why maybe I've got ...well, I would still give him a 3. Actually... D is 3, yeah. And O3.*

On the other hand, the raters in the second category stated that the rubric descriptor "the response is sustained" in band 3 does not allow them to give a higher grade to such performance. Excerpts 40 and 41 show how two raters decided to give a 2 to the same inconsistent recording that received a 3 from NNS24.

> *Excerpt 40, NS17: At the beginning they were kind of sustaining their topic a bit. Then they dropped off and it picked up so it wasn't really sustained. Not a 3 and not a 1 either.*

> *Excerpt 41, NS7: He started strong, when I say strong I mean confident and loud but it drops off. He has some sophistication of his ideas in the beginning. But the pronunciation was pretty poor and required listener effort. For D, I would say 2. LU, I was looking in the threes but second half falls in the twos. And TD because of that falling off at the end, I would give a 2.*

The raters exhibited the same pattern when grading the recordings that had a worse beginning and a better end. Some raters decided to give lower scores due to not sustained discourse, but the raters who tried to average the inconsistent performance tried to "forget" or "forgive" the beginning. One rater, NS30 in Excerpt 42, mentioned that it is not fair to punish the students who do not start talking vigorously from the very beginning because it is not natural for humans.

> *Excerpt 42, NS30: In fact, I think it's a downfall of this kind of test. I do not think that the human brain works like this, where you have to suddenly be on, I think that most people need some time to warm up and this test doesn't account for that.*

Not only the inconsistent nature of students' responses caused the raters to hesitate and change their decisions but also the incongruence between the initial impression the raters had and the grade that the rubric descriptors prompted to give. Again, two strategies were identified – the raters changed their holistic overall grade to match the partial criteria grades, or they changed the partial criteria grades to match the overall holistic one. Excerpt 43 illustrates an honest response of NNS6 who admitted changing the partial criteria grades in order to be able to give the overall score that they initially had. The rater also stated that such reliance on the "gut feeling" is not appropriate.

> *Excerpt 43, NNS6: D is 3 LU is 2 and TD is 2 only because I feel that I should give overall 2. It's strange. It's weird and it's not appropriate for an assessor to think like this, yeah, to rely on my feelings inside, you know this gut feeling. It's not good reasoning, but O I give it 2. I feel because of the vocabulary that is quite poor I would say but the pace and the development of the recording was good enough, so again I should, I feel really ashamed but I gave TD 2 only just to justify O 2, something like that.*

Excerpt 44 also shows a rater who thought that the recording should receive an overall grade of 2, awarded a 3 for two criteria, but then changed the score for Topic Development to a 2 to be able to give an overall 2.

> *Excerpt 44, NNS45: I was moving my scores for D and for TD. I was sure that it was not an O 3 response but I was not sure if I say, if her D or TD were up to 3 because something out of those two was really not that bad and allowed me to understand her so I needed time to determine each one and I just moved those points on the scale for me to realize that.*

Unlike the previous two excerpts, Excerpts 45 and 46 show raters who initially gave a lower score but then changed it to a higher score based on their decision for partial criteria scores.

> *Excerpt 45, NS37: I originally gave her a 3 O and then I moved it up to a 4. Looking at the rubric made me change my mind.*

> *Excerpt 46, NS30: I wanted to give it a 2 but I felt like the rubric would give it a 3.*

In addition, the raters did not always follow the same strategy – to change the overall to fit the partial or change the partial to fit the overall. Sometimes the raters tried to fight their "gut feeling" and abide by the rubric, and sometimes trusted their feelings more. This internal iteration involving raters' willingness to make a decision that is guided by the rubric and not by their feelings but at the same time to assess correctly was causing hesitations in many raters. Excerpt 47 illustrates NNS45 who did not want to give an overall 1, but could not justify a higher score based on the rubric. This rater was also mentioned before, in Excerpt 44, when they chose to change the partial scores to give a grade that matched their overall impression. Based on these two examples, we can make an inference that general impression was more important for NNS45; however, in the first case, the rater was able to find some descriptors in the rubric criteria to justify the change, but not in the second case.

> *Excerpt 47, NNS45: well, I still kind of don't agree that's it a 1 O but when I get back to each scale separately I still end up with the same grade I have given him so that just mathematically a 1 O, it does not sound to me like a 1 response so it is what it is. I don't see how I could change anything here… I think that's about inconsistency between my impression of the whole answer of the student and their specific grade because you go to the grading scales and this, well let's say a 2 on D, 1 on LU, and so on. But, if you did actually understand that person and if you had to just give the overall grade, it might have been a hard one. So, that's the inconsistency that I had to overcome and that caused hesitations.*

To summarize, the raters did not show any specific grading strategies depending on the NS or NNS rater group affiliation. Moreover, based on the description, the TOEFL iBT rubric can function as a holistic or analytic one, but it is not clear which approach the rates should employ. The raters in this study utilized two grading strategies, top-down and bottom-up, to arrive at the initial score or scores. Most of the raters relied on their overall impression about an examinee’s answer (top-down), three raters made more analytic decisions about partial criteria scores first (bottom-up), and two raters used both approaches. Due to the fact that the majority of the raters followed the top-down approach, it can be inferred that it is easier and more natural for raters to employ a more holistic approach to grading. Furthermore, all raters, regardless of the approach, tended to make an initial decision while listening and then support it by scanning the rubric. If a decision was easy (usually extreme cases such as a 1 or a 4), the raters did not need to look at the rubric to confirm their grade. Furthermore, all the raters experienced some moments of hesitation when the overall quality of students' performance was inconsistent and when the score the rubric prompted to give did not match the raters' general impression.

References to Non-Rubric Criteria

The raters in this study verbalized their thoughts while giving their scores to the students while using a well-established scoring rubric, nevertheless, they sometimes tended to either bring in additional non-rubric criteria or biases to facilitate their rating or interpreted existing criteria in their own way. Some raters mentioned that they are aware of their tendencies and tried to control them, whereas others did not realize that they are following some non-rubric criteria while rating. The raters exhibited several recurring patterns, and the most common ones are described below. These decision-making patterns were grouped by the rubric criteria that they can fit under, namely non-rubric criteria for Delivery and non-rubric criteria for Topic Development. There were several patterns that can be attributed to Language Use, such as differences in interpretation of what minor and major errors are and what complex structures entail, but they did not show as much prominence. The raters did not show any specific patterns depending on the NS or NNS rater group.

All of the raters referred to all or most of these non-rubric criteria during the think-aloud session or while reflecting on their grading in the interview. Some raters noted that they potentially could be affected by the following biases subconsciously and not consciously; others explicitly stated that they are aware of the possible effects and tried to control their perceptions. Nevertheless, some raters were not aware and, therefore, did not question their interpretations, and several of them admitted that talking about their process of rating lead them to the realization that some criteria that they used to guide their grading were not on the rubric.

Non-Rubric Criteria for Delivery. The criteria that are described in this section were not included in the rubric, however, some raters referenced to these criteria while providing justifications for their scores for Delivery during the think-aloud protocol or reflecting on grading Delivery during the interview. The most frequently non-rubric criteria were voice quality, accent familiarity, and unfamiliarity.

Voice quality. The first frequently mentioned non-rubric criteria or bias was the quality of examinees' voice, which was controlled or uncontrolled for by some raters. The raters noted that they could be prone to assigning higher scores to those students who had more confident and loud voices and lower scores to softer or quieter speakers equating these features with lack of confidence. Excerpts 48 and 49 show examples of raters who took confidence or lack of it into account while grading students' responses. The raters did not talk about the same student. Excerpt 48 illustrates a rater who mentioned quiet voice and lack of confidence when justifying an overall score of 1. Excerpt 49 shows another rater who referred to confidence but from a different perspective; the rater decided to give a 2 not a 1 due to students' confidence.

> *Excerpt 48, NS7: I'm giving a 1 across the board. He begins with a filler. He is very monotone, very choppy, very quiet voice, not a lot of confidence there.*

> *Excerpt 49, NNS28: He sounded confident, he spoke without any pauses and it was pretty quick but the intonation was poor, so in this case I would give him a 2 just for confidence.*

Not all the raters associated quieter voice with lack of confidence. Some raters said that quieter voice is one factor that makes them listen to students' responses more than once because they do not want to punish examinees for their natural voice quality or recording problems. Excerpts 50, 51, and 52 demonstrate raters' opinions.

> *Excerpt 50, NNS24: He was with low voice, but it should not have affected anything.*

> *Excerpt 51, NNS45: I was listening and checking my volume bar but it was up, yes, he spoke quietly. And then I tried to make my grades but that didn't go so well, that quickly, and then I considered listening again.*

> *Excerpt 52, NNS28: Sometimes I had to lean over, and listen closer because it was difficult to understand what they were saying. But, in general, I guess that it shouldn't affect grades because maybe it's just the way people speak generally with their families and friends, it's just not a physical ability but a physical feature.*

To continue, several raters did not see confidence and quieter voices as major factors and were not affected much by confidence or lack of it. Excerpt 53 shows a rater who awarded the highest grade even though they noticed that that student spoke quietly. Although NNS35 (Excerpt 54) mentions that confidence is important, they say that confidence is not the critical factor that can make them give a higher grade.

> *Excerpt 53, NNS46: The person was talking a bit quiet, but it did not prevent from understanding. Yeah, I think it's a four. I think it's very good.*

> *Excerpt 54, NNS35: For confidence, it's definitely important but there were definitely cases when the person was confident or spoke pretty loudly, but they still had issues with pronunciation and intonation and I still couldn't understand what they were saying, no matter how confident they were. Maybe they were too confident, I don't know.*

In addition, one rater had a very different stance on quieter and louder voices. They noted that softer delivery is more appealing. Excerpt 55 demonstrates this rater's opinion. Interpreting raters' thoughts, it is evident that the rater preferred accuracy over fluency and did not favor filled pauses. Additional inferences can be brought up by the researcher based on the rater's own style of speaking. NNS13 spoke softly, quietly, and slowly (both in L1 and L2), therefore, it can be hypothesized that this rater subconsciously preferred students with a speaking style that matched their own.

> *Excerpt 55, NNS13: I noticed if the speaker spoke quite softly and in a very not loud tone, that impressed me more than the students who were talking very loudly, I even paid more attention to such speech patterns, because actually they were more accurate and more fluent than those who had a loud voice and spoke with filled pauses and had disrupted speech.*

Moreover, some raters expected students to speak in a more confident manner in order to earn higher scores such as a 3 or a 4. These raters believed that confidence in their interpretation is students' ability to speak with few pauses, show their knowledge, and display solid language skills (Excerpts 56 and 57).

> *Excerpt 56, NNS24: In the beginning she talked without pauses and she gave the impression that she is confident about her answer and she knows what she's going to say.*

> *Excerpt 57, NS7: Absolutely, yes. In fact, I wrote confident down, 1, 2, 3 times. I mentioned pacing as well. Volume of course would go into the confidence factor. I wasn't necessarily harsh, there were some female speakers whose voices were softer who I don't think it really influenced me but there were a few, one gentleman. The one who started with the 'uhh'. His voice was just really soft the entire time and I think it was soft because of a lack of confidence and lack of really having a firm grasp of what he wanted to say and how to say it. Volume, definitely. I think pacing is important, yeah. One more thing, the monotone enunciation would, I guess that would fall under the rubric for intonation.*

Other raters drew a line dividing the rubric in half where they saw speakers who can earn higher grades as confident. For example, Excerpt 58 illustrates how NS37 considers confidence when deciding not to award a lower grade of 2. Also, Excerpt 59 shows a rater who mentioned confidence among the reasons for not awarding an overall grade of 3.

> *Excerpt 58, NS37: She was not a 2. She was definitely a 3. A 2 is not very confident.*

> *Excerpt 59, NNS24: O3? Well, I think it's too high for this response 'cause he was not confident, he was not elaborating the answer.*

In addition, quieter speech and audio quality were sometimes associated with the descriptor "requires listener effort" mentioned on the rubric. The raters in Excerpts 60 and 61 stated that it is hard to rate quieter responses.

> *Excerpt 60, NS14: Sometimes pronunciation or quietness makes me be more harsh because I have to work harder to hear it. The harder you have to take to listen, the more your brain works, the harder you'll be on the grading…It was difficult to separate the quality of the recording and the quality of the speaker's responses to rate them.*

> *Excerpt 61, NS17: Audio quality, probably. If I can barely hear what these people are saying, it's pretty hard to give 4s. It's a lot of listener effort if I cannot hear it.*

Excerpt 62 demonstrates an opinion that is connected to the one mentioned above. The rater believes that it is the students' job to speak loudly and confidently because it is hard for the raters to score quieter answers.

> *Excerpt 62, NNS2: I would first say that students who speak very quietly make it extremely complicated for me to understand and, like, whether I want it or not, it contributes to the overall negative impression, so they need to articulate their words clearly and speak loudly. Yeah, difficult though it may be for them, the students, but yeah, it's very important.*

Apart from confidence and quieter voices, there was another factor that can be attributed to the voice quality – how natural or friendly examinees sound. Excerpt 63 describes several thoughts by NS34, who mentioned that the way they perceived one student's tone made it hard for them to score objectively.

> *Excerpt 63, NS34: It was a very unnatural sounding voice in English…. I think because of the pronunciation and tone quality, it makes sense because I think when you hear someone with that unnatural voice, it's hard to imagine that they're quite proficient… a tone quality issue made me not want to listen to it anymore, and could have impacted the score … I found someone with a voice that was outputting and hard to understand, due to how they use vocal sounds, and this may have influenced me to give a harsher rating.*

In Excerpts 64 and 65 the raters refer to how "friendly" or "pleasing" the student sounded talking about the same recording by a female Russian speaker.

> *Excerpt 64, NS30: Her intonation was really natural and familiar. Her intonation was so friendly.*

> *Excerpt 65, NNS10: In terms of D, her accent was very pleasing.*

Even though it was not a recurring pattern, one rater suggested that raters might factor in their unconscious biases stemming from the global political situation (Excerpt 66).

> *Excerpt 66, NS40: More psychological or sub-conscience factors behind certain raters' decisions could be based on their perception of accents and the people behind them in terms of the global political climate, things of this sort,… It is apparent to me that the speaker is an Arabic speaker, therefore, maybe they just have an inherent bias. I would*

> *hate for that to be the case but that's something that could be an explanation I guess. Maybe they found the voice more intimidating or something.*

Accent familiarity. The next frequently mentioned controlled or uncontrolled bias was raters' accent familiarity or unfamiliarity. It should be highlighted that the researcher never used the word accent in any questions throughout all the steps of the data collection. Nevertheless, the accent was mentioned by all the raters except for NS30, who always used the word pronunciation. Similarly, even though rater NS34 mentioned accent several times, they highlighted their ability to tease apart accentedness from comprehensibility. For accent familiarity as a non-rubric criterion, the raters formed patterns based on their familiarity levels with examinees' L1s used in the study.

Familiarity and unfamiliarity. In general, the raters mentioned that they found it easier to score test-takers with familiar accents because they could listen through typical pronunciation issues and understand the message easily. Excerpts 67, 68, and 69 illustrate such opinions.

> *Excerpt 67, NNS10: His accent was Russian and I could understand him and his accent was familiar to me that's why I could catch his ideas*

> *Excerpt 68, NNS45: she's I suppose Russian so I can understand her intuitively, but still it did not spoil anything for me. And the D is a 4 because she's coherent, intelligible, not that strong of an accent did not preclude me from understanding her speech again so okay that's a 4*

> *Excerpt 69, NS17: I hear people with her accent everyday so it wasn't hard to understand... She's Asian-based student, I can tell. I teach them every day so I'm fairly comfortable understanding them.*

Some Russian raters admitted that they could have potentially exhibited a positive bias towards Russian speakers because of familiarity. In Excerpt 70, NNS6 admitted that it was easier for them to understand the ideas produced by a person with a shared L1 background, but the rater noted that they could not speak for the speakers of other L1s. On the other hand, NNS45 in Excerpt 71 controlled for their positive bias towards Slavic speakers. The rater did not narrow down their scope to Russian L1 students because they thought that Ukrainian, Belorussian, and other Slavic speakers have the same Slavic accent.

> *Excerpt 70, NNS6: Russian speakers, yeah, well you can define them from the accent I suppose, and for me, it makes it easier to understand, I don't know about Chinese or Arabic assessors, whether it's easier for them. But it influences the idea. The overall idea…Yeah, it definitely affected my scores, of course. If I had known, maybe Arabic, yeah, yeah I would have understood it better, yeah. Chinese, the same, sure.*

> *Excerpt 71, NNS45: Apart from the thing with a Slavic language native speakers whom I understood well and therefore had to stop myself from automatically grading their D higher than it should be graded.*

On the other hand, some Russian raters thought that the effects of the L1 match are overstated because all non-native speakers can have pronunciation problems (Excerpt 72).

> *Excerpt 72, NNS2: Well, it goes without saying that it's maybe more comfortable for me to listen to the Russian speakers because I'm really familiar wtih this English speech by the Russians, but by and large I believe firmly that if a person has difficulties with pronunciation it will tell in his speech regardless of his nationality. So if a Russian who needs to work on his pronunciation then I will not understand him well. So yeah, I think this degree of familiarity with the Russian speaker--well, not that it overstates it, yeah, I may say that the influence of this familiarity may be overstated because I mean it has to do with your pronunciation features.*

In terms of lack of familiarity, some raters reported that they might have had unconscious negative bias towards Chinese speakers due to the fact that this accent was unintelligible for them. Excerpts 73 and 74 show two raters who describe the difficulty of deciphering unfamiliar accents as “it wasn't English”.

> *Excerpt 73, NNS6: And then again in the beginning, if I didn't know that it was a TOEFL answer I would be even doubtful whether it's English.*

> *Excerpt 74, NNS13: Maye it happened with the woman again who had some influence from her native language so there was a very strong accent and I didn't like it because I wasn't able to understand what she was talking about though she sounded very fluent and the guy too--they sounded fluent, but it wasn't English, so I was rather harsher maybe on some test takers.*

Additionally, Excerpt 75 shows a comment by NS7 that shows some inexplicit stigma that the speaker ascribed to Chinese accent through comparing it to Arabic accent. Many times throughout the study NS7 reiterated their familiarity with Chinese accent due to substantial time spent living in China and teaching Chinese students. NS7 was explaining why some raters can give a higher score to an Arabic student and emphasized that greater exposure to Chinese accent can train a persons' ears so that other accents seem clearer.

> *Excerpt 75, NS7: It didn't require a lot of listener effort. Maybe for the D there might be some bias. If you teach Chinese students, you might lean towards a 4.*

Even though some more unfamiliar raters showed some skepticism about what they called "Asian pronunciation features" (Excerpt 77), at the same time, they admitted that this is something they were controlling for. In other words, they were aware that they are unfamiliar and, therefore, need to pay more attention and employ guessing strategies (Excerpt 76). These raters also stated that they are willing to learn more about L1 transfers from different languages (Excerpt 77).

> *Excerpt 76, NNS6: I don't understand what the person is saying because of his accent and I really had to guess for example "blah blah knowledge" he said something like "blah blah" and I figured out that this has to be "widen my knowledge" the same as with "friendship" what it has to do with the "friendship shshs" and I thought it might be "destroy," so mainly I had to guess.*

> *Excerpt 77, NNS2: Well I do not think it affected my score to any considerable degree because I try to be indifferent--well, of course, as I, well my lamentations about this lack of intelligibility from Chinese--Asian speakers, yeah that all remains in place. I don't take my words back. They might need to work on their pronunciation, But overall I would say that these national peculiarities they don't--at least in my case, they didn't affect the way I scored the response because everybody takes--around the TOEFL is taken around the world, right, and it's only natural that we see so many different accents... That actually takes us back to what I was saying about the pronunciation I might be more lenient on delivery because yeah if the person is fairly intelligible but makes--the flow of his speech is not satisfactory enough or if there are some mistakes specific to some languages yeah like Chinese for instance, like mistakes which are inherent to these speakers because they just can't get rid of them especially if they started learning the language as adults. So,*

> *yeah, I would say if I had to give real scores I would require more training on D, including maybe a special session on how to deal with local accents*

Several raters were familiar with all the accents used in the study and acknowledged the fact that this could have resulted in positive bias. Excerpt 78 exemplifies a rater who was familiar with all L1s but thought they might have some positive biases only towards Russian L1 speakers, with whom they had the most familiarity. The rater in Excerpt 79 also was familiar with all the L1s but did not single out a specific L1 they could have been more positively biased towards. This rater thought that their familiarity with all the L1s in the study could have resulted in overall more lenient rating patterns. Both raters did not control for their extensive familiarity with examinee L1 backgrounds.

> *Excerpt 78, NS34: Because I'm in a legacy Russian-speaking area, there may have been subconscious sympathies with that because this is familiar. There definitely may have been subconscious biases. I find it endearing to find Russian accents and grammatical structures…*

> *Excerpt 79, NS40: Absolutely. My familiarity with their delivery because of getting used to their accents, speech patterns, and working with English language most likely caused me to be more lenient.*

On the other hand, some raters tried to control for their comprehension benefits due to familiarity, and they found it to be a cognitively demanding process (Excerpt 80).

> *Excerpt 80, NS7: With the Chinese students, I think this came up, I think I can understand Chinese intonation pretty well now. Whereas someone who might not have been exposed to the all the classroom time that I've had will have a harder time hearing some of the common pronunciations of certain words and certain structures… I was trying to control, which requires more effort for me to say, okay, I understood that but you know grammatically, the pronunciation it is pretty far off from what a native speaker would be accustomed to. So that sort of process happening in a millisecond in my head but its still going on and it requires energy.*

Several NS raters had extensive familiarity with the way Arabic and Chinese people speak in English, but not all of them were familiar with the way Russians speak in English.

However, some of them noted that they did not have any trouble understanding Russian examinees (Excerpt 81).

> *Excerpt 81, NS30: But I will say that it seemed... the Russian speakers in these samples happen to be very easy for me to understand. But if I had had speakers from countries I'm not familiar with, like other countries, it might have been hard. So, I'm really familiar with Arabic and Chinese speakers, so that was easy. Not so much with Russians, but I felt like they were talking like... easy samples for some reason, so...*

Another rater's opinion may explain why the raters' who were unfamiliar with the way Russians speak did not experience much trouble understanding them. In Excerpt 82, NS34 hypothesized that Russian is phonologically more similar to English.

> *Excerpt 82, NS34: Everyone's biased, right? How the bias plays out is very different. For languages that are phonologically similar to English, Spanish or Russian, the bias tends to be positive because you hear these crystal clear phonetic things. But for other languages that have really different phonological backgrounds, such as Chinese, it can have a big impact, negatively.*

Ability to recognize accents. This sub-section provides examples of how the raters were or were not able to recognize specific L1 accents. It talks about patterns of both NS and NNS raters who were familiar or unfamiliar with examinees' accents.

Not all the raters were sure about what kind of accents they hear even if they listened to students whose accents were familiar or shared their L1. Based on the think-aloud, only four out of nine NNS raters were more than sure that the student there were listening to was Russian. In Excerpt 83, NNS6 draws on her own experience of learning English in Russia when describing a typical introductory statement produced by a student. On the other hand, the same rater was not sure whether a test-taker was Russian. In both cases, NNS6 was grading a Russian L1 student.

> *Excerpt 83, NNS6: But still O I would say that it's 4 despite some minor lapses as it said in the table for example it's not grammar it's grammatic and it's the beginning like in the classical Russian school "Today, our theme is about universities".*

> *Excerpt 84, NNS6: I think that was very easy in this case to assess this because the accent was pretty clear for me there was an accent of a nonnative speaker but that really, first of*

> *all that really sounded like Russian a bit I don't know as a rater because I don't know where the person comes from but this sounds clear for me and yeah. and the speaking was clear so the whole recording was clear for me and the structure and vocabulary that is used, nothing that I needed or I felt that I needed to listen to again.*

One NNS rater hypothesized that a Chinese female student is Spanish or Latin American (Excerpt 85). The same rater also thought that an Arabic male student is Russian (Excerpt 86). It can be hypothesized that the rater might have been misled by the phonological similarity of Russian and Arabic articulation of [r] in English. It was interesting to see that a Russian native speaker made this mistake. Similarly, one NS reported grouping Russians together with Arabic (Excerpt 87).

> *Excerpt 85, NS10: I was trying to listen very attentively, very carefully, and just trying to catch any words but it was really hard. So I think maybe it's because of her accent, or maybe because she's Spanish or Latin American, I don't know.*

> *Excerpt 86, NS10: It was not so difficult for me to listen to him and to catch his ideas. I would choose easy because he is from Russia and his accent is very familiar to me and that's why I could understand him.*

> *Excerpt 87, NS7: No. Well, yeah. Arabic and Chinese. I don't think I noted that any of them were Russian speakers….Because I'm more familiar with them through PIE. I was just going to say, I'm wondering which ones were the Russian ones now, and if I grouped them with the Arabic speakers….I don't know…That's probably what I did. I probably thought that I was listening to an Arabic speaker.*

There were more examples of incorrect L1 identification. One NS labeled a female Arabic student as a speaker of a Latin-based language even though they emphasized their familiarity with Latin-based languages (Excerpt 88).

> *Excerpt 88, NS37: Her D was, I mean she is probably Italian or Greek, I am guessing maybe she speaks a Latin-based language. I am familiar with Latin-based languages, so maybe I overlooked something.*

Additionally, based on the interview, two NNS raters thought Arabic test-takers were Indian (Excerpt 89). One NS rater (Excerpt 90) first stated that they were able to identify all accents, but then they changed their mind saying that they could only identify Chinese (and did that correctly in the think-aloud).

> *Excerpt 89, NNS2: Absolutely. I was able to distinguish a Russian speak--well, one for sure, one Russian speaker for sure. Quite a few Chinese speakers, and I believe that whenever I wasn't sure what accent it was, it was Arabic, yeah, because I thought it might have been Indian.*

> *Excerpt 90, NS37: I could distinguish the Chinese one. I could distinguish the Russian but the Arabic ones I couldn't. And you know what, I think the only one I could say definitively is the Chinese. They have a very distinct accent that's unmistakable. Asian accents are unmistakable. Other accents tend, they can blend; they're not as definitive.*

In sum, almost all raters were able to guess that some examinees had “Asian background”. Regarding NS raters, five out of seven were highly familiar with Arabic and Chinese students. Thus, these raters reported that they were able to distinguish Russians only because they differed from the other two. However, one NS rater, who was highly familiar with Arabic pronunciation, self-reported labeling Russian examinees as Arabic. One NS speaker who was not very familiar with all the accents used in the study identified several examinees’ hypothetical L1 background as Asian when listening to a Chinese student and Latin-based when verbalizing thoughts about an Arabic recording. In terms of NNS raters, two raters thought that Arabic examinees are Indian, one rater identified one Arabic examinee as Russian and a Chinese speaker as Spanish or Latin American. Four NNS raters guessed one or two Russian examinees during the think-aloud. Only one NNS rater reported their ability to identify all three L1s due to familiarity.

Whose bias? Another line of patterns about familiar and unfamiliar accents can be described as “Whose bias?” The name of this section was inspired by articles by Lowenberg (1993) “Issues of validity in tests of English as a world language: whose standards?” and Davies, Hamp-Lyons, and Kemp (2003) “Whose norms? International proficiency tests in English”.

Most of the raters who participated in this dissertation study believed that the biased are those who lack accent familiarity and global citizenship. In Excerpt 91, NS14 equates accent familiarity with losing bias.

> *Excerpt 91, NS14: I don't think specific accents of backgrounds play too much into it, because I had a lot of students from a lot of different places and so I feel like much of that bias has been lost, luckily…I think the person who does not have experience has more of a bias because they are only going off of their background and what other people said about these cultures, or what they see in the media, or what they've learned of other native speaker may or may not be correct. And no background with the culture or the language I would say that they are probably more biased.*

On the other hand, a few raters considered accent familiarity to be a bias. They reported that they do not want to be positively biased towards the examinee groups that they are more accustomed to; therefore, they "take off their ESL hat" (Excerpt 92). In addition to this, some raters felt it is probably better to rate students' exams not from an ESL teacher's perspective, but from a perspective of a person from a university campus (Excerpt 93).

> *Excerpt 92, NS7: That's exactly it, because as a teacher, you become very accustomed to certain grammatical structures that are errors but because you've heard them so many times, you understand them. I think that certainly becomes a filter from which you hear and I think you have to try to get rid of that filter and listen as someone who's going to hear that for the first time. It's extremely difficult to do that, I think. But, I was trying to control for those. When I heard a certain word pronounced a certain way, because I know what they're saying because I have been exposed to more Chinese students per se, I would try to hear that word as it was being said, if that makes sense?*

> *Excerpt 93, NS34: Yes, but I think it should be a random person on a university street with significant international student populations and interactions, instead of a random person that works in the grocery store that may not really interact with international students.*

To continue, most of the raters expressed strong feelings about keeping their ESL hats on while grading students and had an opinion that rating as a random person in the street is a disservice to students. Excerpts 94 and 95 illustrate raters' opinions who thought that controlling for accent familiarity is not rational as it is unknown how an unfamiliar accent will be processed by a "random person in the street".

> *Excerpt 94, NS30: Okay, so I don't take off my ESL hat, and here's why. I think it's a very simple reason. Some people on the street or a random person can understand a Chinese speaker much easier than an Arabic speaker, and vice versa. And I think that's really unfair to put that as a factor, so no. So what I try to do is I definitely listen as an ESL teacher because I just think a lot of raters are overwhelmed by pronunciation issues and there is a lot more going on than pronunciation*

> *Excerpt 95, NS14: I think it does a student a disservice, because it does bring back the biases and difficulties in understanding pronunciation.*

Referring to the same topic, NS17 expressed an opinion that "normal people" have biases and would not be able to award passing grades to non-native speakers (Excerpt 96).

> *Excerpt 96, NS17: I don't think it's fair. Normal people may have biases on comprehension with these student's ratings, so many of these students wouldn't pass if normal people were rating them.*

In addition, NS37 argued that it is rude not to be willing to negotiate the meaning even if there are some comprehension problems (Excerpt 97).

> *Excerpt 97, NS37: you run into somebody on the street, well, that's just a rude person. You know, somebody who doesn't know the language, I'm gonna try to come down to the lowest common denominator so that we can communicate.*

Not only did many raters try to classify people into biased and unbiased based on the amount of accent familiarity but also expressed ideas about who can be the ideal rater. For example, NS34 thought that raters should not be completely unfamiliar because it does not match the reality of university campuses, which are very diverse. NS34's opinion was that a person with some familiarity is ideal, but people who have extensive familiarity are also acceptable (Excerpt 98).

> *Excerpt 98, NS34. Someone with a bit of exposure, is the ideal middle ground. Someone with extensive exposure, you or me, are also valid because we are represented on these populations but we are not the average. I think TOEFL should train on intelligibility and*

> *naivety. Striving for the middle road, is the best, because it is a changing dynamic world and test.*

On the other hand, some NNS thought that extensive familiarity such as shared L1 could have extensive effects on student's grades. For example, NNS46 had an opinion that raters should not rate students whose L1 they share but did not have a strong stance.

> *Excerpt 99, NNS46: And maybe, as I said, if you are Russian, you should not rate Russians. Maybe I am wrong.*

Moreover, in Excerpt 100, NNS24 had an opinion that contradicted the opinion by NNS46 in Excerpt 99 and was more in line with the idea expressed by NS34 in Excerpt 98. NNS24 thought that it is important to have accent familiarity with the way the students with various L1s speak in order to grade their responses confidently and with fewer hesitations.

> *Excerpt 100, NNS24: I wouldn't really like to score Chinese or Arabic, because I'm not really used to their way of speaking. I can give lower scores, or maybe higher. I would have many doubts. If you rate someone, it should be someone closer to the way... I mean, not a native speaker. The teachers should be trained to score specifically Chinese or Russians, they should be differentiated. They shouldn't go to the exam and see Chinese speaker for the first time and score him, or Indian.*

Similarly, several unfamiliar NNS raters also questioned the role of their unfamiliarity and reported having hesitations when listening to unfamiliar accents. For example, NNS2 in Excerpt 101 was wondering if the students' performance causes poor comprehension or it was the lack of familiarity. They were aware that accent is not part of the rubric and wanted to have more guidelines. Nevertheless, even though NNS28 did not see accent among the rubric descriptors (Excerpt 102), they wanted to factor it in by attributing it to Delivery or Language Use category.

> *Excerpt 101, NNS2: Well, I remember these students as I understand of Asian descent and, predominantly I had difficulties with these students because I had hard time understanding their accent. ...So, yeah, for me it was difficult with their pronunciation features, mostly this, because if I have other questions they clearly fall into these categories already outlined for us, and I can--in case I have difficulties I can refer to the scoring rubric and then if I don't understand the accent I'm like "is that on my side or is it on their side?"*

> *Excerpt 102, NNS28: because of the accent, because I wasn't sure how we grade accent is it connected with D or with LU.*

Finally, it is important to note that several raters were annoyed by the fact that the rubric had already factored in rater accent familiarity bias into the rating criteria by referring to the amount of listener effort required. In Excerpt 103 NS37 brings up this issues by citing the rubric. NS37 also stated that listener effort is a rater-specific, not a student-specific criterion. In other words, the amount of listener effort will change based on who is listening to a student's recording.

> *Excerpt 103, NS37: "Consistent pronunciation, stress and intonation difficulties cause considerable listener effort" this is what I call bias that is already factored into the assessment criteria. Maybe it causes for me, but not for others.*

To summarize, the raters approached accent familiarity and unfamiliarity from a variety of different perspectives. Some believed that accent familiarity could positively bias their scores and accent unfamiliarity can negatively bias their scores. Others stated that they try to control for their greater familiarity, but most of the raters believed that unfamiliarity, not familiarity, is the cause of bias. Thus, raters should not be unfamiliar with examinees' L1 specificities.

Non-Rubric Criteria for Topic Development. Among the non-rubric criteria for Topic Development that the raters referred to there were raters' attitudes towards finished and unfinished responses, utilization of writing-like organization, prompt reading, critical thinking and unique ideas, and making an explicit choice when responding to prompts formulated as alternative questions.

Finished or unfinished responses. All the raters divided into three groups based on their approach to grading good responses that were cut-off mid-word at the end (i.e., negative, positive, and neutral). The negative group of raters firmly believed that the students need to come to take such exams prepared and examinees' inability to work with the timer signals this lack of preparedness. These raters tended to lower their scores for Topic Development category if a student did not manage to wrap-up their response (Excerpt 104). Most of the raters held this opinion when assigning higher grades of 3 and 4 (Excerpt 105 and 106). As it was described before, a student cannot get 4 out of 4 for their answer if they receive even one score in band 3. These raters decided on a 3 instead of a 4 only because the student did not have a conclusion before the time was up and backed-up their decision with the descriptor "coherent" in the Topic Development criteria (Excerpt 107).

> *Excerpt 104, NNS2: the person should be prepared to work with the timer and make sure he or she finishes on time. So, that's something to take points off the answer, yeah?*
>
> *Excerpt 105, NNS46: the main point is that the speech was not finished properly. It was abrupt, that's why I think it cannot get three in this case, only two….I guess it's the same as topic development when people could wind up. Which was a cut. And I would agree with those who would say that they could not manage their time well.*
>
> *Excerpt 106, NNS28: And speaking of TD, the start was definitely better than the end because he definitely needed more time to finish his pitch, so in general he still answered the question so I guess that also would be a 3.*
>
> *Excerpt 107, NS17: I assumed it was built into the rubric for topic development. It is coherent, if you finished in time. They could word it more specifically into the rubric, but I think it's already implied.*

On the other hand, the neutral group of raters did not see unfinished responses in either negative or positive way, and their scores were not affected. The positive group of raters tended to see the unfinished responses as something positive; they saw the potential, abundance of ideas, and students' ability to talk longer (Excerpts 108 and 109). Compared to the negative group of raters, these raters awarded higher scores. However, one rater noted that they did give a lower score to an unfinished response even though they see unfinished responses as a positive feature (Excerpt 110).

> *Excerpt 108, NNS6: I wanted to find the answer is it appropriate for response with a score of 4 to be, not to finish the response, it wasn't enough time for the response for this girl to finish her ideas or she forgot completely about the time limit*

> *Excerpt 109, NNS45: if they, let's say they completed the assignment, basically, but their speech was not finished on time or because their buzzer buzzes; then, that does not mean anything to me.*

> *Excerpt 110, NS7: That's funny, because I think I scored one girl because she didn't finish for her TD as a 3 but actually I see it as a positive.*

Organization. All the raters discussed recordings' structure, logic, and organization, but, again, had various opinions. To begin with, describing the Topic Development category, the rubric states, "It is generally well-developed and coherent; relationships between ideas are clear (or clear progression of ideas)". When the raters were making their Topic Development decisions, they tended to interpret "well-developed and coherent" in different ways. One example was already mentioned in the previous sub-section in Excerpt 107 where NS17 thought that unfinished responses do not fall under "coherent" descriptor.

Roughly, the raters can be divided into three groups: (a) pro-organization, (b) pro-logic, and (c) either one works. The raters who favored typical organization structure wanted students' responses to have an introduction, body with a least two reasons, and a simple conclusion. These raters had experience teaching speaking for such exams as TOEFL and IELTS and, therefore, were very disappointed not to see such organization that is taught in every exam preparation book. Several raters expressed an opinion that if a student follows such generic structure, they show that they prepared for the exam (Excerpt 111). Some raters also believed that speech that does not follow such organization pattern could not be considered coherent, well-developed, and academic (Excerpt 112).

> *Excerpt 111, NNS24: For TD, when I was saying that good introduction, introductory, that was important, because it means that, first of all, the person is prepared for this exam and he is used to answering questions in a short time.*

> *Excerpt 112, NS34: I don’t think that organization is different from well-developed, coherent, and has relationship progression of ideas. Organization is more strict, but has these pieces. A conversation response may have a high score, but it is different than an academic response.*

On the other hand, the pro-logic raters argued that normal humans do not speak the way they write and that spontaneous speech is logical, but not meticulously organized. Some of them had a strong stance on making a change and stop teaching “canned” phrases (Excerpt 113). Others noted that natural speech patterns are more authentic but that utilization of templates did not affect their scores either (Excerpt 114). Also, some raters did not treat template-like responses and logical spontaneous speech differently (Excerpt 115).

> *Excerpt 113, NS30: In fact, I think that making an introduction and conclusion sentence is the easiest part of the task. So if a student skips that and merges right in, if they have all the different points. There was an Arabic speaker, I think, I had noted that he just went went straight in, I don't mind that at all. In my mind, that's what I think of as a pure sample, because it's not surrounded by any super-easy canned sentences. That does nothing for me. I wish we would tell them not to do that.*

> *Excerpt 114, NS7: Yeah, I think for this, you could tell when there was that organization. It didn't really affect the score there for me because it seemed more canned, their responses. The ones that really got me were the ones that used the natural transition words. It seemed more natural, seemed more automatic, they were transitioned between their ideas.*

> *Excerpt 115, NNS45: I think I felt okay with the lack of the TOEFL book templates because I think the rubric pretty much says it all about what they expect from a student and that's a coherent, completed, simple thought on a rather simple matter from the person who's asked a question and has a limited amount of time to answer it. I don't think let's say, native speakers would use a template*

Reading the prompt. Reading the prompt was another pattern that exhibited prominence among the non-rubric criteria for Topic Development. Many raters did not consider reading the prompt as a drawback, but they rather saw it as a student making themselves a disservice by wasting their precious time. On the other hand, some raters were more adamant about it and automatically decreased their scores one band down if a students' response started with prompt repetition (Excerpt 116).

> *Excerpt 116, NS17: Once the student read the prompt it automatically dinged it 1 level… If you imagine that these responses have a beginning, middle, and an end, it didn't have a beginning and the ending is missing.*

The differences can be explicated by looking at Excerpts 117 and 118, which were produced by raters who treated prompt reading differently. Both raters listened to the same student (Recording 11) and noticed that the prompt was read; however, arrived at different scores. Both raters agreed on a 2 for Delivery, but differed one point for Language Use (NS17-2 and NS40-3) and had a two-point difference on Topic Development (NS17-1 and NS40-3). As a result, these two raters provided adjacent holistic overall scores.

> *Excerpt 117, NS17: TD was a 1. The first 12 seconds was just reading the prompt. It was just the same idea over and over. ... I'd give it O a 2…TD was not good.*

> *Excerpt 118, NS40: The TD was good as well. The first 15 seconds the student was reading the prompt verbatim so I felt like that wasted some time. So I gave an O 3.*

Making a choice. Making a choice was another criterion that some raters referenced when they were deciding on the Topic Development grade. Due to the fact that both speaking prompts used in the study were alternative questions, the raters wanted the students to be specific and make their choice (Excerpt 119). The raters did not favor students who generally spoke about pros and cons of either side of the question without explicitly stating their preference. For example, Excerpts 120 and 121 show two raters who decided on giving a lower grade for the reason that the student did not choose a side. The raters wanted to hear a clear statement that one option is better than the other.

> *Excerpt 119, NNS2: Then I'm looking for the opinion, yeah, of course, if it's this choice question, then I see whether the student actually makes his or her choice.*

> *Excerpt 120, NNS28: as I understand it, TD she should answer a question, and choose only one option and then give arguments. So she failed to choose one of them, that's why I would give 1 for TD.*

> *Excerpt 121, NS40: The TD would be a 1 because she did not definitively chose a side between group studying and studying independently. I couldn't get a clear idea of how she would feel about that. I felt that there was precious time that was wasted.*

Idea quality. The last recurring pattern that can be attributed to Topic Development is the way the raters treated students' ideas. Some raters explicitly were looking for thoughtful reasoning and critical thinking, while other raters preferred personal examples. The next excerpts show thoughts verbalized by four raters in response to the same student (Recording 8). The first two raters (Excerpts 122 and 123) assessed the student's Topic Development as a 4 even though they ended up giving an overall holistic grade of 3. On the contrary, the last two raters (Excerpts 124 and 125) placed this students' ideas in band 2 just as their overall holistic grade. Although NS2 and NS30 also wanted to hear some thoughtful ideas, NS2 decided that that language is basic, but the response can be classified as a thoughtful one; whereas NS30 called the response basic.

> *Excerpt 122, NS2: what I particularly liked was personal examples that the person ... he gives some very basic information, like it's good because I can talk to the teacher and that is still a point that he made. And then he says I'm shy, I tend to keep quiet, or something like that. ..By and large, I get an impression that basic though the language may be, the answer is very thoughtful. The speaker comes across as someone who has really understood the question. He's not just giving an automatic answer but has really put thought into it.*

> *Excerpt 123, NS34: 4 in TD. There was very simple vocabulary used very well to express a variety of ideas. It was sustained throughout the entire idea.*

> *Excerpt 124, NNS35: Development was somewhat limited and lacked elaboration, vaguely expressed, and repetitions... Okay, I'm gonna give this person a 2 in TD. Some of the ideas were pretty vague and repeated.*

> *Excerpt 125, NS30: he had no trouble moving from one idea to the next. But I thought his topics were extremely general, basic, not thoughtful. They were a listing of points. He did not develop any one point.*

Based on the examples above, we can see that the raters interpreted the descriptor "well-developed" in different ways. The raters did not share the same understanding of what a good quality topic development is. Although several raters wanted to hear thoughtful ideas, only NS30 was aware of what they want to hear and knew that this was not part of the rubric. During the interview, NS30 admitted that she always looks for "ability to express abstract critical thinking higher level ideas" because they are indicators of students' language use ability (Excerpt 126).

> *Excerpt 126, NS30: I don't necessarily care about how much they say, I want to know what they are talking about. That's why I take notes. I just think that critical thinking is not necessary, but if I'm giving even a 3, I want some thoughtful response. And if I'm giving a 4, I definitely want to hear some critical thinking. I want to hear a point that is unique, or a point that is... just explained with really helpful details, something like that. It doesn't necessarily have to be amazing, but it has to be thoughtful and well-developed. I also care about... I don't want a list, to me that is very basic, if they just say for the following reasons, and they just list all three of them and they move to the reasons why it's not helpful. To me that is not developed. It's worth very little, so I am really in for development.*

Several raters reported that interesting ideas distract them from paying attention to the grading process since they get interested and forget that they are not just listening.

> *Excerpt 127, NNS46: Now I'm thinking that I could not quite recollect well LU. So there were some overall impression which was good. But I could not get exactly whether it was close to four or whether it was three or two. So I kind of skipped this when listening. And I was doubting in D as well. A bit doubting. That's why. I think that the main fault is this person's perfect TD, which I find perfect! [laughs] Because it confused me a lot. I thought 'wow, it's so good,' and I kind of forgot about the other two points. So, you can blame this person for this.*

In addition, some raters noted that new, fresh, or unique ideas make grading Topic Development complicated. For example, Excerpt 128 shows a rater who reflected on the way they approach scoring ideas that are new. The rater thought that they need to control whether they add to the grade only because they hear an idea for the first time.

> *Excerpt 128, NNS6: Maybe it makes sense for raters, it makes difference for raters when they hear something new and something interesting. For example, in this response, in this recording for the first time I heard some fresh new ideas to divide the classes according to the level and I though hmm that's interesting, that might be 4 just for the ideas. I don't think that it's fair actually because again like with the students when you hear it for the first time you think it's a plus but when you hear it again and again you get used to this and it's not fresh anymore and you don't add to the score. So again it's tricky. For the first time it sounds more interesting than for the second time and so on and so on and, in the beginning you add to the score and after that you don't and that isn't fair. That's what I felt.*

To summarize, non-rubric criteria that can be grouped under Topic Development criterion included: (a) raters' attitudes towards finished and unfinished responses, (b) utilization of writing-like organization, (c) prompt reading, (d) making an explicit choice when responding to prompts formulated as alternative questions, and (e) idea quality. The raters differed in respect to who employed one or all of these non-rubric criteria while making their decisions about students' responses. The examples of thoughts by different raters who listened to the same student but arrived at different scores illustrate that non-rubric criteria references affected students' scores.

Perceived severity

The way raters approach rating was also influenced by their perceptions of how lenient or severe they are. Raters can be aware of their grading strategy to apply more lenient or more severe patterns of scoring. In the interview, that was conducted after the think-aloud protocols, the raters were asked to reflect on their perceived severity of grading. Table 29 summarizes the raters' responses, where the first seven raters are NS and the last nine raters are NNS. Based on the descriptions in Table 29, most of the raters were able to categorize themselves as either more severe or more lenient raters, but some raters did not have an opinion or considered themselves average. First, perceived severity of NS raters is overviewed followed by NNS raters.

Three out of seven NS raters perceived themselves as lenient graders, one rater thought they are in the middle but more lenient, two believed that they are more severe, and one rater refrained from making a decision. The following examples show that NS30 in Excerpt 129 perceived themselves as a more sever rater, whereas NS7 in Excerpt 130 thought they tended to be more lenient.

> *Excerpt 129, NS30: Severe because I think when I doubt I go down. Also, I think I expressed too many times that I'm looking for critical thinking during my rating. I think it's a good indicator for other language things. I don't know if everyone does that, so yeah, I do consider myself more severe.*

> *Excerpt 130, NS7: Generally speaking, I think I try to err on the side of being generous. And you probably heard this in some of my responses. I felt like there was more. Maybe there was some effective filter through the test taking or you feel like if you were speaking with them conversationally, they would produce better language. So I think I try to see them compassionately, which makes me want to score a little bit higher.*

Table 29. *Raters’ Perceived Severity*

Rater	Perceived Severity
NS7	Lenient; Harsher on D and LU, lenient on TD
NS14	Middle but more lenient; Harsher on LU, lenient on TD and D
NS17	Harsh; lenient on D, harsh on TD, in-between on LU
NS30	Severe; harsher on TD and LU, lenient on D
NS34	Lenient; harsher on TD, lenient on D and LU
NS37	Neither, just fair
NS40	Lenient; more lenient on D
NNS2	Severe; more lenient on D, harsher on LU
NNS6	Lenient, wants to be stricter; lenient on LU
NNS10	Lenient; harsher on TD
NNS13	Severe; lenient on LU
NNS24	Middle, but more lenient; Harsher on LU
NNS28	Balanced, but more lenient; lenient on LU
NNS35	Balanced, but more lenient; lenient on LU, harsher on D
NNS45	Severe; no preference
NNS46	Severe; harsher on TD and D

Note. D stands for Delivery; LU stands for Language Use; TD stands for Topic Development.

Regarding NNS raters, two out of nine perceived themselves as more lenient, three raters thought they are in the middle but more lenient, four believed they are more severe. Excerpt 131 exemplifies NNS45 rater who described themselves as more severe, and Excerpt 132 shows NNS6 who thought they are more lenient but would like to become more severe.

> *Excerpt 131, NNS45: I think I'm more of a "bad cop" side. I'm trying to recall my decision-making process right now and I think that in most of the cases where I was not sure how to rule, I ruled not in favor of the student.*

> *Excerpt 132, NNS6: Of course, yeah, you're a human, you try to be fair, but it's not easy. I want to be harsher. You know, overall, I have the impression that I was more lenient, so I want to be harsher. I don't regret any of my decisions, but if somebody like experienced assessor tells me that you should be, of course, I would consider this advice. Of course, I would reconsider my assessment process.*

Overall, both NNS and NS raters perceived themselves as someone who can be inclined to award more lenient or more severe scores. It is important to see how these raters' perceived severity levels are reflected in their actual scoring patterns, which is addressed in the mixed methods results section.

Perceived rating criteria importance

Another factor that influences raters' approach to grading can be their perceptions of how critical each rubric rating criteria is. Such perceptions can indicate that raters make their decisions relying on one rubric category more than the other. As a result, such differences in perceptions can bring variance in raters' scores. In the interview, the raters were asked to reflect on their perceived rating criteria importance. Information in Table 30 summarizes raters' beliefs about criteria importance. Based on the raters' responses, most of them could see themselves treating some rubric criteria as more or less important.

Table 30. *Raters' Perceived Criteria Importance*

Rater	Perceived Criteria Importance
NS7	All important, but D is the most important since it influences the LU and TD. More attention to D and LU.
NS14	Equally important.
NS17	TD is the most important, then LU. D is the least important since so familiar with accents.
NS30	TD is the most important. LU goes hand in hand with TD. D is least important since can hear ideas through pronunciation.
NS34	Equal, TD is the most important. D and LU are not that important since still can understand.
NS37	Equally important.
NS40	TD is more important. D and LU a less important since can infer.
NNS2	Equally important.
NNS6	Equal, but LU is not as important as D and TD.
NNS10	TD is more important since the interlocuters need to understand the ideas.
NNS13	General description is important, all the rest are interconnected. TD is less important since not everyone can have excellent ideas on the spot.
NNS24	General description is important, all the rest are interconnected. LU is more important since it is a language test.
NNS28	All are important but TD is the most important.
NNS35	D and TD are more important since can disregard LU errors and understand the ideas
NNS45	Equal but TD is more important
NNS46	D is more important because if you do not understand the person you cannot assess LU and TD

Note. D stands for Delivery; LU stands for Language Use; TD stands for Topic Development.

Looking at the NS raters, four out of seven (NS17, NS30, NS34, NS40) believed that Topic Development criterion is the most important. These raters deemed Delivery and Language Use less important since the raters were familiar with various non-native speech patterns, therefore, they could understand ideas delivered with flaws in grammar and pronunciation. One rater, NS7, considered Delivery to be the most important category since it is the medium through which the other two categories can be assessed. Two NS (NS14, NS37) refrained from assigning more or less importance to any rating criterion. Excerpt 133 introduces thoughts by NS7 who ascribed more importance to Delivery. Excerpt 134 shows a rater explaining how their familiarity with non-native speech can prompt them to be more lenient on Language Use as the rubric does not have specific examples. Thus, the rubric permits the rater to draw their own line between what errors obscure the meaning or not.

> *Excerpt 133, NS7: I think D though might be the most important. Because, I think, for listening purposes, I think that if someone is more fluent, it's less choppy, it's intelligible and there's not a lot of pauses, I think it's just much easier to listen to. So, it influenced my scores on the others as well.*

> *Excerpt 134, NS34: I think I'm more lenient on language-use because the rubric says, "makes some errors and noticeable but does not obscure meaning". I've spent a number of years outside English-speaking countries, so I think I can accommodate vocab and grammar use and still understand people. This may make me more lenient, than someone who hasn't lived outside of English-speaking countries. I'm a strategic communicator. If I do not understand someone, I will try to understand someone, I will try to negotiate that communication.*

Turning to NNS raters, the patterns were not very straightforward. One rater, NNS2, considered all of the categories to be equally important. Similarly, NNS28 and NNS45 supported this opinion that all rubric categories have the same level of importance; however, they still allotted more weight to the Topic Development (Excerpt 135). In addition, NNS10 also viewed Topic Development as a more valuable criterion because idea exchange is the purpose of any kind of communication.

> *Excerpt 135, NNS45: I was trying to think of them as equal because that was the instruction to the test is. But, for me, the TD was the main one, I think. So, the basic point is that if the person did or not get that point across and then I could estimate how well they did do it, how much they did or did not elaborate linguistically. So, I think TD is like, how do you say, the first among the equal, yeah*

Furthermore, NNS46 shared the same opinion with NS7 mentioned before; they believed that Delivery is the number one in importance since flawed pronunciation prevents comprehension. Two categories were important for NNS6 and NNS35 – Delivery and Topic Development. Just like NNS46, these raters believed that poor pronunciation makes it impossible to see the ideas and language, but at the same time they thought that they do not pay much attention to language errors if the ideas are clear. Excerpt 136 demonstrates the opinion that Delivery and Topic Development have more importance.

> *Excerpt 136, NNS35: I think they have an equal importance, maybe LU is not as important as D and TD. If D sucks, then the rest just goes into trash. You can't even make sense of the topic, and you don't even notice the language they're using because you just have a hard time really hearing them…LU was a lot about mistakes, errors in their speech, but if I could still understand what they were saying, maybe I would disregard some of the mistakes that they were making. So, if I could see the topic being developed, and I could see they were making their points, even if they had mistakes in grammar, vocabulary, or whatever, I would probably not pay that much attention to those.*

Additionally, there were two raters, NNS13 and NNS24, who referred to the General Description as the most important category (Excerpt 137). To make it clear, General Description is the name of the TOEFL iBT rubric category that refers to the overall holistic score. For example, here is the description of band 3, "The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:". NNS13 and NNS24 believed that each answer is a system and that all the categories are interconnected to the extent that it is impossible to tease them apart as they are affecting each other. We can see that these raters wanted to rely on the more holistic description that encompasses all the categories together. These raters also used the top-down rating approach that was described before. However, a closer look at the rubric shows that General Description for all the bands does not comprise Language Use but only includes descriptors about the topic, ideas, and pronunciation. Based on this, an inference can be made that even though these two raters reported paying more attention to the General Description, they can be attributed to the raters who value Delivery and Topic Development based on what descriptors the General Description contains.

> *Excerpt 137, NNS24: For me they are not equally important, so maybe general description is most important for me, because you cannot judge from one or two categories. That's why, it is systematic, it sees the answer as a system. If the language is good, but the question is not for this person, he cannot show his knowledge, so the general impression is, so the general impression is always important for me. Cause it is bits and pieces put together, that's why in this case. Also, I think general description combines all the other three, that's why.*

To summarize, both NS and NNS raters tended to ascribe some weight to one or more criteria. The NS raters formed a more consistent pattern of treating the Topic Development category as the most important one. The NNS raters held more diverse patterns and mainly believed that all the criteria are important, but at the same time they showed more preference to Delivery and Topic Development. For the most part, both groups of raters showed a pattern of not mentioning the Language Use category among their primary criteria, whereas Topic Development and Delivery had more weight from the raters' perspective.

Summary

The results reported in this chapter have demonstrated that NS and NNS are following similar cognitive processes during rating. Based on raters' listening strategies, they were classified into the following types: note-takers, hard listeners, multitaskers, and re-listeners. In addition, the raters followed two main grading strategies: top-down and bottom-up. Ten top-down raters preferred to have a holistic idea before assigning partial grades, whereas three bottom-up raters preferred to assign the partial scores before calculating the holistic score, and two raters employed both approaches. Moreover, the raters showed the use of non-rubric criteria for Delivery and Topic Development due to their personal views or differences in their interpretations of rubric descriptors. Two most prominent non-rubric criteria for Delivery were voice quality and accent familiarity. Topic Development showed a variety of non-rubric criteria such as raters' views on finished and unfinished responses, organization, reading the prompt, making a choice, and idea quality. Finally, the raters' statistical and perceived severity tended to match and, in general, most of the raters considered Topic Development and Delivery to be more important rubric criteria than Language Use.

Mixed Methods Results

This section introduces the results from the side-by-side comparison analysis (Onwuegbuzie & Teddlie, 2003) of quantitative and qualitative data. This type of analysis helps to see how quantitative and qualitative data can complement each other and provide further insights. The first sub-section compares severity statistics from Facets rater measurement report for the 16 raters who participated in the qualitative inquiry with their perceived overall and criteria severity elicited during the interview. Second, the statistical values of rater consistency (infit statistics from Facets rater measurement report) for these 16 raters are overviewed in the light of raters' decision-making strategies in order to see if there are any recurring patterns distinguishing stable and unstable raters.

The results for each side-by-side comparison are presented for NS and NNS raters separately. First, comparisons are provided for seven NS raters followed by comparisons for nine NNS raters. Although the results are presented for each rater group independently, no group-specific patterns were prominent. Nevertheless, it is important to note that the present investigation had an exploratory nature, therefore, the limited numbers of raters in the NS and NNS group did not allow the researcher to clearly define any group-specific patterns. Thus, at the end of each result section, the findings are summarized for L2 speaking assessment raters in general, regardless of their NS or NNS status.

Statistical and Perceived Severity

Facets analyses indicated no radical or consistent differences between NS and NNS speakers regarding their severity. Based on the *t*-test, NS and NNS speakers did not show any significant differences in overall and criteria severity. To further understand rater cognition, raters' perceived severity elicited during the interview can be compared to raters' statistical severity measures. Table 30 shows 16 raters who participated in both quantitative and qualitative part of the study with seven NS raters listed first and followed by nine NNS raters. Their statistical severity measures in logits were obtained from different analyses: (a) one severity measure from the analysis when all the categories combined (i.e., Overall, Delivery, Language Use, Topic Development) and (b) four severity measures from the analyses for each category separately. In large, Table 30 shows that raters' perceived severity tended to match raters' statistical severity calculated when all the category scores were considered together, but no consistent matches were found comparing raters' statistical and perceived severity per category. More detailed comparisons for NS and NNS raters are provided below.

Examining the NS raters' general severity (when all the categories were analyzed together) and raters' perceived severity, we can see that their statistical and perceived severity matched for most of them (Table 31). For example, NS7, NS14, and NS40 perceived themselves as more lenient raters and showed lower severity logits, while NS17 and NS30 perceived themselves as more sever raters and showed higher severity logits. Statistical and perceived severity did not quite match for NS34 who perceived themselves as a lenient rater but showed more stricter rating patterns. NS37 did not have any opinion on their general and category severity; they showed more lenient scoring patterns than average.

Table 31. *Raters' Statistical and Perceived Severity*

Rater	Severity All scores	Severity O	Severity D	Severity LU	Severity TD	Perceived Severity
NS7*	-.41	-1.50	-1.89	-1.18	-1.24	Lenient; Harsher on D and LU, lenient on TD
NS14*	-.37	-1.38	-1.13	-1.47	-2.04	Middle but more lenient; Harsher on LU, lenient on TD and D
NS17*	.92	-.07	-.14	-.06	-.23	Harsh; lenient on D, harsh on TD, in-between on LU
NS30*	.60	-.40	-.03	-.62	-.71	Severe; harsher on TD and LU, lenient on D
NS34	.21	-.95	-.63	-.99	-.96	Lenient; harsher on TD, lenient on D and LU
NS37	-.15	-.71	-.58	-1.29	-1.91	Neither, just fair
NS40*	-.23	-1.17	-.99	-1.59	-1.72	Lenient; more lenient on D
NNS2*	.95	-.07	-.08	-.12	-.17	Severe; more lenient on D, harsher on LU
NNS6*	-1.17	-2.24	-2.61	-2.14	-2.39	Lenient, wants to be stricter; lenient on LU
NNS10*	-2.19	-3.45	-2.68	-3.40	-3.78	Lenient; harsher on TD
NNS13*	.14	-.51	-1.30	-.53	-1.29	Severe; lenient on LU
NNS24	.50	-.78	.07	-.73	-.90	Middle, but more lenient; Harsher on LU
NNS28*	-.14	-.94	-1.01	-1.05	-1.66	Balanced, but more lenient; lenient on LU
NNS35*	-.44	-1.47	-2.11	-1.73	-.97	Balanced, but more lenient; lenient on LU, harsher on D
NNS45*	.20	-.94	-.38	-.95	-1.38	Severe; no preference
NNS46	.04	-1.05	-.71	-1.03	-1.47	Severe; harsher on TD and D

Note. O = Overall; D = Delivery; LU = Language Use; TD = Topic Development; * indicates that raters' perceived and statistical severity matched.

Looking at the NNS raters' general severity (when all the criteria scores were considered together) and their perceived severity, we can see that their statistical and perceived severity also matched for many of them. For example, NNS2, NNS13, and NNS45 perceived themselves as severe raters and showed higher severity logits, while NNS6, NNS10, NNS28, and NNS35 tended to perceive themselves as rather lenient raters and actually showed statistically more lenient scoring patterns. In addition, NNS46 thought they are severe and were statistically on the severe side of the logits, but too close to the average that it should not be considered a complete match. The statistical and perceived severity did not match for NNS24 who was leaning towards seeing themselves as a more lenient scorer but was more severe from the statistical perspective.

In summary, raters' perceived and statistical severity matched for 12 out of 16 raters (5 NS and 7 NNS), did not match exactly for one rater, did not match for two raters, and could not be compared for one rater. Based on this evidence, we can infer that many raters were aware of their tendency to award more lenient or more severe scores. If this is the case and the raters are consciously being more lenient or less severe, the raters need to have more training in order to adjust their patterns to become closer to being statistically interchangeable.

Rater Consistency and Decision-Making Strategies

This section overviews consistent and inconsistent raters based on their cognitive processes while grading L2 speaking performance. First, consistent raters are described and then inconsistent.

Six raters who participated in think-aloud protocols and interviews were marked as consistent because their internal consistency measures (infit) did not exceed the 1.20 cut-off. They were three NS raters (NS14, NS30, NS34) and three NNS raters (NNS24, NNS35, NNS46). The first similarity that these raters had is that they did not look at the rubric while scoring. It should be admitted that NNS35, NNS23, and NS34 looked at the rubric several times while listening at the beginning, but then did not do that again. Moreover, NNS35 explicitly stated that trying to skim the rubric while listening was distracting and impeded comprehension. Additionally, NS14 and NS30 had a strong belief that skimming the rubric while listening is a disservice to the student because it prevents raters from paying full attention to examinees' speech. They believed that raters must listen attentively to what test-takers are saying and not multitask.

Another similarity among these raters was that they re-listened to unfamiliar accents (NNS35, NNS24, NS34, NS14) or tried hard to tune to speakers' speech quickly (NNS46). Only one of them, NS30, deemed re-listening unfair, but at the same time, this rater had extensive familiarity with all the student L1s in the study and did not experience any issues processing any accents. The third similarity is that none of the raters controlled for accent familiarity. Moreover, all three NS raters saw controlling for familiar accents unfair, to be a non-rubric criterion, disservice to the student, and something that can make them inconsistent.

In terms of grading strategies, four raters (NS14, NS30, NNS24, NNS46) utilized the top-down processing meaning that they thought about an overall preliminary grade or a range of grades before providing partial scores. Two raters (NS34, NNS35) used both top-down and bottom-up approaches, but employed the bottom-up approach infrequently, only in difficult situations when it was hard to decide on a grade. Regarding their non-rubric criteria references, the raters divided into two groups based on their favorable approach to Topic Development category, where NS14, NS30, and NNS35 were pro-logic, and NNS46, NNS24, and NS34 were pro-organization.

It is important to note that even though these six raters did not show any erratic scoring patterns, three of them showed some overly-consistent scoring (with Facets infit scores lower than .80). Lower infit values show that the raters either awarded the same scores to examinees of different ability or did not use the full scale. Looking at how these raters differed from the group can help to understand why this happened. For example, NNS46 was overly consistent (.69) for the Topic Development category, and they mentioned applying non-rubric criterion for the Topic

Development by lowering the scores for unfinished responses. In addition, NNS24 was overly consistent (.73) for the language use category. This raters' opinion about this speaking exam differed since they thought that a language test such as TOEFL primarily needs to value the Language Use category (described in Table 30). NS30 showed overly consistent judgments about topic development (.70), and it was noticed that this rater consistently applies a non-rubric criterion "critical thinking" to approach Topic Development scoring. Thus, it can be inferred that all three raters had more stricter expectations for each category for which they tended to show overly-consistent scoring patterns. These stricter rules that the raters tended to apply did not permit using higher scores for those criteria compared to other raters.

Turning to the inconsistent raters, there were five of them, namely NNS28, NNS13, NNS10, NS37, and NS7. Based on their cognitive processes while scoring, they formed three patterns – (a) all of them used the bottom-up approach to grading, (b) four of them engaged in multitasking (e.g., skimming the rubric) while listening to examinees' performance, and (c) four of them mentioned that they may add confidence as part of their Delivery scores.

Summary

This research question aimed to ascertain how qualitative and quantitative data complement each other. It looked at how statistical and perceived severity of raters matched and presented an overview of the decision-making strategies for consistent and inconsistent raters. The results showed that raters' statistical and perceived severity matched for the majority of the raters who participated in the qualitative part. This section also illustrated that inconsistent raters tended to multitask while listening to examinees' recordings, used a bottom-up grading approach, and added confidence as a criterion to rely on when grading students' Delivery.

Chapter 5: Discussion

The present study utilized a mixed methods approach to investigate variability in raters who scored L2 speaking performance. It is important to analyze rater variation since it can undermine the strength of the evaluation inference in the validity argument for speaking performance assessments. The evaluation inference is the building block of all other inferences (Kane, 2006), therefore, the flaws in the evaluation inference weaken all the subsequent inferences and, as a result, threaten the validity as a whole. Based on the common validity threats for the evaluation inference described in the literature (e.g., Crooks, Kane and Cohen, 1996; Xi, 2010), this dissertation looked at (a) raters' status, native (NS) versus non-native (NNS), (b) raters' accent familiarity, and (c) raters' cognitive processes while scoring L2 speaking performance to investigate how raters interact with the scoring criteria and what role NS/NNS status, cognitive processes, and accent familiarity play in this process.

This mixed methods study utilized different quantitative and qualitative data including scores and comments for 46 raters, as well as data for 16 raters from think-aloud protocols and interviews. To answer the first research question, the quantitative analyses of NS and NNS raters' scores and comments were used. For the second research question, raters' patterns of decision-making processes were described using the themes from both think-aloud protocols and interviews. The third research question looked at how quantitative and qualitative data complement each other.

In this study, two groups of raters NS (*n* = 23) and NNS (*n* = 23) scored speaking recordings by examinees from three L1 backgrounds Arabic (*n* = 25), Chinese (*n* = 25), and Russian (*n* = 25). Facets analyses were conducted in order to examine the severity and consistency of NS and NNS raters. Building on the statistical properties obtained using Facets, NS and NNS raters were compared in terms of their overall consistency, overall severity, rating criteria scoring consistency, and rating criteria difficulty. Furthermore, these 46 raters were asked to provide their comments for each participant they rated. Based on raters' comments that they provided while grading, the NS and NNS raters were compared regarding the directionality of their comments. Finally, 16 raters (NS = 7, NNS = 9), that represented a subset of those raters, who participated in the scoring described above, rated 12 recordings Arabic (*n* = 4), Chinese (*n* = 4), and Russian (*n* = 4) while thinking aloud and then answered interview questions. The themes from both qualitative sources were described to compare raters' cognitive processes while grading. Also, raters' statistical severity was compared to raters' perceived severity obtained during the interview, and raters' consistency measures were compared to raters' decision-making processes.

In this chapter, the results from all research questions are interpreted and discussed in relation to previous research where possible. First, the chapter discusses the findings from comparing NS and NNS raters. It is followed by a discussion of raters' cognitive processes. The chapter ends with a discussion of raters' accent familiarity.

NS and NNS raters

This sub-section discusses the differences between NS and NNS speakers. First, raters' consistency and severity are discussed. Next, raters' rubric criteria utilization are discussed. This section draws upon quantitative, qualitative, and mixed methods findings.

Consistency and severity. To compare the NS and NNS groups of raters, their consistency values and severity logit measures from Facets rater measurement report were compared. In general, no radical or consistent differences were found between NS and NNS raters regarding their consistency or severity. First, based on the *t*-tests, NS and NNS rater groups did not show any significant differences in consistency and severity. Furthermore, looking at the minimal effect size for consistency (Cohen's *d* = 0.09), the similarity of these two rater groups can be further supported. However, unlike rater consistency, rater severity showed a small effect size (Cohen's *d* = 0.34), which tells us that even though the *p*-values were not significant with the current group size (23 raters in each group), some differences could potentially be detected if larger rater groups were compared. It is possible that the NNS raters may show slightly more lenient average rating patterns due to the fact that they were 0.24 logits more lenient than NS. Additionally, when the severity patterns of NS and NNS raters were compared by examinee L1 group (i.e., Arabic, Chinese, and Russian), the NNS raters consistently showed lower severity measures grading each examinee L1 group meaning that they were more lenient raters regardless of the examinee L1 background. However, the tendency of NNS raters can be disputed since one rater from the NNS group, NNS10, can potentially be an outlier as their severity logit was -2.19, which is -1.03 logits higher than the severity logit of another lenient NNS rater. Even though NNS10 consistently maintained their severity across examinees, it can be argued that this rater was an outlier, an exceptionally lenient rater rather than average. If this is the case, the NNS raters might not show more lenient patterns of rating in further research. Furthermore, the NNS raters in this study scored examinees who shared their L1 and showed significantly more lenient patterns (this finding is further discussed in the next sub-section). The fact that they were more lenient towards the Russian examinees affected their overall severity patterns. Thus, it can be suggested that there would not be any significant differences between NS and NNS raters if the NNS group does not rate examinees with a shared L1.

The findings of this study that the NS and NNS rater groups did not differ in terms of their consistency and severity are in line with the studies by Brown (1995), Kim (2009), Xi and Mollaun (2009), Zhang and Elder (2011, 2014), and Wei and Llosa (2015), where the NS and NNS groups of raters did not show any differences according to Facets consistency and severity measures. This consistency of findings suggests that both NS and NNS speakers are suitable to evaluate non-native speaking performance. Thus, the study supports the inclusion of qualified, trained NNS as assessors for spoken exams as NNS raters are comparable to NS raters. Furthermore, in light of current globalization trends and moving towards English as an international language, inclusion of more NNS raters will allow assessment of English in a more comprehensive way as suggested by Hill (1996), Canagarajah (2006), and Gu and So (2015). Overall, the results of the present study suggest that the inclusion of more NNS as raters of oral performance will not pose a greater threat to test validity since NS and NNS groups of raters has consistently shown comparable rating performance in terms of rater consistency and severity.

The language assessment literature has consistently shown rater severity variation to be a recurring pattern across performance assessment. Thus, further exploration of rater severity was done for the subset of 16 raters (NS = 7, NNS = 9) raters, who participated in the qualitative part of the study. In an attempt to investigate whether the raters are aware of their severity patterns, their statistical severity (based on Facets outputs) and perceived severity (based on their answers to an interview question) were compared. The results indicated that for most raters, 12 out of 16 (NS = 5, NNS = 7), the perceived and statistical severity levels matched meaning that the raters who showed to be more severe from the statistical perspective also perceived themselves as more severe; and those raters who thought that they were more lenient were also lenient statistically.

There were no obvious patterns for NS and NNS groups; accordingly, this finding is discussed for all raters in general.

Based on the comparisons of raters' statistical and perceived severity, it can be suggested that the raters were aware of their underlying severity levels and were conscious about their tendency to award more lenient or more severe scores. Raters' awareness of their own rating patterns was speculated to be a type of bias by Winke and Gass (2013) who looked how accent familiarity of NS raters affects their scores. If a similar interpretation is applied to raters' severity, then it can be suggested that the raters were aware of their biased scoring, which in this case are more severe or lenient ratings. Since rater severity can be equated to construct-irrelevant variance that should not be part of examinees' scores, raters need to have more specific training to adjust their patterns in order to increase the chances to become statistically interchangeable. Since many testing companies use MFRM models to adjust their test scores for rater severity, consideration of rater severity variation is especially important for those testing companies that do not utilize MFRM models since large variation can significantly affect examinees' scores.

Rubric criteria utilization. The raters in the present study utilized TOEFL iBT independent speaking rubric to score examinees' speaking performance. This rubric had bands from 0 to 4 and four rubric criteria (i.e., Overall, Delivery, Language Use, and Topic Development). The scores for the Overall category for bands 1 through 3 were given based on the mode (the most frequent number), and a score of 4 was given when all three categories, Delivery, Language Use, and Topic Development, are given a 4.

According to the descriptive statistics from the Facets rating rubric criteria report, NS and NNS did not utilize the rubric criteria differently. The Delivery category was one of the most harshly scored criteria on the rubric. It was scored 0.06 logits more severely by the NS raters (0.17) than by NNS raters (0.11). Both rater groups scored the Language Use criteria with almost average severity and it was scored -0.01 logits more leniently by the NS raters (-0.07) than the NNS raters (-0.06). Also, the Topic Development category was the most leniently scored by both rater groups. The NNS group scored it less leniently (-0.17) than the NS group (-0.22). In addition, based on the infit measures, both rater groups utilized rating criteria stably. Unfortunately, these findings cannot be compared to some of the studies that used Facets to investigate the differences between NS and NNS raters (Kim, 2009; Zhang & Elder, 2011) since these studies used unguided holistic rubrics. In addition, even though Xi and Mollaun (2009, 2011) and Wei and Llosa (2015) used Facets and TOEFL iBT speaking rubrics, the results of this study cannot be compared since those studies used this rubric as a holistic one. Similarly, these results cannot be compared to Carey et al. (2010) as it looked only at pronunciation scores.

The finding that both groups of raters scored Delivery similarly cannot be easily compared to Brown's (1995) study where the rating criteria included Pronunciation and Fluency. In Brown's study, NNS raters were substantially harsher on the Pronunciation category (NS -0.37, NNS 0.60); nevertheless, both rater groups showed similar severity levels for Fluency with NNS speaker being more lenient (NS 0.02, NNS -0.13). The Delivery category used in the current study includes descriptors that can be categorized as both pronunciation (e.g., articulation, stress, intonation) and fluency (e.g., pacing, lapses, fluidity), therefore, cannot be directly compared to Brown's study. Lastly, Brown's study compared a limited number of NNS raters (*n* = 9) to a larger pool of NS raters (*n* = 24) and did not report category infit statistics.

It is also challenging to compare this study's results to Zhang and Elder (2014) who also compared NS and NNS raters. Their study utilized a rubric with the following criteria: Category

1 (Accuracy, Range), Category 2 (Size, Discourse Management), and Category 3 (Flexibility and Appropriacy). Based on the rating category descriptors, Accuracy focused on pronunciation, grammar, and vocabulary errors, whereas Range focused on grammar and vocabulary diversity, therefore, their Criteria 1 cannot be compared to Language Use category used in this study since it included pronunciation. Category 2 also cannot be compared to the Topic Development criteria used in the present study since Size was operationalized as "Size of contribution made by the candidate" (Zhang & Elder, 2014, p. 325), which was not part of Topic Development. Finally, Criteria 3 is not applicable to any of the rubric categories in the present study.

Overall, there is no study that resembles the current dissertation well enough in order to make direct comparisons of the findings. The findings of the present study suggest that NS and NNS raters were comparable in terms of criteria severity and consistency. Since the raters awarded similar difficulty values to each rubric criterion, it can be implied that the raters correctly interpreted the wider concepts of subdividing speaking performance into Delivery, Language Use, and Topic Development parts. This result also demonstrates not only the fact that both NS and NNS raters had a similar rubric interpretation but also that the rater training was effective in terms of explaining the rubric and how it worked for the raters.

The quantitative comparisons of NS and NNS raters suggested that they utilized the rubric criteria consistently and effectively, however, the quantitative analyses of raters' comments showed some variance. Each rater provided comments for each examinee scored; these comments were coded and counted. The coding scheme developed based on the scoring rubric allowed the researcher to classify raters' comments into three groups: General, Delivery, Language Use, and Topic Development. Raters' comments were coded in a way that the number of words in each comment did not matter as the coding was focused on the number of rubric features each rater attended to. For example, a lengthy comment “That's a nice response, the speaker doesn't come across as having any trouble expressing her ideas” was coded as two separate comments – one for General and one for Topic Development. At the same time, a shorter comment “pauses and grammar errors” was also coded as two separate comments – one for Delivery and one for Language Use. Such coding allowed comparisons not in terms of the number of words but based on the speaking performance features that the raters attended to.

Considering raters’ overall speaking performance feature attention (i.e., all the coded comments together), both groups of raters provided almost the exact number of comments (NS – 1674; NNS – 1648), which shows that both groups were actively involved in rating and able to provide similar level of feedback. This finding provides yet another piece of evidence to suggest that there were no differences between NS and NNS raters that adds to the previously mentioned statistical evidence about rater consistency and severity. This finding contradicts Zhang and Elder (2011), where 20 NNS located in China provided fewer comments (713) than 19 NS raters (935); and Kim (2009) where 12 Korean NNS raters left a significantly fewer number of comments, 1,172, compared to 12 Canadian NS speakers – 2,123. It is not easy to assess the comparability of the number of comments in Zhang and Elder since the number of raters was not equal. Furthermore, Zhang and Elder did not discuss the differences in overall counts of rater comments. On the other hand, Kim explained that the NNS raters gave fewer comments with the fact that “providing students with detailed evaluative comments on their performance is not as widely used in an EFL context as traditional fixed response assessment” (Kim, 2009, p. 204). Even though most of the NNS raters in this dissertation were EFL teachers, the study did not show such difference. It can be hypothesized that Chinese and Korean EFL contexts differ from the Russian EFL context, but yet another explanation is possible – the change over time since the studies were conducted approximately more than nine years (as it is unknown what year those studies were actually conducted in before publishing). Because a lot of time has passed, the trend of more focus on fixed response assessment might not anymore be true for EFL contexts in general. Due to the fact that both NS and NNS raters provided similar number of comments, it can be suggested that it is not uncommon to provide feedback on speaking performance in Russian ESL context. This inference can also be backed-up by the statements of NNS raters who participated in the qualitative part of this study as they sometimes mentioned giving written and oral feedback on their students' speaking performance. Overall, the fact that the NS and NNS raters in the present study provided a similar number of comments indicates no differences between NS and NNS raters in terms of overall attention to speaking performance features as measured by the total number of coded comments.

Comparing NS and NNS raters regarding their feature attention based on the comments for each coded category (i.e., General, Delivery, Language Use, and Topic Development), again, both groups of raters provided almost same number of comments per category. The comments for each coded criterion were counted and percentages from total number of comments per rater group were calculated in order to account for the slight count differences in total number of comments provided by each rater group. Delivery received 643 comments from the NS raters (38% from total number comments by NS group) and 637 from the NNS raters (39% from total number comments by NNS group); the NS raters left 312 comments for Language Use (19%) and the NNS raters left 350 (21%); for Topic Development, the NS raters typed 597 comments (36%) and the NNS raters typed 503 (31%); for General, the NS raters provided 122 comments (7%) and the NNS raters provided 158 (10%).

Based on the numbers above, when the NS and NNS rater groups were compared in terms of how much attention they paid to each coded comment category, the numbers came out to be similar. Both rater groups paid more attention to Delivery and Topic Development and less attention to students' performance on Language Use; they also provided the minimum amount of General comments, which refer to students' overall performance (e.g., good, poor). The themes from the qualitative data also back-up the hypothesis that Delivery and Topic Development were the most important criteria for the raters. Both NS and NNS raters expressed opinions that Topic Development is the most important because it is the purpose of any communication to exchange ideas. Delivery was also important because it is the medium of receiving the communicative message and can cause misunderstanding. On the contrary, the raters did not ascribe similar importance to Language Use since they found it easy to decipher awkward vocabulary and grammar or guess the meaning from basic words. Again, these opinions were shared by both NS and NNS raters. It can be implied that both rater groups had a similar interpretation of criteria importance when grading L2 speaking performance since their comments had almost the same percentage distribution per coded criteria type, and they expressed similar opinions based on the qualitative data. This finding is another piece of evidence that can be used to suggest that there were no differences in the way NS and NNS raters treated rating criteria importance.

Now, the results of the present study are compared to three other studies that looked at comments provided by NS and NNS raters who scored L2 speaking performance. The finding of the study that NS and NNS raters provided a similar number of comments per coded category contradicts the findings of Zhang and Elder (2011), where NS provided more comments for all coded categories (i.e., Fluency, Content, Interaction, Demeanor, Compensation Strategy, and Other General Comments) except for Linguistic Resources category. Comparing the counts, the researchers drew a conclusion that the differences were significant for all except for Fluency and Content categories. Zhang and Elder suggested that these differences serve as evidence of lesser comparability between NS and NNS speakers; however, Zhang and Elder did not normalize the counts and did not take into consideration the overall differences of the number of comments provided by rater groups and the differences in the number of people in each rater group.

Similarly, the results of the present study can hardly be contrasted with Kim (2009) as the study compared the raw counts for raters' comments that were illustrated in a bar graph. Thus, no direct comparisons can be made in terms of all 19 categories mentioned in Kim's study. However, Kim used percentages from total comparing the most salient comment categories. For example, the most salient category for the NS group was Language Use (13%) followed by Pronunciation (11%), Vocabulary (11%), Fluency (9%), and Specific Grammar Use (6%); whereas the percentages for the NNS raters were as follows: Pronunciation (15%), Vocabulary (14%), Intelligibility (7%), Overall Language Use (7%), and Coherence (5%). If these percentages are summed based to match the comment categories in the present study, Kim's NS raters emphasized Language Use (30%) and Delivery (20%) and NNS raters relied on Delivery (22%), Language Use (21%), and Topic Development (5%). Based on this scarce evidence that considers only about 50% of raters' comments, it can be suggested that the present study does not fully support Kim's findings since, unlike Delivery, Language Use was not among the most preferred rating categories for the raters in the present dissertation. Such difference can be explained by the fact that the raters in Kim's study did not have explicit training and used a holistic rubric with limited descriptors, which was justified by the purpose of the study (e.g., band 1 description was "Overall communication is generally unsuccessful; a great deal of listener effort is required).

The results of the present study can be somewhat compared to Zhang and Elder (2014). Even though their study used stimulated recall protocols and not comments while rating, the researchers provided rating criteria counts in percentages for the three criteria coded. Again, the present study cannot be directly compared due to the fact mentioned before (Criteria 1: Accuracy and Range included pronunciation); however, it is important to note that 38% or NNS raters' comments belonged to this category. Another important fact that needs attention is that the NS rater group in Zhang and Elder (2014) talked about features that could not be coded as referring to any of the rubric rating criteria (39%), which demonstrates NS raters' deviation from the rubric.

Further comparison of NS and NNS rater groups were made by ascribing each rater to a specific type based on the majority of their comments for each coded category (i.e., General, Delivery, Language Use, and Topic Development). This classification was applied in order to see if the raters within NS and NNS groups were more inclined to focus on a specific category on the person, not the average level. A rater was determined to be one-category oriented if the majority of the comments (50%) were devoted to one category and no other criteria received more than 35%. The raters were also classified into two-category oriented when the majority of their comments spread out between two categories, and raters whose comments were allotted to all the categories more or less evenly were marked balanced.

In terms of one-category oriented raters, there were two raters in the Language Use-oriented category (NS-0, NNS-2), one in the General-oriented category (NS-0, NNS-1), 13 Delivery-oriented raters (NS-3, NNS-10), and eight Topic Development-oriented raters (NS4, NNS-4). Regarding two-category oriented raters, there were 16 Delivery and Topic Development oriented raters (NS-13, NNS-3), two Delivery and Language Use oriented raters

(NS-1, NNS-1), and one Language Use and Topic Development oriented rater (NS-1, NNS-0). For the balanced type, there were six raters (NS-2, NNS-4). This finding is in-line with Brown (2000) May (2006) and Orr (2002) where raters preferred to attend to a set of salient response features, which probably were aligned with their perceived criteria importance and criteria harshness. Rater variation in terms of the rating criteria also corroborates findings of Vaughan (1991) who identified different essay reading styles. Vaughan identified reading styles such as "first-impression-dominates style" or the "grammar-oriented style" regardless of the similar rater training.

Based on these descriptive comparisons of personal rater types based on the criteria, there were some observable differences between NS and NNS groups of raters since more NNS raters tended to prioritize Delivery, whereas more NS emphasized both Delivery and Topic Development in their comments. These results suggest that NNS raters might treat the rubric category of Delivery as a more important one, and the NS can have a tendency to value more both Delivery and Topic Development. Taking into account the fact that there were not many Language Use oriented raters and the fact that Language Use received fewer comments by both rater groups (NS-19%, NNS-21%) compared to Delivery and Topic Development comments which had more than 30% in each rater group (see the previous section), it is suggested that Language Use category had less weight or importance when both NS and NNS raters evaluated L2 speaking performance. To reiterate, the similar pattern was observed in the qualitative data produced by a sub-set of raters. They thought that it is not that difficult to infer meaning even if the grammar and vocabulary were not the best. On the contrary, the raters believed that Topic Development and Delivery are more important for spoken communication because they represent the ideas and the medium of exchanging them.

Overall, the descriptive comparison of NS and NNS rater groups did not show any quantitative differences in terms if the number of comments provided. However, these comment counts should be interpreted with caution since examinees' spoken proficiency and raters' negative/positive directionality were not considered when the raters' comments were counted. It is possible that the raters left more negative comments regarding Delivery or Topic Development for lower-level examinees and fewer positive comments about Delivery or Topic Development of higher-ability examinees. These variables could have affected the dispersion of raters' comments across the categories.

Raters' Cognitive Processes

The present study looked at raters' cognitive processes while grading L2 speaking performance. To investigate them, 16 raters (NS = 7, NNS = 9) individually participated in think-aloud protocols followed by interviews. These raters were selected from the 46 raters who participated in the quantitative part because they showed either consistent or inconsistent scoring patterns based on infit statistics from the Facets measurement report. Content analysis was used to identify raters' patterns of decision-making in terms of their listening and grading strategies, non-rubric criteria references, perceived severity, and perceived category importance. To clarify, the listening strategies were defined as what the raters did during the time a student's recording was playing, and grading strategies referred to raters' decision-making processes after listening to a student's recording. Non-rubric criteria references included raters' references to additional features that characterize examinees' speaking performance that were not on the rubric. Perceived severity referred to raters' opinions about themselves as lenient or severe raters elicited during the interview. Criteria importance described the importance that the raters ascribed to a rubric criterion based on raters' responses to an interview question. In large, there were no explicit strategic differences between NS and NNS group rater. The raters utilized a variety of strategies during scoring, and they differed from each other because of their individual differences and not due to NS or NNS rater group affiliation.

The qualitative data showed that raters employed a number of listening strategies while listening to examinees' recordings. Based on the patterns, the raters were sub-divided into note-takers, hard listeners, re-listeners, and multitaskers. The raters did not form any particular NS or NNS groups. There was always a justification for each strategy used. For example, rater type *note-takers* argued that it helps to be an active listener and not to rely on working memory, whereas rater type *hard listeners* favored simple attentive listening, sometimes with eyes closed in order to give examinees full attention and not to be distracted. It is questionable which strategy is better. Based on the mixed methods analyses of rater consistency and listening strategies, note-taking did not form any patterns, both consistent and inconsistent raters did or did not do this. It is suggested that raters can choose which strategy to follow, of a note-taker or a hard-listener. Looking at the re-listeners and multitaskers, the data showed that many consistent raters re-listened to students' recordings, but the multitaskers tended to be inconsistent raters. It is suggested that reading the rubric should be monitored closer in future research. In addition, differences in individual listening strategies can affect how fast raters get tired. It is possible that raters who take extensive notes can show signs of rater fatigue faster. Thus, future research should address how rater' listening strategies are related to rater fatigue.

Regarding grading strategies, two prominent approaches were top-down (holistic grade first and then specific grades) and bottom-up (specific grades were used to calculate the holistic grade). Usually, 12 raters who utilized the top-down approach formed their initial opinion while listening. Three raters who utilized bottom-up approach also thought about partial scores during listening. Out of 12 top-down raters, two sometimes referred to using the bottom-up approach if they found it difficult to decide on a holistic score. These findings match the results in Pollitt and Murray (1996), who examined oral interviews and arrived at two approaches used by raters: intuitive and analytical. In addition, the fact that two raters used both approaches corroborate the findings provided by Ang-Aw and Goh (2011), who also traced the same two patterns, intuitive and analytical, but also added the mixed one. These findings are also in line with research on rater behavior (Joe, 2008; Joe, Harmes, & Hickerson, 2011) which classifies raters into analytic and holistic. Additionally, previous research has associated holistic judgments with experienced raters who made intuitive judgements based on the internalized rubric, whereas novice raters had more analytic approaches. The present study is in line with these findings as it makes a connection between raters' approach and raters' infit stability measures since all the inconsistent raters used the bottom-up approach to grading.

Non- rubric criteria showed patterns of raters who used non-rubric criteria for delivery and for topic development. In general, such presence of non-rubric criteria shows that raters are susceptible to the non-rubric criteria coming from their individual differences such as personality, beliefs, personal reference standards, professional and academic background, which goes along with findings of other studies (Brown, 2000; Brown, Iwashita, & McNamara, 2005; Joe, Harmes, & Hickerson, 2011; Kim, 2015; Wei & Llosa, 2015; Winke, Gass, & Myford, 2012).

For Delivery, the non-rubric criteria themes included accent familiarity and voice quality. Being a multifaceted phenomenon, accent familiarity is discussed in the next section. In terms of voice quality, the raters reported that they attended to how confident or not examinees are. Usually, softer or quieter voices were equated with lack of confidence. In addition, the raters struggled with applying the notion of "listener effort" for quieter examinees or examinees whose recording quality was not the best due to students' breathing into the microphone. Some raters could easily discern what caused them to strain their ears; others were hesitant because they needed to decide if that was the quieter voice, recording quality, or actual examinees' performance that caused listener effort. Voice quality also encompassed raters' references to how pleasing or off-putting examinees voices are. Several raters used the words "pleasing, friendly, and endearing" to describe examinees' voices; this might mean that personal attitudes can also make a rater prone to be more or less lenient on an examinee depending on how they perceive their voice. One rater mentioned that a very unnatural sounding voice made them want to rate that examinee lower. The present study is not the first one to highlight the fact that raters pay attention to examinees' voice quality. Raters' references to voice quality can also be seen in the literature, but for the most part, such comments were coded as rare, for example, "comments that occurred fewer than 20 times were excluded as categories (e.g., "low volume," "soft voice," "little confidence," "poor time management," and so on)." (Kim, 2005, p.51). On the other hand, Zhang and Elder (2011) coded them as Demeanor that included confidence, lack of confidence, nervousness, shyness, quietness, maturity, sense of humor. Since raters' references to test-takers' voice quality is a construct-irrelevant factor, further research should be done to see to what extent students' voice quality can impact their scores.

Now, the themes of raters' non-rubric criteria for Topic Development are discussed. This type of non-rubric criteria was mentioned more than the quality of examinees' voice but less than accent familiarity. The NS and NNS raters did not show any specific patterns. Overall, the raters had different thoughts about idea quality and various interpretations of the descriptors such as "coherent", "well-developed," and "logical".

Not all the raters had a similar interpretation of what kind of ideas can or cannot be given higher scores. Some raters thought that the presentence of personal examples is enough to gauge students' L2 proficiency. On the other hand, one rater, NS30, continuously reiterated that critical thinking ideas are the best descriptors of students' language ability. This finding is parallel to Brown (2007) who stated that some raters might believe that "maturity of ideas" is relevant to students 'university success. In addition, raters' decisions for the Topic Development category were contaminated by their attitude towards finished and unfinished responses. Some raters punished students for the inability to finish before the time runs out and others do not lower the grades. One rater used this reasoning and justified it with the fact that the rubric description "coherent" explains their actions since a response that is not wrapped-up on time is not coherent. Not all the raters shared the same understanding of unfinished responses. Some raters valued that students were willing and able to talk more than the time allowed; they thought that this is what shows students' potential.

In addition, raters had a differing understanding of what "relationships between ideas are clear" means. Some raters (pro- organization) thought about more essay-organized information (introduction, body, conclusion), whereas others (pro-logic) raters believed that speaking should not be like writing and that people do not speak with using such organization in real-life. In addition, those raters valued the logically organized responses more since they considered introduction and conclusion "canned" phrases that anyone could memorize and use for any response. Overall, Topic Development descriptor interpretations have shown a variety of ways that raters can understand "coherent," "well-developed," and even "logical". For instance, two NNS raters (who showed more lenient rating patterns) mentioned the rhetorical organization as a factor that can play a role in understanding the logic behind examinees' responses. These raters noted that what can be absolutely logical sequence in one language, is not a logical sequence in English, and the raters were trying to answer the question, “Is it me or them?”. These raters’ thoughts are supported in the literature. Wei and Llosa (2015) described that Indian raters were better than American raters at understanding rhetorical features of Indian examinees.

In conclusion, more attention should be paid to how raters make their decisions, specifically to how they interpret rubric descriptors and if raters’ interpretations can bring any systematic or irrelevant variance. The present study did not show any differences in rater decision-making patterns due to NS/NNS group affiliation. As discussed above, the raters showed variation because of individual differences. Previous research that focused on rater variation among the NS raters also showed strategic rater differences in writing (Eckes 2008, 2012; Vaughan, 1991) and speaking (Ang-Aw & Goh, 2011; Orr, 2002). It is suggested that NS and NNS speakers can vary within their group in terms of criteria interpretation and the standards they apply to each sub-scoring criteria, therefore, differ in score assignment.

Raters’ accent familiarity

The present study also looked at raters’ accent familiarity as it can cause unexpected rater variability. For this purpose, the study chose specific L1 accents that were hypothesized to be more familiar or less familiar for the NS (North American) and NNS (Russian) raters. The examinee L1 backgrounds chosen were Arabic and Chinese (more familiar for NS and less for NNS) and Russian (more familiar to NNS and less to NS). Arabic and Chinese L1s are familiar to NS raters due to the fact that they represent common L1s in academic context in North America, but are almost not represented in academic contexts in Russia. In addition, the inclusion of Russian raters and examinees allowed investigating the effects of L1 match between Russian examinees and Russian raters. To assess raters’ familiarity, the study collected raters’ self-reported familiarity scores (familiarity with L1 identification) and raters' familiarity scores after they listened to 24 short excerpts by examinees (familiarity without L1 identification). Familiarity with L1 identification means that the raters knew what L1 they are reporting their familiarity for; without L1 identification means that the raters were unaware of L1 background of the speakers that they reported their familiarity for. Accent familiarity was collected in these ways to have a better understanding of raters' accent familiarity. As the data showed, accent familiarity with L1 identification and without L1 identification showed similar patterns but were not exactly the same which will be further discussed below.

Accent familiarity of NS and NNS. First, raters' familiarity is overviewed in terms of NS and NNS rater groups. As hypothesized, the NS and NNS differed, on both measures, in their accent familiarity of examinee L1s due to their experiences. NS raters were more familiar with Arabic and Chinese accents than NNS raters and NNS raters were more familiar with Russian accent than NS. However, both groups showed differences in numbers for accent familiarity reported with and without L1 identification. When percentages were compared, only NS raters' familiarity for Chinese test-takers did not change a lot (86% with identification and 82% without identification). The NS raters showed a lower familiarity with the Arabic L1 (changed from 83% to 76%) and a greater familiarity with the Russian L1 (from 60% to 69%); the NNS participants revealed a much higher familiarity with the Arabic speakers (from 42% to 59%) and Chinese (from 44% to 56%), while a lower familiarity with the Russian L1 (from 95% to 81%). Based on these fluctuations, it was hypothesized that not all the raters could recognize speakers' L1s when reporting familiarity without L1 identification, especially for Russian and Arabic L1s, and that NS raters were better at identifying Chinese speakers than NNS raters.

To discuss this probability that not all raters can identify L1 of familiar and unfamiliar examinees, qualitative findings can be used. During the think-aloud some NS and NNS raters guessed examinee L1s correctly or incorrectly and one of the interview questions asked the raters to reflect on their ability to distinguish L1s used in the study (it is important to note that no questions were asked about examinee L1 during the think-aloud and the interview questions about examinee L1s were the last). Five out of seven NS raters were highly familiar with Arabic and Chinese and they reported that they were able to distinguish Russians because they differed from the two they know; however, this does not mean that they could identify that those speakers were Russian. In addition, one NS, who was highly familiar with Arabic pronunciation, reported labeling Russian examinees as Arabic. One NS speaker who was not very familiar with all the accents used in the study, during the think-aloud, identified several examinees' hypothetical L1 background as Asian when listening to a Chinese student and Latin-based when verbalizing thoughts an Arabic recording. In terms of NNS raters, two raters thought that Arabic examinees are Indian, one rater identified one Arabic examinee as Russian and a Chinese speaker as Spanish or Latin American. Only one NNS rater reported their ability to identify all three L1s due to familiarity. One NNS mentioned noticing Slavic examinees clear delivery. Four NNS raters guessed one or two Russian examinees during the think-aloud. The raters' guesses were consistently made about the same two examinee recordings; however, the raters did not mention anything else about the other two Russian recordings used in the study. One rater mentioned that a typical introduction "Today our theme is" prompted them to guess it is Russian. All NS raters were able to identify Chinese speakers as speakers with "Asian background" and so did many NNS raters as well.

Looking at the quantitative and qualitative findings together, it can be seen that some NS and NNS reported their accent familiarity for what they thought was Russian while listening to Arabic speakers and vice versa, which caused both NS and NNS raters' familiarity with these L1s fluctuate when it was reported with and without L1 identification. In addition, NNS raters' familiarity with Arabic speakers could have fluctuated because they thought they were reporting their familiarity with Indian, not Arabic L1. Similarly, qualitative findings support the quantitative suggestion that NS raters could identify Chinese speakers without much difficulty, therefore, their accent familiarity remained stable. Also, it can be suggested that NNS raters' accent familiarity for Russian L1 that was reported without L1 identification decreased because they might have thought they were reporting their familiarity for a different L1 background. Overall, the qualitative findings help us explain why NS and NNS raters' familiarity fluctuated when reported with and without L1 identification.

The fact that not all the raters could identify examinees' L1s correctly might have also caused some variation in their scores if the raters had some predisposed biases; however, it is just a hypothesis. For example, one Russian NNS rater in the study reported that it was very easy for them to grade a recording and justified it by their accent familiarity with examinee L1 only because they thought the student was Russian, not Arabic as the student really was. Similarly, the same effect could have been present in ratings by a NS rater who had extensive familiarity with Arabic speakers; however, did not distinguish Russian and Arabic speakers. Both cases exemplify an interesting effect of raters' familiarity when there was no actual accent familiarity.

This finding also needs to be addressed when operationalizing accent familiarity, which still has not been clearly defined in the literature. So far, accent familiarity in speaking assessment studies was determined by various factors, for example, shared L1 (Xi & Mollaun, 2009), shared L2 (Winke et al., 2013), length of residence (Carey et al., 2011), or teaching experiences (Huang, 2013). The fact that not all raters can identify examinees' L1 correctly can influence raters' exhibited accent familiarity, which may cause this exhibited accent familiarity not to match their predetermined accent familiarity.

Familiar and unfamiliar groups of raters. Now, accent familiarity is discussed based on the Facets results when the raters were grouped as familiar and unfamiliar regardless of their native-speaker status. To group the speakers as familiar or unfamiliar, their composite familiarity scores (with L1 and without L1 identification) for each L1 were summed and the raters were sorted. Next, the raters who marked their familiarity as "A Lot" or "Extensive" were added to the familiar group and the rest formed the unfamiliar group; this was done for each L1 separately. Then, these two rater groups were added to added to Facets. Facets rater measurement report did not show any differences between familiar and unfamiliar raters added to the model. From the quantitative perspective, it can suggest that raters' accent familiarity does not play a role; however, qualitative findings can help to interpret this fact.

First, we should keep in mind that not all the raters were able to identify examinee L1s and their exhibited and expected familiarity could have differed. In addition, qualitative findings show that NNS raters were trying to control for the lack of accent familiarity and how NS speakers were trying or not trying to control for extensive accent familiarity. NNS raters' familiarity with Russian L1 will be discussed later (L1 Match). NNS who were unfamiliar with Arabic and Chinese L1s were trying to control for the lack of this familiarity. They were trying to guess and infer what speakers from unfamiliar L1 backgrounds were saying. To give a more specific example, one NNS speaker mentioned that even though she knew she was not making a real TOEFL decision, she felt responsible as this is a life-changing test. In other words, the raters did not want to negatively affect examinees' lives only due to raters' lack of familiarity with examinees' accent. In addition, even though some NNS raters posed a question, "Is it me or is it them?", many of them questioned their ability to understand unfamiliar accents rather than making a conclusion that the person is unintelligible. NNS raters' re-listening patterns, they tended to listen twice, close their eyes while listening, and listen very carefully to make sure they understood the person when they were unfamiliar with the accent. These raters reported that the first listen allowed them to tune to the pronunciation and they employed inferencing and guessing strategies during the second listen to understand examinees' ideas. The majority of the NNS raters stated that they would like to receive more training on unfamiliar accents such as an explanation of typical phonological specificities, grammar, and vocabulary in order to become more impartial scorers.

Similarly, several NS raters, who were familiar with the accent they were grading, tended to re-listen to make sure that they are not affected by pronunciation specificities. Many of the NS raters specified that they are aware of their familiarity and would not like to control for it as it can cause instability of their ratings. In addition, these NS raters mentioned that biased are the ones who do not have the familiarity, not the ones who have it. They argued that controlling for the greater familiarity promotes pronunciation stigma. However, not all NS shared the same opinion, one NS rater explicitly stated that they controlled for their extensive familiarity trying to listen to recordings without an ESL hat and more like a random person in the street. Another NS rater noted that a random person in the street who does not try hard to understand a non-native speaker is just simply rude.

Even though the quantitative findings prompted to make a decision that accent familiarity does not affect raters' scores, the qualitative findings show the depth and breadth of the accent familiarity phenomenon. This finding should also be considered when operationalizing accent familiarity since unfamiliar and familiar raters can be aware of their familiarity and controlling for it. In addition, this fact should also be seen through the lens of specific rating criteria, for example, one of the descriptors on the TOEFL iBT rubric for Delivery is “listener effort”. As one of the NS raters mentioned, the presence of this descriptor factors in rater-specific biases into scoring since this is a rater- not speaker-dependent descriptor, meaning that the students’ scores can fluctuate based on who listens to their performance. Finally, unfamiliar NNS raters in the study expressed their willingness to learn more about unfamiliar accents, and most of the familiar NS raters did not want to control for their familiarity since they considered themselves unbiased because of this familiarity. Such opinions reflect some trends in L2 speaking assessment research. For example, Xi and Mollaun (2011), Carey et al. (2011), and Wei and Llosa (2015) advocated for exposing raters towards language varieties. Specifically, Wei and Llosa (2015) suggested that variety speakers can develop training materials for unfamiliar raters.

L1 match. Another facet of accent familiarity is when it is interpreted as L1 match between examinees and raters. The present study looked at L1 match from the statistical perspective by comparing Facets measurement reports for NS and NNS raters based on ratings per examinee L1.

Based on Facets severity means, the NNS rater group was more lenient across L1s; however, it exhibited statistically significant lenient ratings when scoring Russian L1 (-.54 logits, Cohen’s $d = 0.65$). This finding reveals that Russian NNS raters tended to exhibit positive bias when grading examinees who shared their L1. At the same time, these statistics can be interpreted in a different way. Based on Facets severity means, the NS rater group was more severe across L1s; however, it exhibited statistically significant severe ratings when scoring

Russian L1 (.54 logits, Cohen's *d* = 0.65). This finding reveals that NS raters tended to exhibit negative bias when grading examinees from a relatively unfamiliar L1 background. Now, turning to qualitative results, this finding is interpreted.

The majority of NNS raters mentioned that it was easy for them to score Russian examinees since pronunciation did not preclude understanding of the message. Some raters admitted the fact that they could have exhibited an unconscious bias towards Russian L1 students since comprehension was never a problem. One NNS rater expressed the thought that Russians should not rate Russians due to extensive familiarity with the pronunciation, language patterns, and culture. Another rater admitted that they wanted to give the higher Delivery scores for Slavic speakers, but controlled for that. On the other hand, the NS speakers did not express any difficulties rating Russian L1 speakers during the think-aloud and did not express any concerns during the interviews. However, we should also keep in mind that it is not clear how well they could distinguish Russian students from others. Nevertheless, based on the integration of quantitative and qualitative findings, it can be suggested that the NNS exhibited more lenient patterns of scoring towards Russian L1 examinees.

To further discuss why the NNS raters awarded more lenient scores to the examinees with shared L1, we can draw on the comments by the NNS raters about the clarity of Delivery of Russian speakers and the presence of "listener effort" descriptor in the Delivery category. The NNS raters noted that it was very easy for them to understand the typical phonological inconsistencies of Russian speakers. Now, looking at the rubric, it can be seen that the rater effort is described for three first bands and not mentioned for band 4; description for band 3 hedges listener effort with "may," for band 2 it is "needed," for band 1 it is "considerable". Based on the results for the raters' cognitive processes, let us imagine this situation: A NNS speaker listens to a Russian speaker and starts estimating their overall performance as something around 2 or 3. The rater skims Delivery section for band 2 that says, “Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.” Then the rater can certainly agree with the first statement but then rejects all other five statements because “listener effort” modifies all of them. Thus, this rater is more likely to move a band higher to see, “Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times though overall intelligibility is not significantly affected.” Now, there are more statements to agree or disagree with. This rater agrees with the first statement and, for example, disagrees with the second one about fluidity of expression, but the rest of the description is again modified by “listener effort”. In large, due to the fact that this rater did not experience any listener effort, they disagreed with five statements in band 2, moved on to band 3, and disagreed with only one statement, which gives the rater an impression that band 3 is a better description. Thus, this rater is more likely to decide to give a 3 even though there was no fluidity of expression, which, in turn, affects all the next grading steps of this rater. A higher score for Delivery can lead to higher scores on other categories.

Additional insights about why NNS in this study exhibited more lenient patterns of grading towards examinees with a shared L1 background can be drawn from the results of subdividing raters into types based on their comments. In other words, NNS raters could have awarded more lenient scores based on their interactions with the rating criteria. The results showed that compared to three NS raters, 10 NNS raters had the majority of their comments about Delivery category. This can be interpreted as Delivery being a more salient or important category for NNS raters. Then a possible explanation can be as follows: A rater, who considers Delivery an important category, scores this category higher due to the absence of “listener effort” and, therefore, biases all the other criteria score decisions resulting in a higher holistic score. This offers a possible explanation of why NNS raters showed a tendency of scoring Russian L1 speakers higher. Thus, such interaction of raters and the rating criteria needs to be addressed by rubric revisions and rater training.

Chapter 6: Conclusion

This chapter includes the implications of the findings of this study, limitations with directions for future research, and concluding remarks. First, the implications of the study are presented including methodological and practical. Next, the limitations of the study are acknowledged and followed by directions for future research. Lastly, the dissertation ends with concluding remarks.

Implications of the Study

The results of this mixed methods research study have several methodological and practical implications for the field of L2 speaking performance assessment. This study sought out the sources of variability for raters with respect to their native speaker status (NS/NNS), cognitive processes while scoring, and accent familiarity with examinees' L1s. The study broadens the understanding of rater variation that can have an effect on raters' ability to rate accurately, consistently, and with a uniform severity level. This study also adds to the accumulating body of literature on rater variation and bias in assessment contexts and provides valuable insights.

Methodological. This study offers the following methodological implications. First, the study employed a mixed methods design to compare the quantitative and qualitative findings that complement each other and provide a deeper and broader perspective on rater variation. The study showed that qualitative findings provided a better picture of raters' behavior even when there was no statistical effect. In addition, unlike previous studies that researched NS and NNS raters, L1 match, accent familiarity, and rater types that relied on one or two sources of data, the present dissertation broadened the scope of data analyzed to include raters' scores, comments, think-aloud protocols, and interviews. This study offers an example of how multiple sources of information provide a better picture on rater variation.

Second, this dissertation, grounded in L2 assessment research, contributes to the limited number of studies in the L2 assessment literature that explored the effects of accent familiarity on raters. Unlike speech perception research on accent familiarity that usually uses short speech excerpts, untrained raters, and unguided scales, the present study simulated real-life assessment practices, namely recruited raters with ESL/EFL teaching experience who were skilled enough to qualify to be TOEFL raters, provided individual rater training, utilized a well-established TOEFL iBT speaking rubric, and used 1-minute long test-takers' responses. Future studies can also replicate research undertaken in the realm of speech perception by modifying the procedures in order to follow the steps specific to L2 speaking assessment in order to provide a better vision of the role of accent familiarity in the field of L2 speaking assessment.

Third, the study offers additional insights into operationalizing accent familiarity. The study showed that not all NS and NNS raters of L2 speaking can distinguish examinees' L1 accents. Moreover, the study illustrated that some raters tried to control for the lack of accent familiarity by re-listening to recordings and guessing words from context. Additionally, it was described that some raters who have extensive accent familiarity also tried to control for it and listen to test-takers' answers as people who are not that familiar. Thus, raters' accent familiarity in the present study was shown to be not a clear-cut phenomenon, which can be affected by raters' ability to recognize accents as well as raters' conscious attempts to control for their familiarity or unfamiliarity. Future studies in L2 speaking assessment and speech perception can take these findings into account when operationalizing accent familiarity, intelligibility benefits, and L1 match benefits for research purposes.

Finally, the present dissertation is one of the few studies that looked at raters through a cognitive approach. The study demonstrated that the raters differed in how they arrived at scores, what they did while listening, how they interpreted the rubric, and whether or not they relied on various non-rubric criteria. Since it is crucial for the validity and fairness of performance assessments to know how raters approach rating and interact with the rubric and speech samples, future studies can use the findings presented in the dissertation to further explore the effects of raters' behavior on examinees' scores.

Practical. From the L2 speaking assessment standpoint, the study offers several practical implications for testing companies, language programs, and teachers who are involved in L2 speaking assessment. The implications contain the inclusion of more NNS speakers as raters, guidelines for rubric development, suggestions for rater training of L2 speaking assessors, and advice for teaching test preparation courses.

For rater recruitment, the study provides backing to the fact that proficient and experienced NNS can exhibit severity and consistency levels comparable to NS raters, which means that NNS can be used for scoring speaking exams. In light of current globalization trends and movement towards English as an international language, inclusion of more NNS raters will allow assessment of English as a global language. Even though there were detected L1 match differences, the study still argues for the inclusion of NNS as raters, but suggests providing more specific training for NNS raters who can be prone to exhibit some degree of positive bias. The special training can include more comprehensive guidelines for NNS raters about how to approach scoring examinees with shared L1.

The results of this study illustrated that raters' interpretations of the rating scale could affect the scores they assign. Specifically, the most widely-interpreted scoring categories were

Delivery and Topic Development. Based on this finding, suggestions can be made for future rubric development. It is advisable to provide explicit interpretations for the descriptors used in the scoring rubric. For example, descriptors such as well-developed ideas, coherent speech, and listener effort can be defined and illustrated with examples.

In addition, the aforementioned findings about how raters interact with the scoring rubric, guidelines for improving rater training materials can be outlined. One suggestion can be to train raters using additional materials with more fine-grained, specific descriptions of each rubric band and category in order to provide the raters with a framework of correct interpretations of short statements on the scoring rubric. For example, the additional training materials can provide the following extended descriptions for interpreting "well-developed ideas" descriptor correctly: "*Well-developed ideas* descriptor on the rubric means that a speaker has one or two specific reasons to support their opinion on the topic (e.g., I prefer preparing for exams alone because I can focus better on the materials and work more efficiently) and elaborates each reason with one or two examples or details (e.g., I can concentrate better because no other people distract me from studying by asking questions). A well-developed idea can be developing only one side of the argument based on advantages (e.g., I prefer to study in a group because it is faster) or disadvantages (e.g., I do not like studying in groups because it is time-consuming) as well as elaborating on advantages or disadvantages of both choices (e.g., I think that both approaches to studying can be beneficial because …). Well-developed ideas can have a set structure (e.g., introduction, body conclusion) or be logically connected to each other. Responses that provide more detailed explanations of only one reason are equal to responses that have two less extensively elaborated reasons." Moreover, it can be even more beneficial to have several annotated scripts of students' answers that illustrate general rubric descriptors. Additional training materials can help raters to achieve a similar framework of reference and, in turn, can limit rater variation. However, due to the fact that long and detailed scoring rubric guidelines can overload raters' cognition, the usual condensed descriptors can be used while scoring. Overall, it is important to ensure that raters achieve similarity in interpretation of rubric descriptors in order to avoid threats to tests' validity and fairness.

Another suggestion for rubric development can be based on the finding that raters had dissimilar opinions about rubric category importance. Rubric developers should explicitly mention if any analytic criteria possess greater weight in determining the final holistic score. For example, "Delivery and Language Use comprise 50% of the total grade and Topic Development constitutes the other half", or "Delivery, Language Use and Topic Development are equally important for determining the holistic grade, and the weight of each category is 1/3 of the final score". That being said, if a holistic score description is presented along with analytic descriptions, the holistic description must mention all the analytic criteria in order not to give raters an impression that some analytic criteria have either more or less importance in determining the holistic grade. Overall, the role of each analytic criterion should be explicitly specified to ensure the absence of variations in raters' interpretations.

The results of the study illustrated that raters could have their own personal opinions about their average severity and rubric criteria importance. The findings of the study can be used to create a questionnaire to survey raters employed at testing companies or language programs about their perceived severity and rubric criteria importance coupled with reasons behind those. Based on the information collected by such a questionnaire, employers can estimate if their raters' needs for more personalized training. Furthermore, the results of this special survey can be distributed among the employees to increase raters' awareness of their personal rating patterns and learn more about other raters' opinions. Additionally, it is advisable to hold in person or online rater professional development sessions to discuss the results of such surveys, which can be helpful to raise raters' awareness and understanding of what could be affecting their rating behavior on the subconscious level.

Finally, the implications of the study can be extended to teaching test preparation courses. Based on the fining that the raters in the present study paid more attention to Delivery and Topic Development, more emphasis should be paid to developing and improving students' skills in those areas such as articulation, intonation, idea development, and coherence. This suggestion does not imply that teachers can completely disregard teaching grammar and vocabulary but rather that the goal for students should be the ability to delivery clear, well-developed ideas and not to reach absolute grammatical accuracy.

Limitations and Directions for Future Research

This section presents some limitations of the study and makes suggestions for future research. Each limitation is followed by a suggestion on how future research can address this limitation.

In the current study, there was only one language group of raters (Russian) who were NNS raters and who shared their L1 with the test-takers. Thus, no generalizations can be made that other NNS raters with different L1s will have the same direction of severity/leniency towards a shared accent. Future research should focus on exploring L1 match effects for raters with other L1 backgrounds to understand if similar L1 match benefits apply.

In addition, the study did not attempt to answer the question regarding whether some accents might be phonologically closer to raters' L1 and thus easier to understand. For example, the pronunciation of sounds in languages such as Ukrainian, Serbian, or Belorussian are close to the pronunciation of sounds in the Russian language; therefore, raters who speak Russian might have a better understanding of speakers coming from those language backgrounds and also exhibit L1 match effects. It is suggested to explore if L1 match effects can be expanded to other languages that are related.

Another limitation is that the study did not encompass the differences in raters' personal attitudes towards any interlanguage phonology; therefore, it did not reveal any insights into how certain interlanguage phonology could possibly be perceived as more attractive or better sounding. As mentioned before, such a direction can also be dependent on linguistic stereotypes and on how phonologically close rater's L1 and examinees' L1 are. The raters in this study reported that they might have been subconsciously affected by this feature that is not relevant for determining test-takers' L2 speaking proficiency. This interesting finding should be further researched in an attempt to see if raters' perceived amiability of examinees' voice can affect raters' grades.

One of the variables that was not controlled in the study is variance within the same examinee accent (Arabic and Chinese). It is possible that the test-takers whose recordings were utilized in the study may have come from different regions of the same country; therefore, they could have different accents. This variable was not controlled in the study because it is not controlled in the target domain (real testing situations). In addition, there was no available information that could help accurately describe the specific regional accents of the participants. In order to provide the readers with the information about regional accent variability, an attempt was made to describe the sample using impressions of native speakers of these languages. Future research can better control such variation within the same examinee accent.

The speaking construct in this study was represented from the perspective of monologic independent speaking tasks. It should be acknowledged that the findings and implications of this dissertation are limited to semi-direct tests and differences may occur in face-to-face speaking tests. Future research should consider investigating decision-making processes of raters of face-to-face tests such as OPIs.

In this study, accent familiarity was defined based on raters' self-reports and raters' familiarity scores for examinee recordings. It is suggested that future studies revise the operational definition for accent familiarity to further explore this multi-faceted phenomenon. In addition, it is suggested to explore how much accent familiarity is accrued by officially employed TOEFL or IELTS raters with substantial rating experience. This information will help L2 speaking assessment researchers to see if developed accent familiarity with most common test-taker L1 can play a role.

Another limitation of the study is the way raters' comments were collected. The comment counts for each rater should be interpreted with caution as the directionality of rater comments was not normalized by examinee proficiency level and raters' negative/positive directionality was not considered when the raters' comments were counted. It is possible that the raters could have left more negative comments regarding Delivery or Topic Development for lower-level examinees and fewer positive comments about Delivery or Topic Development of higher-ability examinees. In addition, if some raters within each rater group left more negative comments for lower-level students and fewer positive comments for students whose ability was higher or vice versa, the numbers could have been skewed. These variables could have affected the dispersion of raters' comments across rubric criteria. Future research studies collecting comments from raters can investigate how raters' attention to rubric categories and the number of negative/positive comments can vary based on examinee proficiency levels.

Finally, the investigation of raters' cognitive processes while rating had an exploratory nature. The present study outlined the possible reasons for rater variation but did not provide a clear framework or confirm how much influence the differences in raters' decision-making processes can have on raters' scores and scoring patterns. Based on the findings in the study, future research can formulate hypotheses to conduct future research to investigate rater variation in the field of L2 speaking assessment.

Concluding Remarks

This dissertation study has shed light on additional potential factors that can cause rater variability during score decision-making process and affect L2 speaking test score uses and interpretations. This study offers an example of how the use of the argument-based validity framework (Chapelle et al., 2008; Kane, 2004; 2006; 2013) in L2 assessment encourages validity research the results of which will lead to improvements in testing procedures. The argument-based validity framework helped to situate the study theoretically and see how the aimed research questions challenge the assumption of the interpretive argument that raters utilize the rating scale appropriately. The results of the study outline potential sources of rater variation that can further be explored to strengthen the validity argument of L2 speaking performance assessment.

The investigation of differences between native and non-native raters, their cognitive processes while rating, and the effects of rater accent familiarity on scores contribute to L2 speaking assessment literature. The arguments presented in this study suggest the inclusion of more non-native speakers as L2 speaking raters and prompt further research on rater variability regarding raters' approaches to scoring, rubric interpretation, and utilization of non-rubric criteria. The work accomplished in this study provides a foundation for future research that can adopt and expand on the ideas from the present study furthering understanding of the causes of variability in L2 speaking raters. It is hoped that the evidence from the present dissertation and these future studies will lead towards establishing a framework of factors that contribute to rater variability in L2 speaking assessment.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Organization.

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgment on national-level oral examination tasks. *RELC Journal*, *42*(1), 31-51.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257.

Barkaoui K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27(4), 515–535.

Barkaoui, K. (2007a). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review*, *64*(1), 99–134.

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*, 54–74.

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment*. John Wiley & Sons, Inc.

Barnwell, D. (1989). 'Naive' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*(2), 152-163.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Bonk, W. J., & Ockey, G. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106,* 707–729. http://dx.doi.org/10.1016/j.cognition.2007.04.005

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1–15.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports*, 3, 49-84.

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *Studies in language testing 19: IELTS collected papers: Research in speaking and writing assessment* (pp. 98–139). Cambridge: Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1), i-157. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-05-05.pdf

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201–219. doi:10.1177/0265532210393704

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (2nd ed.) (pp. 313-321). New York: McGraw-Hill.

Chalhoub-Deville, M. & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*, 383–391.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*, 16–33.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York & London: Routledge.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…, *Language Testing*, *29*, 19–27.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*, 3–13.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Clark, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Clark, J. L. D. (1972). *Foreign language testing: Theory and practice*. Philadelphia: Center for Curriculum Development.

Connor-Linton, J. (1995). Looking behind the curtain: Why do L2 composition ratings really Mean. *TESOL Quarterly*, *29*(4), 762-765.

Constable, E., & Andrich, D. (1984). Inter-Judge Reliability: Is Complete Agreement among Judges the Ideal?. *ERIC*

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W., & Piano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice,31*, 10–20.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessment. *Assessment in Education*, *3*(3), 265-285.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67-96.

Davies, A. (1989). Is International English an interlanguage? *TESOL Quarterly*, *23*(3), 447-467.

Davies, A. (2011). Does language testing need the native speaker? Lang*uage Assessment Quarterly*, *8*, 291–308.

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003) Whose norms? International proficiency tests in English. *World Englishes*, *22*(4), 571-584.

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1) 117 –135. http://dx.doi.org/ 10.1177/0265532215582282

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition*, *19*(01), 1-16.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289–303.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155-185.

Eckes, T. (2009a). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2009b). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang.

Eckes, T. (2010). The TestDaF implementation of the SOPI: Design, analysis, and evaluation of a semi-direct speaking test. In L. Araújo (Ed.), *Computer-based assessment (CBA) of foreign language speaking skills* (pp. 63–83). Luxembourg: Publications Office of the European Union.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. doi:10.1080/15434303.2011.649381

Engelhard, G. J., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model (College Board Research Report No. 2003-1). Princeton, NJ: *Educational Testing Service*. Retrieved from http://www.ets.org/Media/Research/pdf/RR-03-01-Engelhard.pdf

Fayer, J. M. & Krasinski, E. (1987). Native and non-native judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326.

Fulcher, G. (2003). *Testing second language speaking*. London, England: Pearson Longman.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of non-native speech. *Language Learning*, *34*(1), 65-87.

Greene, J. C, Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*, 186–203.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second language communication. *Language Learning*, *41*(1), 1-24.

Haladyna, T. M. & Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hamo-Lyons, L. I. Z., & Davies, A. (2008). The Englishes of English tests: bias revisited. *World Englishes*, *27*(1), 26-39.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*(2), 163-180, doi:10.1177/0265532211421161

Hesse-Biber, S. N., & Leavy, P. (2006). *The practice of qualitative research*. Thousand Oaks, CA: Sage Publications.

Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing*, 5, 29–50.

Hill, K. (1997). Who should be the judge?: The use of non-native speakers as raters on a test of English as an international language. In Huhta, A., Kohonen, V., Kurki-Suonio, L., & Luoma, S., editors, Current developments and alternatives in language assessment: Proceedings of LTRC 96 (pp. 275–290). Jyväskylä: University of Jyväskylä and University of Tampere.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403–424.

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41*(3), 770-785.

Joe, J. N. (2008). Using verbal reports to explore rater perceptual processes in scoring: An application to oral communication assessment. ProQuest.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 239-258.

John W. Creswell, J. W. & Zhou, Y. (2016). What is mixed methods research? In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 35-50). Cambridge: Cambridge University Press.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In Quirk, R. & Widdowson, H., editors, *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge: Cambridge University Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement*, *38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: Greenwood.

Kane, M. (2010). Validity and fairness. *Language testing*, *27*(2), 177-182.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3-17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198-211.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice*, *18*(2), 5-17.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, *9*(3), 249-269.

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating US undergraduates’ attitudes toward international teaching assistants. *TESOL Quarterly*, *49*(4), 681-706.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review, 64*(3), 459-489.

Kim, Y. (2009). An investigation into native and non-native teachers’ judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187–217. doi:10.1177/0265532208101010

Kim, Y. (2009). Exploring rater and task variability in second language oral performance assessment. In Hill, K., & Brown, A. (Eds.). *Tasks and Criteria in Performance Assessment Proceedings of the 28th Language Testing Research Colloquium*. Frankfurt: Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, *28*(2) 179–200.

Kobayashi, H. & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning*, *46*, 397–437.

Kobayashi, T. (1992). Native and non-native reactions to ESL compositions. *TESOL Quarterly*, *26*(1), 81–112.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–60.

Linacre, J. M. & Williams, J. (1998). How much is enough? *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 12, 653.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2000). Comparing "partial credit" and "rating scale" models. *Rasch Measurement Transactions*, *14*(3), 768.

Linacre, J. M. (2017). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com

Lowenberg, P. H. (2000). Assessing English proficiency in the global context: The significance of non-native norms. In Kam, H. W., editor, *Language in the global context: Implications for the language classroom* (pp. 207–228). Singapore: SEAMEO Regional Language Center.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters. *Language Testing*, *19*, 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54-71.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*, 425–44.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127–145. doi:10.1080/15434303.2011.565845

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. F. (1997). “Interaction” in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–65.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: American Council on Education / Oryx Press.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching*. Pearson Higher Ed.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(01), 111-131.

Myford, C. M. & Wolfe, E. W. (2000). Monitoring sources of variability within the test of spoken English Assessment System (*TOEFL Research Report No. RR-65*). Princeton, NJ:

Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-faceted Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-faceted Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189-227.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217-237.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

Plano Clark, V. L., & Badiee, M. (2010). Research questions in mixed methods research. *Mixed Methods in Social and Behavioral Research*, 275-304.

Purpura, J. E. (2014). Cognition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-25). John Wiley & Sons, Inc.

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*, 113-125.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. Experimenting with uncertainty: Essays in honour of Alan Davies, 11, 82-96.

Saito, K., & Shintani, N. (2016b). Do native speakers of North American and Singapore English differentially perceive second language comprehensibility? *TESOL Quarterly*, *50*, 421-446.

Savignon, S. J. (1972). *Communicative competence: An experiment in foreign language teaching*. Philadelphia: The Center for Curriculum Development.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*(3), 355–91.

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly*, *40*(4), 715-738.

Selinker, L. (1972). Interlanguage. IRAL-International Review of Applied Linguistics in *Language Teaching*, *10*(1-4), 209-232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shi, L. (2001). Native- and non-native-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303–325.

Sidaras, S. K., Alexander, J. E. D., & Nygaard, L. C. (2009). Perceptual learning of systemic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306–3316.

Spolsky, B. (1973). What does it mean to know a language; or how do you get someone to perform his competence? In J. W. Oiler & J. C. Richards (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher* (pp. 164-176). Rowley, MA: Newbury House Publishers.

Stansfield, C. W. & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, *72*, 129–141.

Stansfield, C. W. & Kenyon, D. M. (1992b). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*, 347–364.

Strauss, A., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. London, UK: SAGE publications.

Suto, I. (2012). A critical review of research methods used to explore rater cognition. *Educational Measurement: Issues and Practice*, *31*, 21–30.

Teddlie, C, & Tashakkori, A. (2006). A general typology of research design featuring mixed methods. *Research in the Schools*, *13*(1), 12-18.

Timothy C. Guetterman, T.C. & Salamoura, A. (2016). Enhancing test validation through rigorous mixed methods components. In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 153-176). Cambridge: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge, UK: Cambridge University Press.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. HampLyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–25). Norwood, NJ: Ablex.

Wei, J., & Llosa, L. (2015). Investigating Differences Between American and Indian Raters in Assessing TOEFL iBT Speaking Tasks. *Language Assessment Quarterly, 12*(3), 283-304.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*, 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263-287.

Weigle, S. C. (2002). Assessing writing. Cambridge, UK: Cambridge University Press.

Weil, S. A. (2001). Foreign-accented speech: Encoding and generalization. *Journal of the Acoustical Society of America*, *109*, 2473 (A).

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave MacMillan.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305-335.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*. doi:10.1002/tesq.73.

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *ETS Research Report Series*,(2), i-67.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231–252. doi:10.1177/0265532212456968

Wolfe, E. W. (1995). A study of expertise in essay scoring. Unpublished doctoral dissertation, University of California, Berkeley.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*, 83-106.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. doi:10.1111/emip.2012.31.issue-3

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*, 465-492.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 8*, 370.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.

Xi, X., & Mollaun, P. (2009). How Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?. *ETS Research Report Series*, 2009(2), i-37.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222–1255. doi:10.1111/j.1467-9922.2011.00667.x

Yang, R. (2010). A many-facet Rasch analysis of rater effects on an Oral English Proficiency Test. Retrieved from ProQuest Dissertations Publishing (3570410)

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31–50. doi:10.1177/0265532209360671

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgments of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325.

Ziegler, N. & Kang, L. (2016). Mixed methods designs. In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 51-82). Cambridge: Cambridge University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale: Lawrence Erlbaum Associates.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4 (2), 223-233.

Appendix A

Student Background Survey

My gender is ______

My first language is ______

Translation

Информация об участнике

Мой пол ____

Мой родной язык ____

Appendix B

Rater Background Questionnaire

Adapted from Language Experience Questionnaire (Harding, 2012) and Rater Language Background Questionnaire (Wei & Llosa, 2015)

Part 1 (Recruitment Stage)

Directions: Fill out the questionnaire to the best of your knowledge.

General.

1. Age: _______

2. Gender: Male/Female

3. In which country were you born? ___________________________

4. Have you ever lived in another country for more than 3 months? __Yes __No. If no, go to #8

5. Where?____

6. For how long?____

7. For what purpose?____

8. Educational background (fill out those that apply):

Undergraduate degree in ____

Certificate in ____

Master's degree in ____

Doctoral degree in ____

Other ____

9. Have you taken any courses on assessment of speaking and writing? Yes/No If no, skip #10

10. How many assessment courses have you taken? _____

11. What kind of courses were they? ____

____

Languages.

1. What is your native language/mother tongue? _______________________

2. Other languages spoken:

Additional language 1 _____________________

Additional language 2 _____________________

Additional language 3 _____________________

Additional language 4 _____________________

3. Please rate your ability to use these languages (low/intermediate/advanced/almost native).

Additional language 1 _____________________

Additional language 2 _____________________

Additional language 3 _____________________

Additional language 4 _____________________

4. Is English your native language/mother tongue? __Yes __ No. If yes, go to the next section.

5. For how long have you studied English? ___________

6. Have you studied English abroad? Yes/No. If no, go to #8

7. Where? ____

8. Please rate your ability to use English in academic settings by checking the appropriate level in the table below.

	Low	Intermediate	Advanced	Almost Native
Listening				
Speaking				
Reading				
Writing				

Teaching experience.

1. For how many years in total have you taught English?_________
2. In what countries have you taught and for how long? ____________________________

____

3. Students from what countries have you had in your classroom?____

Rating experience.

1. Have you scored any standardized language tests before (e.g.,, IELTS, TOEFL)? Yes/No. If no, go to #4

2. If yes, what is/are the test(s) that you scored? _________________________________

____

3. What are the language skills that you scored? (Select all that apply):

__ Speaking __ Reading __ Listening __ Writing

4. Did you receive any formal training as a rater? Yes/No. If no, skip #5.

5. Briefly, describe the training that you received_____________________________________

____

Part 2 (After the Study)

Thank you for participating in my project! This is the last step!

Answering these questions, think about your familiarity BEFORE your participation in my research study.

This final survey has two parts:

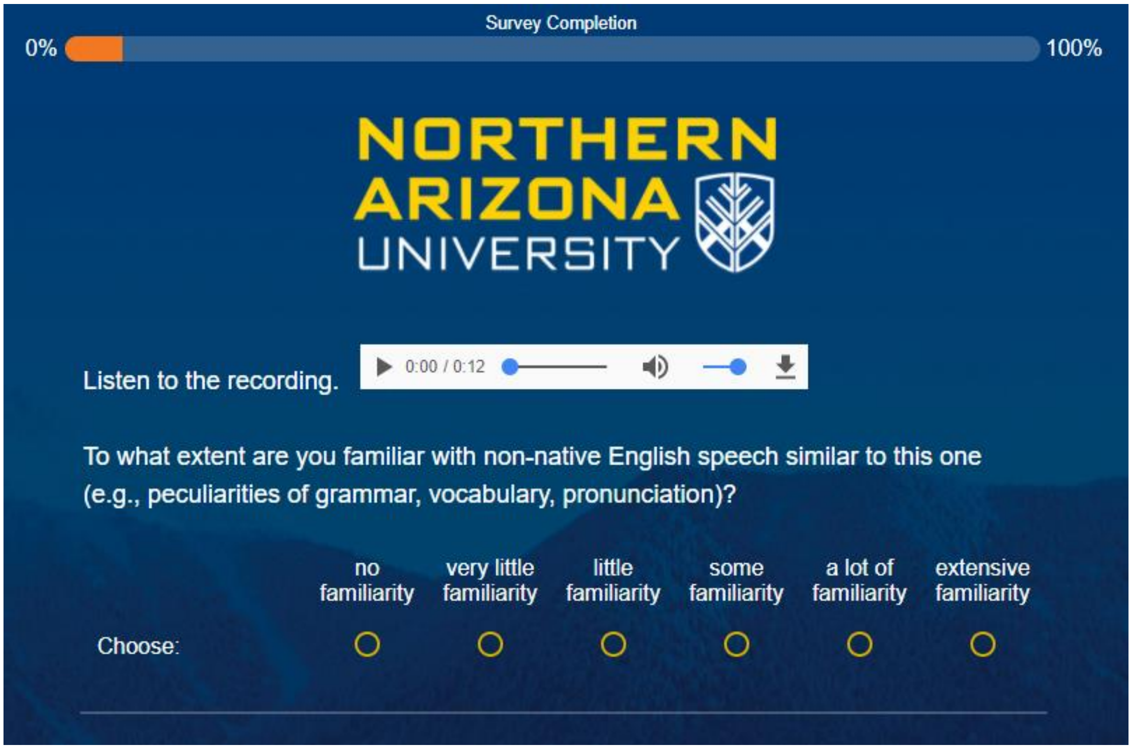
Part #1: Listen to recordings (12 seconds each) and estimate to what extent you are familiar with non-native English speech similar to the one heard (e.g., peculiarities of grammar, vocabulary, pronunciation).

Part #2: Answer 4 questions about your familiarity with non-native speech.

Familiarity Scale

Directions: Listen to the recording.

To what extent are you familiar with non-native English speech similar to this one (e.g., peculiarities of grammar, vocabulary, pronunciation)?



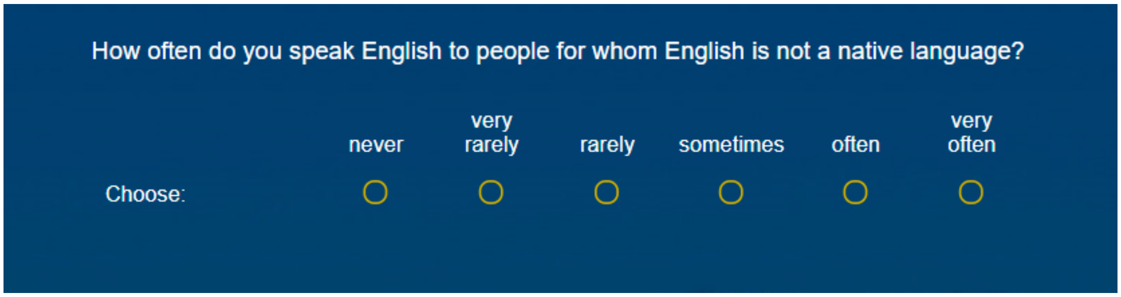
Listen to the recording.

To what extent are you familiar with non-native English speech similar to this one (e.g., peculiarities of grammar, vocabulary, pronunciation)?

	no familiarity	very little familiarity	little familiarity	some familiarity	a lot of familiarity	extensive familiarity
Choose:	○	○	○	○	○	○

Familiarity with accents spoken by nonnative English speakers.

1. How often do you speak English to people for whom English is not a native language?



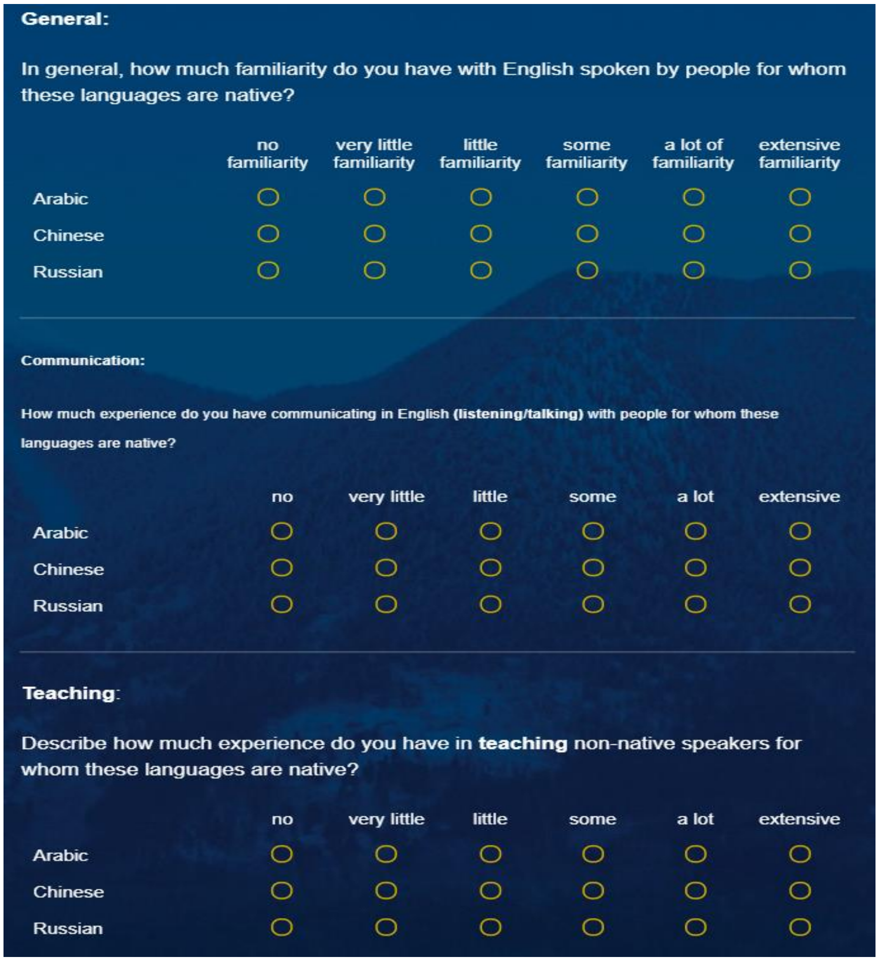
How often do you speak English to people for whom English is not a native language?

	never	very rarely	rarely	sometimes	often	very often
Choose:	○	○	○	○	○	○

2. In general, how much familiarity do you have with English spoken by people for whom these languages are native?

3. How much experience do you have communicating in English (listening/talking) with people for whom these languages are native?

4. Describe how much experience do you have in teaching non-native speakers people for whom these languages are native?



General:

In general, how much familiarity do you have with English spoken by people for whom these languages are native?

	no familiarity	very little familiarity	little familiarity	some familiarity	a lot of familiarity	extensive familiarity
Arabic	○	○	○	○	○	○
Chinese	○	○	○	○	○	○
Russian	○	○	○	○	○	○

Communication:

How much experience do you have communicating in English **(listening/talking)** with people for whom these languages are native?

	no	very little	little	some	a lot	extensive
Arabic	○	○	○	○	○	○
Chinese	○	○	○	○	○	○
Russian	○	○	○	○	○	○

Teaching:

Describe how much experience do you have in **teaching** non-native speakers for whom these languages are native?

	no	very little	little	some	a lot	extensive
Arabic	○	○	○	○	○	○
Chinese	○	○	○	○	○	○
Russian	○	○	○	○	○	○

Appendix C

Speaking Prompts

Task #1

Prepare: 1 minute; Speak: 1 minute

Preparing: Read the following question and then prepare your answer. You may take notes on this paper. Your response will be scored according to:

- Development of ideas
- Pronunciation
- Grammar and vocabulary

Question: You have an exam next week. Do you want to study alone or in a group? Include reasons and examples to support your answer.

Task #2

Prepare: 1 minute; Speak: 1 minute

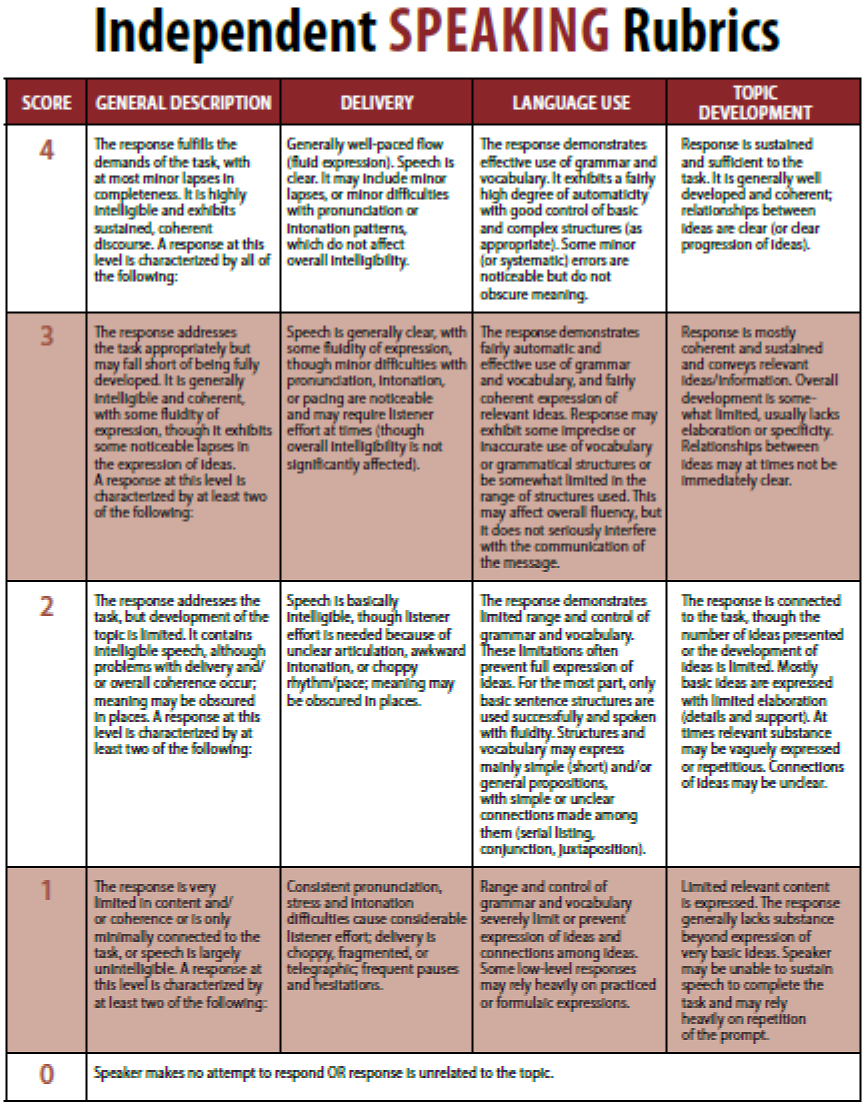
Preparing: Read the following question and then prepare your answer. You may take notes on this paper. Your response will be scored according to:

- Development of ideas
- Pronunciation
- Grammar and vocabulary

Question: There are different ways to teach students. Some universities have large classes with many students. Other universities have small classes. Which of these classrooms is better for learning? Use specific examples to support your answer.

Appendix D

TOEFL Independent Speaking Rubric



Independent SPEAKING Rubrics

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

Use this link to access a PDF version of the rubric
https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Appendix E

Benchmarking guidelines adapted from Weigle, S. (2002)

Goal: Find 2 language samples that typify each score band of a rubric for each L1.

Procedure

1. Read language production prompt.
2. Read rubric scales and descriptors.
3. Randomly select a small collection of language samples.
4. Rate the small collection of language samples. Try to identify features that exemplify strengths or deficiencies mentioned in the rubric descriptors.
5. Record score rationale for the selected band-typical samples.
6. Compare the ratings of the score band-typical samples with pre-scored ratings in order to determine whether the ratings agree.
8. If so, make notes of which samples typify each score band. Write notes to explain why each sample fits into the specific band. Make sure that the samples are absolutely representative of that band.
9. Repeat steps 3-8 until the goal is reached.

Appendix F

Rater Training Script

The researcher followed the script and verbally guided the raters through the Qualtrics online system (e.g., "Now click Next").

Hi! How are you doing today? Thank you for agreeing to participate in my research study. Today we will spend about 1 hour 30 minutes. Our session will have three parts:

1. Rubric and task familiarization
2. Training and practice
3. Rating

Let's begin with the first one!

Rubric and task familiarization

You will be scoring speech recordings from English learners in response to prompt #1 and prompt # 2. Students had 1 minute to prepare and 1 minute to speak. The questions were "…". Now you can take the time to read through the prompt. You can make comments and ask questions if you wish.

Here is the rubric, which will be used to score the recordings. It assesses delivery, topic development, and language use. It also has a general description of overall performance. The possible scores can vary from 0 to 4. As you can see, 0 means that there is no response or response is not on the topic, we will NOT have any 0 score recordings, all the recordings should be assigned a score from 1 to 4. Scores from 1 to 4 describe the quality of appropriate responses. Scores for each category might be the same or might fall into different bands.

Now you can take the time to read through the rubric paying attention to each criterion in each score band. You can make comments and ask questions (give the time needed, approximate length 5 min).

Ok, now let's review the rubric together. I'll explain the salient features for each category.

First, the salient features that distinguish 3 and 4 are that the score of 4 must have all three elements in its band, and three, however, should have two elements in its band and one in a band lower or higher.

And according to the rubric, a 4 is a fluent, clear, intelligible answer which might be a bit flawed. It is a well-developed answer, with clear and connected ideas. Grammar and vocabulary are good but might be a bit flawed which does not obscure the meaning.

A 3 is not as easy to understand as a 4 and it might require some listener effort. It's grammar, vocabulary and topic development are good but a bit limited.

2 is more difficult to understand and it requires listener effort. Grammar and vocabulary affect the expression of ideas in a negative way. Topic development is basic, not elaborated, vague, repetitive with unclear or not connected ideas.

A 1 can have a lot of pauses, hesitations and pronunciation mistakes and it needs a lot of listener effort. Its grammar is severely limited. The ideas are very basic and maybe repeating the prompt, using memorized expressions and be repetitive.

And as I already mentioned, we will not have any recordings with the score of 0 because zero means no answer.

So, 4 is the best, 1 is the worst and 3 and 2 are in the middle. You can look at the descriptions of 2 and 3 in order to find how you would differentiate them. (Give time, elicit an answer). Discuss each criteria (delivery, language use, topic development) in more details.

Now we are finished. Let’s move on to training.

Training

Now we will have the training. I’ll play 6 one-minute recordings overall. Six of the recordings will be in response to task 1 and six in response to task 2.

These recordings are from a proficiency test, so students did not study this topic in a classroom and their ideas are on the spot ideas. Also, students who took this test come from various language backgrounds, so you can expect hear students from different L1 backgrounds.

You will hear the recording once from the beginning to the end. Then, you can listen to the recording again for as many times as you want and pause it if needed.

After you are comfortable with the recording, I will tell you what score it was assigned.

Then, using the rubric, you will express your opinion why this recording was given this particular score based on the sub-score ratings.

Do you have any questions?

We will be following this outline to help us structure the procedure (raters have it on the screen):

1. Imagine you are a high-stakes TOEFL rater.
2. Play the recording as many times as you need until you are comfortable with the recording.
3. I give you the holistic score it was assigned.
4. Using the rubric, express your opinion why this recording was given this score in terms of sub-scores.

Let’s start.

Recording #1. Let’s listen. You may take notes if you wish.

Now you can listen again and pause if needed.

This recording was given a score of 2. Why do you think it was given this score?

Let’s move on to recording #2.

The same pattern with the rest of the recordings.

Learning to use the rubric

Now that we have had training, we will have some practice. I’ll play 8 more recordings. 4 of the recordings will be in response to task 1 and 4 in response to task 2.

You will hear the recording once from the beginning to the end. Then, you can listen to the recording again for as many times as you want and pause it if needed.

After you are comfortable with the recording, you will give your score for each sub-category based on the rubric.

Then, I will give you the score it was assigned and if your initial score was different, you will try to adjust your grading, in needed

Do you have any questions? Let’s start.

We will be following this outline to help us structure the procedure (raters see it on the screen):

1. Imagine you are a high-stakes TOEFL rater.
2. Play the recording as many times as you need until you are comfortable with the recording.
3. Based on the rubric, give your score for each sub-category.
4. I give you the score it was assigned and if your initial score was different, you will try to adjust your grading.

Calibration Practice

Now that we have had training, we will have some practice. I’ll play 10 more recordings. 5 of the recordings will be in response to task 1 and 5 in response to task 2.

You will hear the recording once from the beginning to the end. Then, you can listen to the recording again for as many times as you want and pause it if needed.

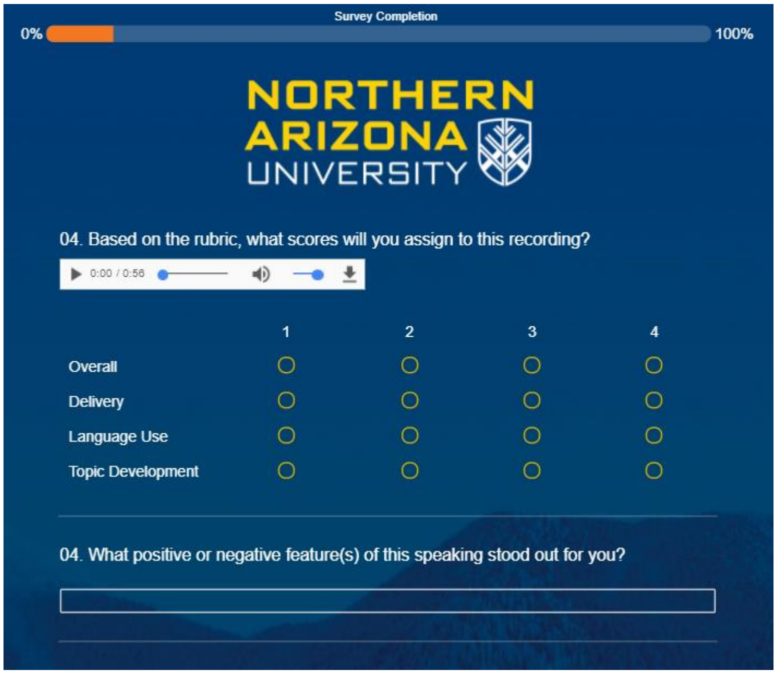
After you are comfortable with the recording, you will give your score for each sub-category based on the rubric and we will move on to the next recording.

Do you have any questions? Let’s start.

We will be following this outline to help us structure the procedure (raters see it on the screen):

5. Imagine you are a high-stakes TOEFL rater.
6. Play the recording as many times as you need until you are comfortable with the recording.
7. Based on the rubric, give your score for each sub-category.
8. Move on the next recording

Now we are about to grade two last recordings, here I would also like you to practice giving comments for each recording. After you are done grading, type your comment or comments in the box and then read them to me. The comments can be about anything that stood out for you in the students’ answer such as grammar, vocabulary, pronunciation, ideas, or something else.



04. Based on the rubric, what scores will you assign to this recording?

	1	2	3	4
Overall	○	○	○	○
Delivery	○	○	○	○
Language Use	○	○	○	○
Topic Development	○	○	○	○

04. What positive or negative feature(s) of this speaking stood out for you?

Rating

Now that we have had training and practice, you will do the rating. Your rating will include X recordings. You should grade the recordings in the order that they are presented to you and mark your scores. You cannot go back and change your scores. You can take a break if needed between grading Task 1 and Task 2.

Appendix G

Rating Plan

Examinees	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1A*	*	*	*	*	*	*	*	*	*	*
2C	*	*	*	*	*	*	*	*	*	*
3 R	*	*	*	*	*	*	*	*	*	*
4 A	*	*	*	*	*	*	*	*	*	*
5 C	*	*	*	*	*	*	*	*	*	*
6 R	*	*	*	*	*	*	*	*	*	*
7 A	*	*	*	*	*	*	*	*	*	*
8 C	*	*	*	*	*	*	*	*	*	*
9 R	*	*	*	*	*	*	*	*	*	*
10 A	*	*	*	*	*	*	*	*	*	*
11 C	*	*	*	*	*	*	*	*	*	*
12 R	*	*	*	*	*	*	*	*	*	*
13 A	*	*	*	*	*	*	*	*	*	*
14 C	*	*	*	*	*	*	*	*	*	*
15 R	*	*	*	*	*	*	*	*	*	*
16 A	*	*	*	*	*	*	*	*	*	*
17 C	*	*	*	*	*	*	*	*	*	*
18 R	*	*	*	*	*	*	*	*	*	*
19 A	*	*	*	*	*	*	*	*	*	*
20 C	*	*	*	*	*	*	*	*	*	*
21 R	*	*	*	*	*	*	*	*	*	*
22 A	*	*	*	*	*	*	*	*	*	*
23 C	*	*	*	*	*	*	*	*	*	*
24 R	*	*	*	*	*	*	*	*	*	*
25 A	*	*	*	*	*	*	*	*	*	*
26 C	*	*	*	*	*	*	*	*	*	*
27 R	*	*	*	*	*	*	*	*	*	*
28 A	*	*	*	*	*	*	*	*	*	*
29 C	*	*	*	*	*	*	*	*	*	*
30 R	*	*	*	*	*	*	*	*	*	*
31 A	*	*								
32 C	*		*							
33R	*			*						
34 A	*				*					
35 C	*					*				
36 R	*						*			
37 A	*							*		
38 C	*								*	

39 R	*									*
40 A		*	*							
41 C		*		*						
42 R		*			*					
43 A		*				*				
44 C		*					*			
45 R		*						*		
46 A		*							*	
47 C		*								*
48 R			*	*						
49 A			*		*					
50 C			*			*				
51 R			*				*			
52 A			*					*		
53 C			*						*	
54 R			*							*
55 A				*	*					
56 C				*		*				
57 R				*			*			
58 A				*				*		
59 C				*					*	
60 R				*						*
61 A					*	*				
62 C					*		*			
63 R					*			*		
64 A					*				*	
65 C					*					*
66 R						*	*			
67 A						*		*		
68 C						*			*	
69 R						*				*
70 A							*	*		
71 C							*		*	
72 R							*			*
73 A								*	*	
74 C								*		*
75 R									*	*

Appendix H

Think-Aloud Protocol Script

The researcher followed the script and verbally guided the raters through the Qualtrics online system (e.g., “Now click Next”).

Hi! How are you doing today? Thank you for agreeing to participate in my research study. Today we will spend about 2 hours. Our session will have three parts:

1. Tasks, rubric, and benchmark review with think-aloud practice
2. Rating
3. Interview

Let’s begin with the first one!

Benchmark review and think-aloud practice

Let’s review the tasks and the rubric. Now we will review the benchmarks that you have already listened to in the first part of my study, but we will change one thing, and I will explain it right now, but before I do it, I would like to remind you that these recordings are from a proficiency test, so students did not study this topic in a classroom and their ideas are on the spot ideas. Also, students who took this test come from various language backgrounds, so you can expect hear students from different L1 backgrounds.

I’ll play 8 one-minute recordings overall. 4 of them are in response to task 1, which was about exam preparation preference, and 4 of them are in response to task 2, which was about university classes.

You will hear each recording once from the beginning to the end, and I will tell you the overall score and then, here is what gonna be different.

Then, using the rubric, you will express your opinion why this recording was given this particular score. The score that I provide to you is the overall score for the recording, but I would ask you to try to predict what scores it was possibly given for each sub-category: delivery language use and topic development.

When you are giving scores, you will be thinking aloud. What I mean by “think aloud” is that I want you to say out loud everything that you would say to yourself silently while you think. Just act as if you were alone in the room speaking to yourself. Please provide as thorough a justification as possible.

Do you have any questions?

We will be following this outline to help us structure the procedure (raters have it on the screen):

1. Imagine you are a high-stakes TOEFL rater
2. Play the recording once (you may listen again later)
3. I give you the holistic score it was assigned
4. Verbalize thoughts that you might have about this recording, give your own sub-category scores to fit the holistic grade, provide brief justifications based on the rubric

Let’s start.

Recording #1. Let’s listen. You may take notes if you wish.

Now you can listen again and pause if needed.

This recording was given a score of 2. Why do you think it was given this score, what would your scores for each sub-category be based on the rubric?

Probes:

Ok. Do you have anything else to add?

Let’s move on to recording #2.

The same pattern with the rest of the recordings.

Rating

Now that we have reviewed the benchmarks and practiced verbalizing your thoughts, we will have the rating. I’ll play 12 more recordings.

You will listen to each recording once from the beginning to the end and explain what you were doing while listening and what thoughts you had.

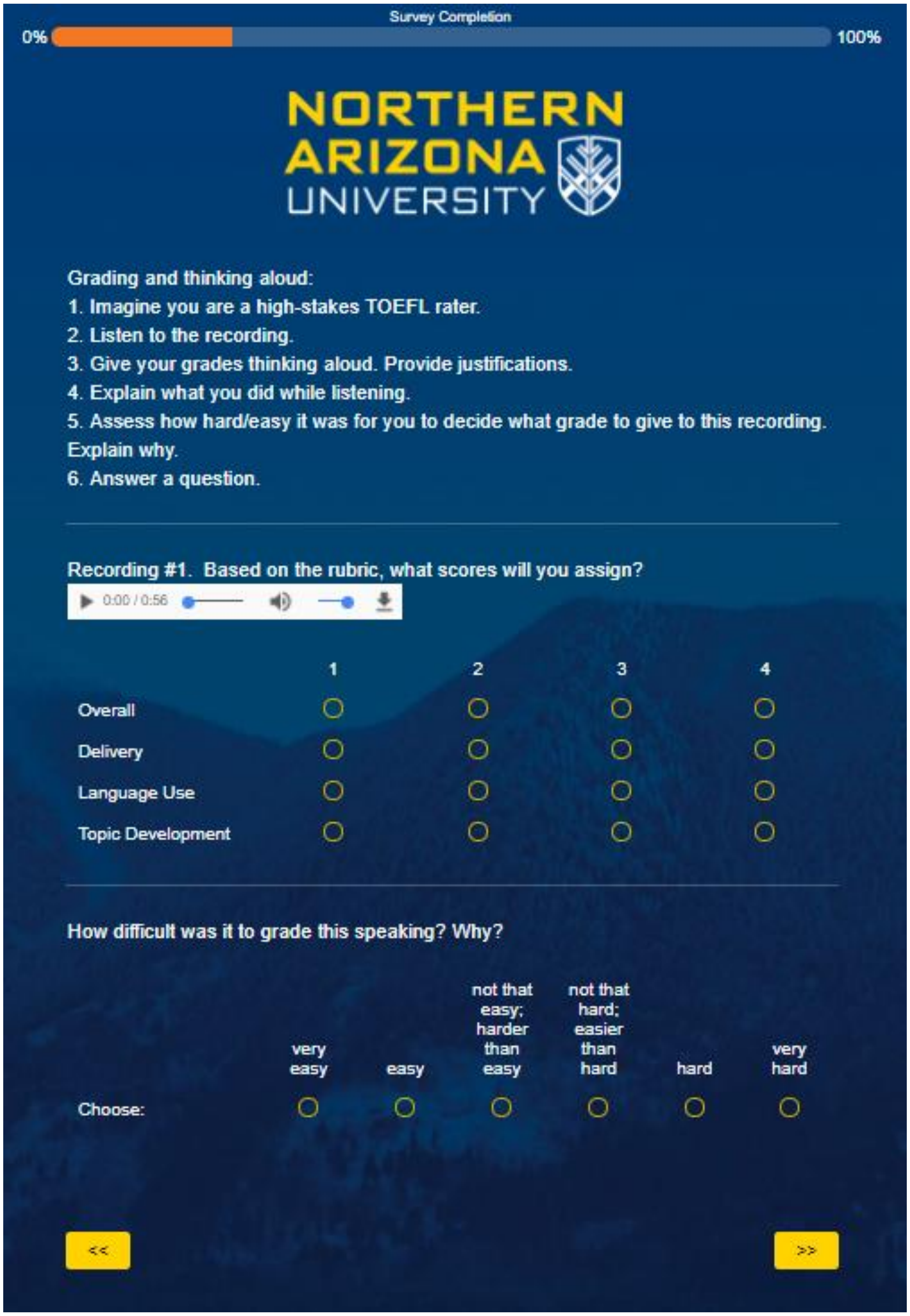
Then, you can listen to the recording again as many times as you want and pause it if needed. You will continue verbalizing your thoughts out loud. Explain your thought processes (e.g., hesitations, change of mind, etc.) and actions (e.g., looking at delivery band 4, trying to decide between 3 and 4, etc.) while deciding to assign sub-category scores and justify you score decisions.

Then, you will tell me if it was easy or hard to grade this recording and support it with a very brief reason why.

After that, I will give you the assigned score of some other rater who also scored this recording but gave a different score than you did. And then we will discuss what allowed that rater to go with a higher or a lower score than you. I am not saying that your scored are incorrect, I am just trying to get some help from you to try to understand why those raters had a different opinion.

Do you have any questions?

We will be following the same outline to help us structure the procedure (raters have it on the screen):



Grading and thinking aloud:

1. Imagine you are a high-stakes TOEFL rater.
2. Listen to the recording.
3. Give your grades thinking aloud. Provide justifications.
4. Explain what you did while listening.
5. Assess how hard/easy it was for you to decide what grade to give to this recording. Explain why.
6. Answer a question.

Recording #1. Based on the rubric, what scores will you assign?

	1	2	3	4
Overall	○	○	○	○
Delivery	○	○	○	○
Language Use	○	○	○	○
Topic Development	○	○	○	○

How difficult was it to grade this speaking? Why?

	very easy	easy	not that easy; harder than easy	not that hard; easier than hard	hard	very hard
Choose:	○	○	○	○	○	○

Let’s start.

Recording #1. Let’s listen. You may take notes if you wish.

Probes:

Continue verbalizing your thoughts.

Now you can listen again and pause if needed.

And what would you give for Delivery?

Can you give some justification for your Topic Development score?

Can you explain your thought processes? What was your strategy?

Could you give some reasons for that?

This recording was given a score of 4. What do you think allowed you to go a band higher/lower? Why do you think it is different?

Was it easy or hard for you to grade this recording? Why?

What are your thoughts about the recording based on the rubric?

What grades would you give for each sub-category?

What were you doing while listening?

What thoughts did you have during listening?

You said "…" what do you mean by that?

You said "…" what exactly do you mean?

Can you give an example?

Why do you think so?

Ok, can you tell me more?

And?

So?

So, you are saying that …. is ….?

So, you want to say that …. is …?

So, you mean that …. is …?

So, what you are saying is ….?

Ok. Do you have anything else to add?

Let's move on to recording #2.

Thank you for your input! I really appreciate it! Let's move on to the short interview.

Appendix I

Interview Questions and Script

1. How was your rating experience today?

2. Did you have any specific test-takers which were hard or easy to rate?

3. In general, do you consider yourself a severe or a lenient rater?

4. You as a rater, do think you have any specific strategy or pattern of rating?

5. There were some recordings when you were hesitant what score to give. What do you think was causing your hesitations?

6. During our session, today, do you think you might have been harsher or more lenient on some test-takers?

7. Looking at the rubric, do you think that each sub-category is equally important?

8. Do you think you might tend to be harsher or more lenient on some sub-categories?

9. Can you reflect on your scoring of delivery?

10. Can you reflect on your scoring of language use?

11. Can you reflect on your scoring of topic development?

12. Do you think that any factors that are NOT on the rubric might have affected your score decisions?

13. There were some Arabic, Chinese, and Russian speakers among the recordings that you graded, could you distinguish them?

14. Do you think your familiarity (or lack of it) with the way Arabic, Chinese, and Russians speak (e.g., peculiarities of grammar, vocabulary, pronunciation) affected your scores?

15. Do you have any other insights, thoughts, or suggestions about improving rubrics, rater training, or exam fairness?

Thank you for participating in the think-aloud session, now we are going to have the interview. I'll ask you some questions about your today's rating experience and some thoughts you have about it and about you as a rater.

1. How was your rating experience today? Like overall impression, what you think was easy or hard or maybe some insights that you got today, or maybe like some other thoughts?

Other probes:

Why are you saying it was hard/easy?

You said "…" why do you think it was hard?

Ok, can you tell me more?

Can you give an example?

Why do you think so?

2. Hypothetically, did you have any specific test-takers which were hard or easy to rate? Why so? Just generally, who was harder to rate and who was easier to score?

Other probes:

Can you give an example?

You said "….", what did you mean?

Can you expand on that?

What makes you say so?

Any other reasons for a recording being hard to grade?

So, your opinion is …? (pause to elicit continuation)

Why was that?

3. There were some recordings when you were hesitant what score to give, like when you were not sure what to give, what do you think was causing your hesitations? Can you recall?

Could you tell me more?

What exactly do you mean by…?

You said "…", can you explain that?

What makes you think so?

4. In general, do you consider yourself a severe, like a harsh or a lenient, like liberal rater? And what makes you think so?

Other probes:

You said that you "…", could you explain?

Are you always a severe/lenient rater?

You said "…" can you give an example?

What do you mean by "…"?

Can you give more details?

So, you mean …? (pause to elicit continuation)

What makes you think so?

Why neither lenient nor harsh?

5. During our session, today, do you think you might have been, like hypothetically, harsher or more lenient on some test-takers? Why?

Other probes:

When you say "…" you mean….? (pause to elicit continuation)

Why do you think so?

Can you tell me more?

Why do you think it was like that?

So, you are saying …? (pause to elicit continuation)

6. Looking at the rubric, for you personally, do you think that each sub-category is equally important or not? Why? Like do you think that one is more important than the other one or they all have the same level of importance?

Other probes:

How would you order TD, LU and D according to their importance?

Do you think that other raters share the same opinion as you or they might have other thoughts?

So, you want to say …? (pause to elicit continuation)

Why do think so?

Any examples?

Any reasons?

7. Another one about sub-categories. Do you think you might be harsher or more lenient on some sub-categories like topic development, delivery or language use? And why or why not?

Other probes:

So, you are saying …? (pause to elicit continuation)

So, why do you think that "…" is the most important?

So, why do you believe that "…" is #1 for you?

Any reasons for that?

Because …? (pause to elicit continuation)

8. You as a rater, do think you have any specific strategy or pattern of rating, like what do you do when you are listening and what you attend to, and how you arrive at that or this grade? Like do you have a specific strategy or system, like what do you do first, second, etc.? What you pay attention to first and then or like if you take notes?

Other probes:

Why do you listen for … first?

What makes you go to … after that?

Any reasons for that?

Do you take notes?

9. Can you reflect on your scoring of delivery?
Other probes:
Why?
Can you tell me more?
What exactly do you mean?

10. Can you reflect on your scoring of language use?
    Other probes:
    Why?
    Can you tell me more?
    What exactly do you mean?

11. Can you reflect on your scoring of topic development?
    Other probes:
    Why?
    Can you tell me more?
    What exactly do you mean?
    What do you do when it is hard to understand a student's idea?

12. Do you think that any factors that are not on the rubric might have affected your score decisions?
    Other probes:
    Why?
    Can you tell me more?
    What exactly do you mean?

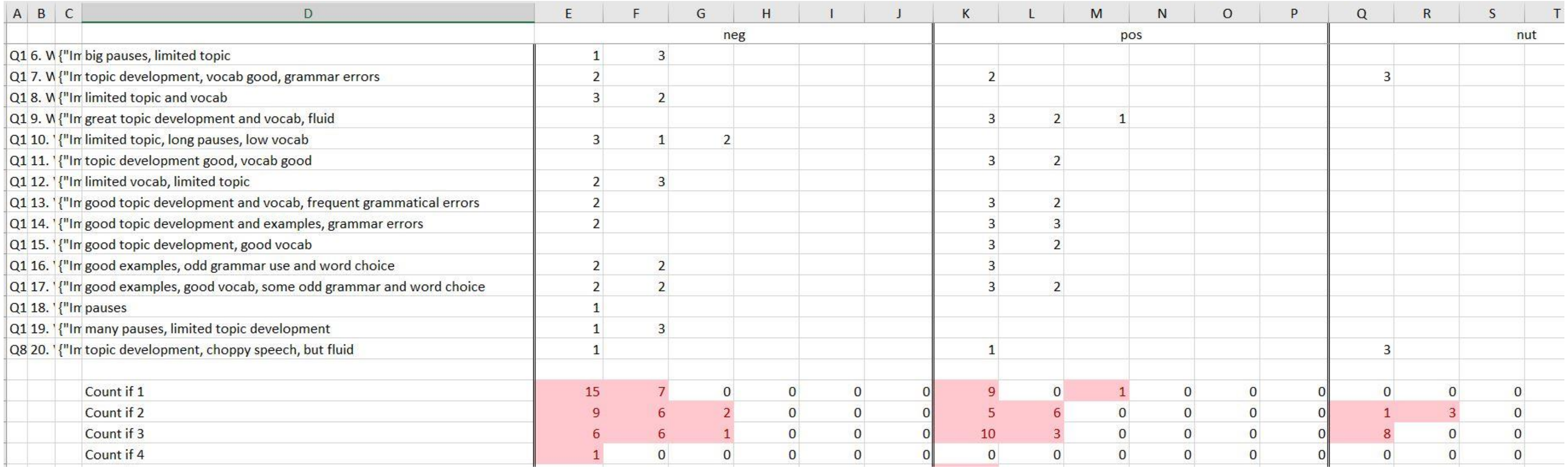
13. There were some Arabic, Chinese, and Russian speakers among the recordings that you graded, could you distinguish them?
    Other probes:
    Why?
    Can you tell me more?
    What exactly do you mean?

14. Do you think your familiarity (or lack of it) with the way Arabic, Chinese, and Russians speak (e.g., peculiarities of grammar, vocabulary, pronunciation) affected your scores?
    Other probes:
    Why? What makes you think so?
    Can you tell me more?
    What exactly do you mean?

15. Do you have any other insights, thoughts, or suggestions about improving rubrics, rater training, or exam fairness?

Appendix J

Example of Coded Comment Lines



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
				neg						pos						nut			
Q1	6. W	{"Im	big pauses, limited topic	1	3														
Q1	7. W	{"Im	topic development, vocab good, grammar errors	2						2						3			
Q1	8. W	{"Im	limited topic and vocab	3	2														
Q1	9. W	{"Im	great topic development and vocab, fluid							3	2	1							
Q1	10. '	{"Im	limited topic, long pauses, low vocab	3	1	2													
Q1	11. '	{"Im	topic development good, vocab good							3	2								
Q1	12. '	{"Im	limited vocab, limited topic	2	3														
Q1	13. '	{"Im	good topic development and vocab, frequent grammatical errors	2						3	2								
Q1	14. '	{"Im	good topic development and examples, grammar errors	2						3	3								
Q1	15. '	{"Im	good topic development, good vocab							3	2								
Q1	16. '	{"Im	good examples, odd grammar use and word choice	2	2					3									
Q1	17. '	{"Im	good examples, good vocab, some odd grammar and word choice	2	2					3	2								
Q1	18. '	{"Im	pauses	1															
Q1	19. '	{"Im	many pauses, limited topic development	1	3														
Q8	20. '	{"Im	topic development, choppy speech, but fluid	1						1						3			
			Count if 1	15	7	0	0	0	0	9	0	1	0	0	0	0	0	0	
			Count if 2	9	6	2	0	0	0	5	6	0	0	0	0	1	3	0	
			Count if 3	6	6	1	0	0	0	10	3	0	0	0	0	8	0	0	
			Count if 4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Appendix K

Coding Scheme for Rater Comments

1 - delivery
2- language use
3 - topic development
4 - general

Negative: Anything that has negative connotation even if it is mitigated. Examples: mistakes, bad accent, some lack of development, repetition, a few mistakes, (only) minor difficulties/mistakes/issues/lapses, few ideas present, lack of bad mistakes, language use had mistakes but never impaired meaning; a couple small issues with language use; a little choppy.

Positive: Anything that has positive connotation even if it is mitigated. Examples: mostly coherent, fluent enough, good accent, the ideas are fairly clearly stated; fairly understandable; Basically Intelligible; somewhat sustained; topic development ok.

Neutral: comments without negative/positive descriptors. Examples: vocab, grammar, topic, pronunciation.

1. Delivery: pronunciation, stress, intonation, listener effort, choppy, fragmented, telegraphic, pauses, hesitations, (un)intelligible/intelligibility, articulation, rhythm, pace, fluidity of expression/fluid expression, lapses, generally well-paced, understandable, impossible to understand

2. Language Use: grammar and vocabulary, grammatical structures, syntax, linking words, limited use of language, cohesive devices, transitions; complex/long sentences

3. Topic Development: content, (number / development/repetition/ vagueness/clarity/expression/specificity) of ideas, task completion, repetition of the prompt, connected to the task/sufficient to the task, elaboration (specific ideas, details, support, examples), substance, coherence (coherent/ well-organized/ sustained/ clear relationships between ideas/progression of ideas); wasted time repeating the question; great structure

4. General: good, nice, mistakes, clear (can be delivery or ideas), limited response (can be language or ideas), lack of bad mistakes (can be any kind of mistakes), limited response (limited in what?), confusing, poor, obscured,

More directions:
The following comments that can be seen in the data should be coded as follows:
Delivery: accent, confident, too fast/quick speech/pace/delivery, quiet
Topic Development: creative/unique ideas, thoughtful answer, essay-structure with introduction and conclusion; sophisticated ideas, incomplete, too many ideas

Two or more features connected by *and* or *with* should be coded separately.
Examples:
pauses and hesitation 1 1
poor grammar and lack of vocabulary 2 2
lack of reasons and examples 3 3
some grammar and pronunciation mistakes 2 1
That's a nice response, the speaker doesn't come across as having any trouble expressing her ideas. 4 3

Additional note: change the color (use red) of the words in the comments that are difficult to classify/cause hesitations/can be seen from two perspectives/etc.

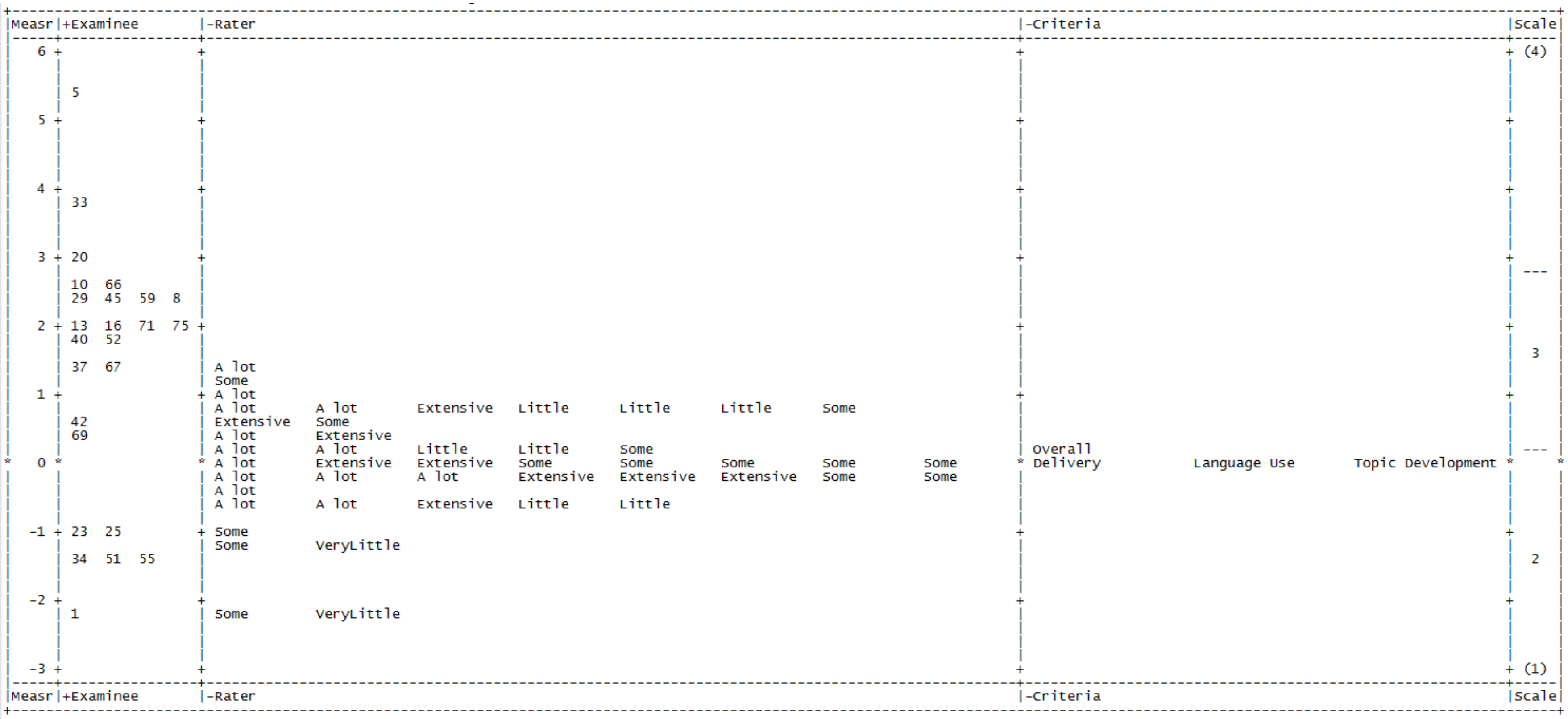
Appendix L

Coding Scheme for Qualitative Data

Rater	Perceived severity	Category importance	Listening	Grading	Biases/Beliefs	Concerns
NS34 Can distinguish all, very familiar with A and R, but not that much with C	Lenient, bc of background, willing to understand, strategic communicator Lenient on LU	TD and D, especially TD, not much about LU	Take notes about everything, glance at the rubric looking for keywords Challenging pronunciation – re-listen to adapt to the accent	Overall impression (sure) or piece by piece (hesitant) Clear cut - Sure -glance at the rubric, Hesitant – read through the rubric and give partial scores First TD, then D	Uses the word accent and accented, but distinguishes accentedness and intelligibility/ comprehensibility Potential biases – voice quality and Russian accents which are endearing Doesn't take off his ESL hat These tasks are so easy that they do not need complex grammar, it is just the matter of the amount of errors Complex structures in writing and speaking are not the same Organization matters, there should be a thesis statement, and linked ideas around it because Academic English has to be organized Hypothetical biases: current state physical or emotional (e.g., thirsty) Hypothetical biases: Experience with language populations Hypothetical biases: sympathy Hypothetical biases: experience teaching based on toefl prep books Hypothetical biases: mic or recording issues, background noise Unfinished discourse should not be a factor Not bothered by fillers Fresh ideas – no Can be biased because of the negative L1 transfer of Chinese No familiarity -biased	Challenging when the ability in D, LU and TD are mixed What are we grading? Speaking or performance?

Appendix M

Familiarity Variable Maps per L1 (6 groups and 2 groups)

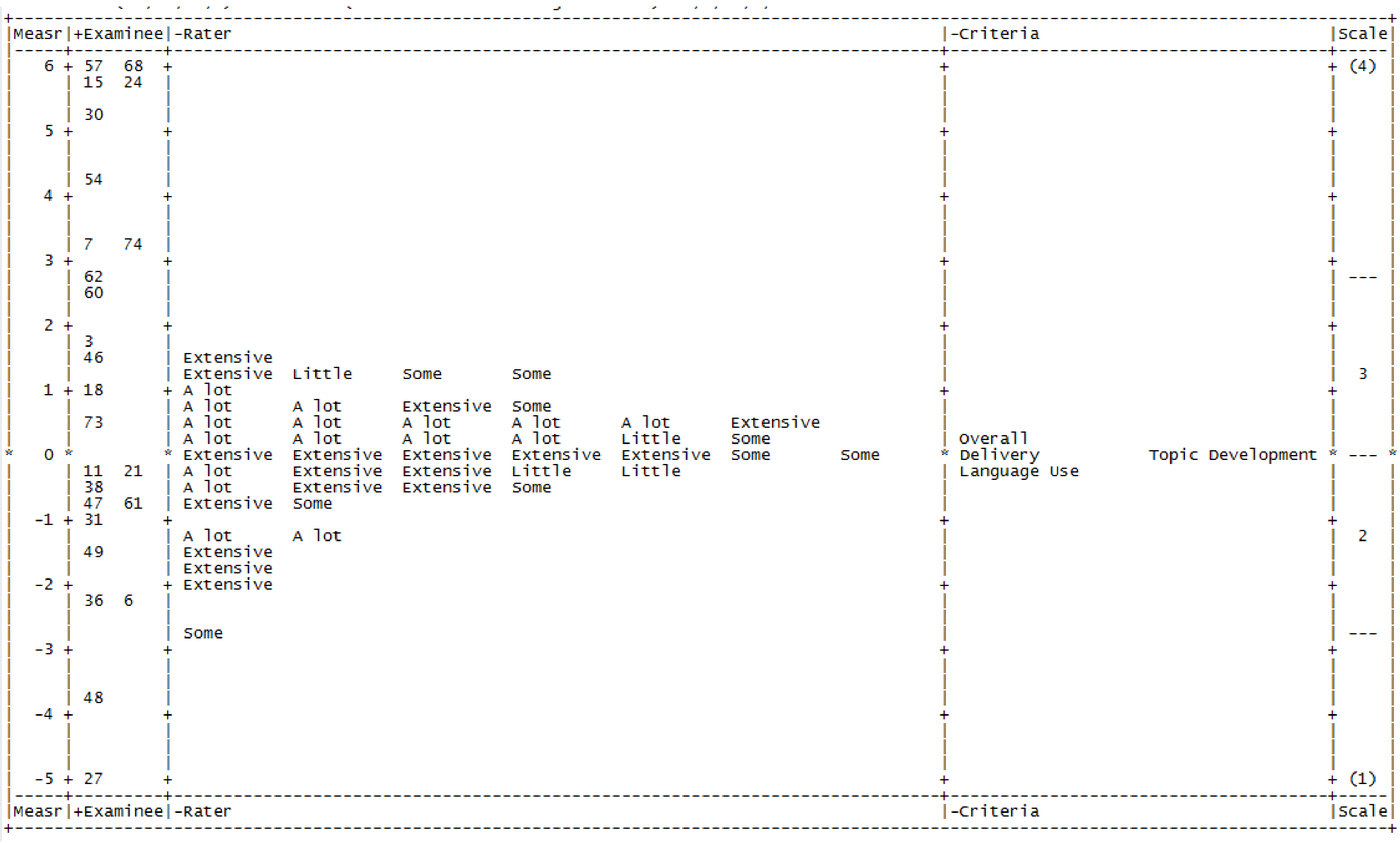


```
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee       |-Rater                                                                                          |-Criteria                                               |Scale|
|-----+----------------+------------------------------------------------------------------------------------------------+--------------------------------------------------------+-----|
|   6 +                +                                                                                                +                                                        + (4) |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     | 5              |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|   5 +                +                                                                                                +                                                        +     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|   4 +                +                                                                                                +                                                        +     |
|     | 33             |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|   3 + 20             +                                                                                                +                                                        +     |
|     |                |                                                                                                |                                                        | --- |
|     | 10  66         |                                                                                                |                                                        |     |
|     | 29  45  59  8  |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|   2 + 13  16  71  75 +                                                                                                +                                                        +     |
|     | 40  52         |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |  3  |
|     | 37  67         | A lot                                                                                          |                                                        |     |
|     |                | Some                                                                                           |                                                        |     |
|   1 +                + A lot                                                                                          +                                                        +     |
|     |                | A lot       A lot       Extensive   Little      Little      Little      Some                   |                                                        |     |
|     | 42             | Extensive   Some                                                                               |                                                        |     |
|     | 69             | A lot       Extensive                                                                          |                                                        |     |
|     |                | A lot       A lot       Little      Little      Some                                           | Overall                                                | --- |
*   0 *                * A lot       Extensive   Extensive   Some        Some        Some        Some        Some       * Delivery           Language Use       Topic Development*     *
|     |                | A lot       A lot       A lot       Extensive   Extensive   Extensive   Some        Some       |                                                        |     |
|     |                | A lot                                                                                          |                                                        |     |
|     |                | A lot       A lot       Extensive   Little      Little                                         |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|  -1 + 23  25         + Some                                                                                           +                                                        +     |
|     |                | Some        VeryLittle                                                                         |                                                        |     |
|     | 34  51  55     |                                                                                                |                                                        |  2  |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|  -2 +                +                                                                                                +                                                        +     |
|     | 1              | Some        VeryLittle                                                                         |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|     |                |                                                                                                |                                                        |     |
|  -3 +                +                                                                                                +                                                        + (1) |
|-----+----------------+------------------------------------------------------------------------------------------------+--------------------------------------------------------+-----|
|Measr|+Examinee       |-Rater                                                                                          |-Criteria                                               |Scale|
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Arabic L1, 6 groups

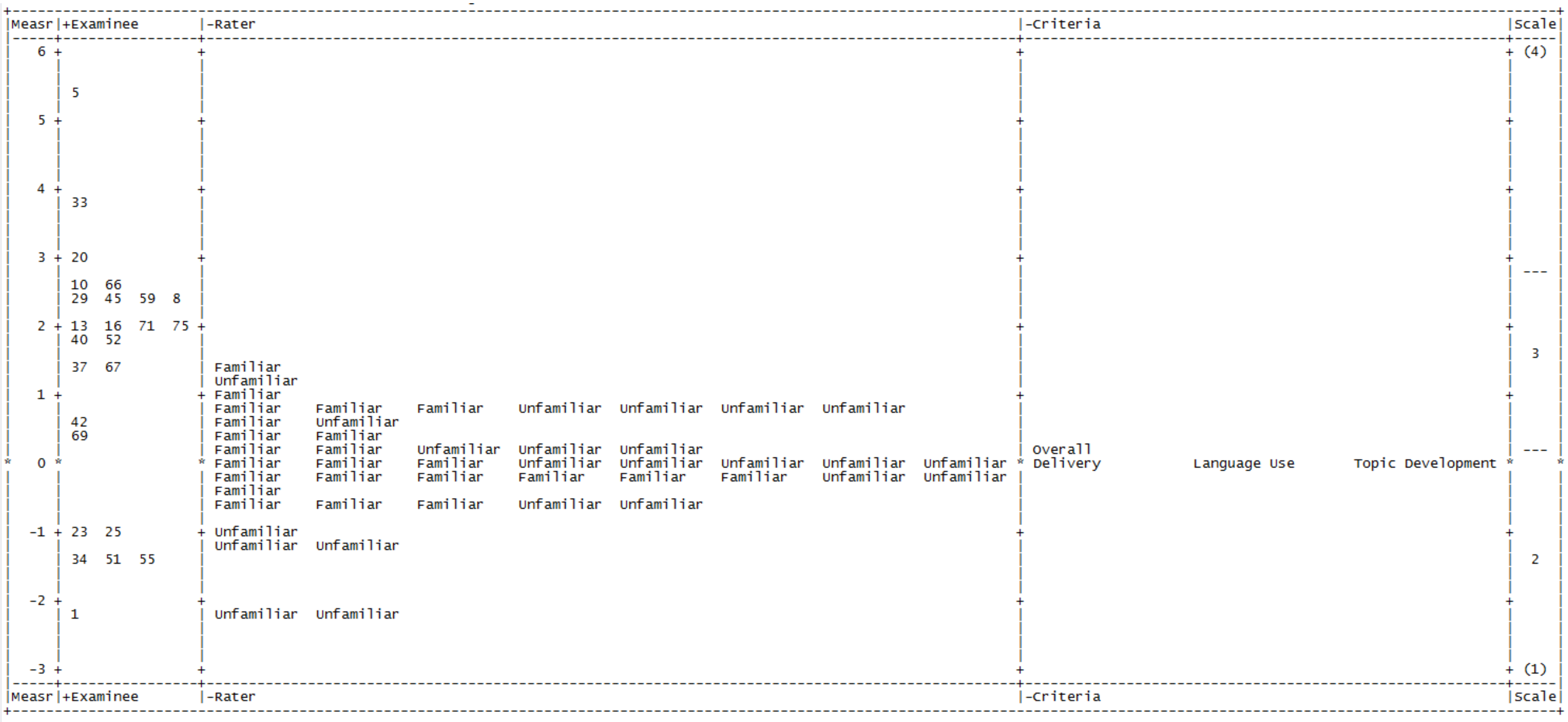
```
+-----------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee   |-Rater                                                                  |-Criteria          |Scale|
|-----+------------+------------------------------------------------------------------------+-------------------+-----|
|   5 +            +                                                                        +                   + (4) |
|     |            |                                                                        |                   |     |
|     | 14         |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|   4 +            +                                                                        +                   +     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     | 50         |                                                                        |                   |     |
|   3 + 56         +                                                                        +                   + --- |
|     | 65         |                                                                        |                   |     |
|     | 28  53     |                                                                        |                   |     |
|     | 43         |                                                                        |                   |     |
|     | 63         |                                                                        |                   |     |
|   2 + 22  32  72 +                                                                        +                   +     |
|     | 2          | Extensive   Little                                                     |                   |     |
|     |            |                                                                        |                   |     |
|     |            | Extensive                                                              |                   |  3  |
|     | 26  44     | Some                                                                   |                   |     |
|   1 + 64         + Extensive                                                              +                   +     |
|     |            | A lot       A lot       Some                                           |                   |     |
|     |            | A lot       A lot       Little      Little      Little      Some       |                   |     |
|     | 70         | A lot       A lot       Extensive   Little      Little      Some       | Delivery          |     |
|     | 17  35     | A lot       Extensive   Extensive   Little      Some        VeryLittle | Overall           |     |
*   0 *            * Extensive                                                              * Language Use      * --- *
|     |            | A lot       A lot       Extensive   Little                             |                   |     |
|     |            | Extensive   Little                                                     |                   |     |
|     | 41  58     | A lot       Extensive   Some                                           | Topic Development |     |
|     | 12  9      | Little      Some                                                   |                   |     |
|  -1 + 4          +                                                                        +                   +     |
|     | 39         | A lot       A lot       Extensive   Some        VeryLittle             |                   |     |
|     |            |                                                                        |                   |  2  |
|     |            | A lot                                                                  |                   |     |
|     |            |                                                                        |                   |     |
|  -2 +            + Extensive   Little                                                     +                   +     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|  -3 +            +                                                                        +                   + --- |
|     | 19         |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|  -4 +            +                                                                        +                   + (1) |
|-----+------------+------------------------------------------------------------------------+-------------------+-----|
|Measr|+Examinee   |-Rater                                                                  |-Criteria          |Scale|
+-----------------------------------------------------------------------------------------------------------------------+
```

Chinese L1, 6 groups



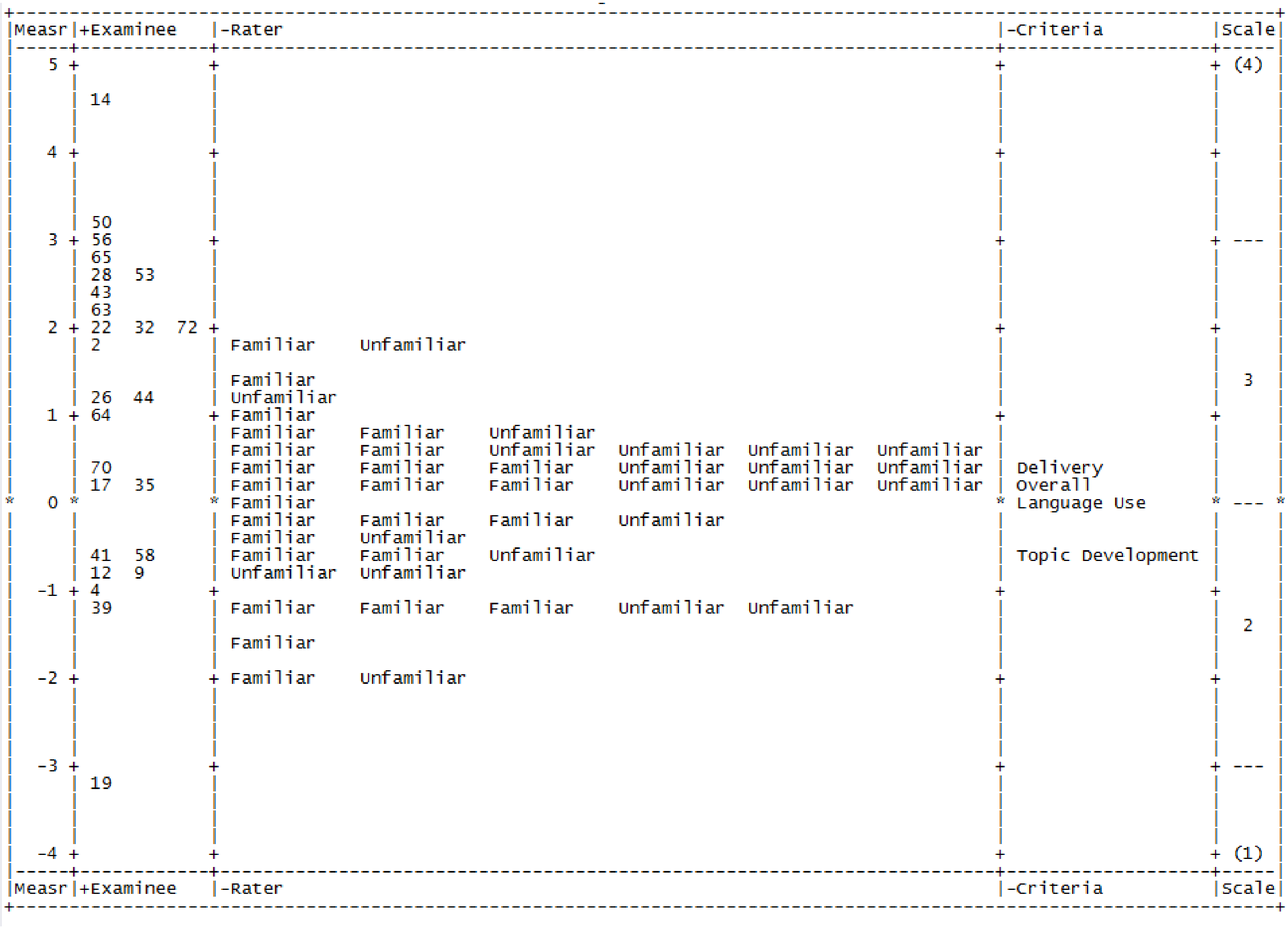
```
+------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee|-Rater                                                                       |-Criteria                             |Scale|
|-----+---------+-----------------------------------------------------------------------------+--------------------------------------+-----|
|   6 + 57  68  +                                                                             +                                      + (4) |
|     | 15  24  |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|     | 30      |                                                                             |                                      |     |
|   5 +         +                                                                             +                                      +     |
|     |         |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|     | 54      |                                                                             |                                      |     |
|   4 +         +                                                                             +                                      +     |
|     |         |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|     | 7   74  |                                                                             |                                      |     |
|   3 +         +                                                                             +                                      +     |
|     | 62      |                                                                             |                                      | --- |
|     | 60      |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|   2 +         +                                                                             +                                      +     |
|     | 3       |                                                                             |                                      |     |
|     | 46      | Extensive                                                                   |                                      |     |
|     |         | Extensive  Little     Some       Some                                       |                                      |  3  |
|   1 + 18      + A lot                                                                       +                                      +     |
|     |         | A lot      A lot      Extensive  Some                                       |                                      |     |
|     | 73      | A lot      A lot      A lot      A lot      A lot      Extensive            |                                      |     |
|     |         | A lot      A lot      A lot      A lot      Little     Some                 | Overall                              |     |
*   0 *         * Extensive  Extensive  Extensive  Extensive  Extensive  Some       Some      * Delivery           Topic Development * --- *
|     | 11  21  | A lot      Extensive  Extensive  Little     Little                          | Language Use                         |     |
|     | 38      | A lot      Extensive  Extensive  Some                                       |                                      |     |
|     | 47  61  | Extensive  Some                                                             |                                      |     |
|  -1 + 31      +                                                                             +                                      +     |
|     |         | A lot      A lot                                                            |                                      |  2  |
|     | 49      | Extensive                                                                   |                                      |     |
|     |         | Extensive                                                                   |                                      |     |
|  -2 +         + Extensive                                                                   +                                      +     |
|     | 36  6   |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|     |         | Some                                                                        |                                      | --- |
|  -3 +         +                                                                             +                                      +     |
|     |         |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|     | 48      |                                                                             |                                      |     |
|  -4 +         +                                                                             +                                      +     |
|     |         |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|     |         |                                                                             |                                      |     |
|  -5 + 27      +                                                                             +                                      + (1) |
|-----+---------+-----------------------------------------------------------------------------+--------------------------------------+-----|
|Measr|+Examinee|-Rater                                                                       |-Criteria                             |Scale|
+------------------------------------------------------------------------------------------------------------------------------------------+
```

Russian L1, 6 groups



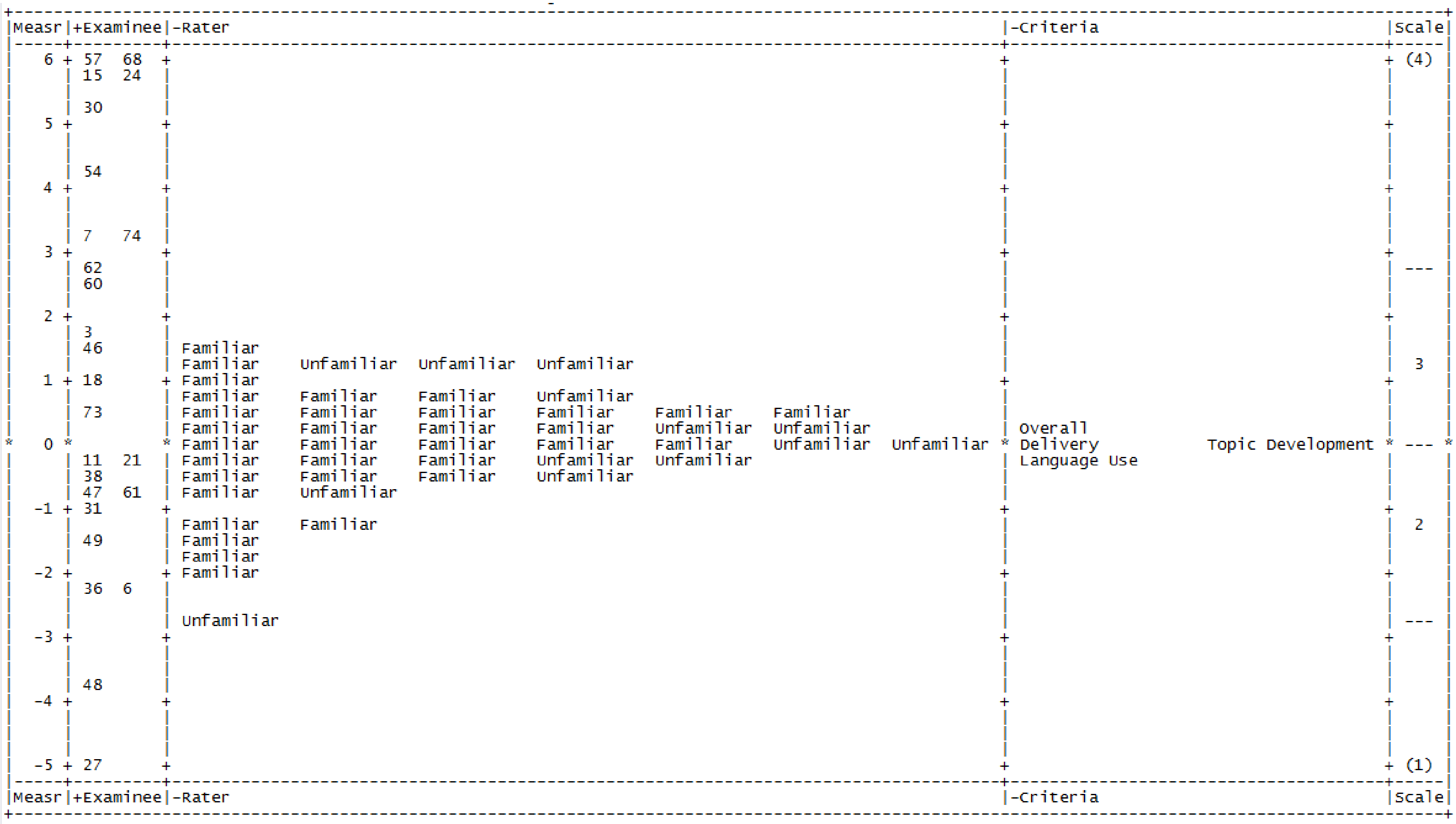
```
+----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee        |-Rater                                                                                            |-Criteria                                                  |Scale|
|-----+-----------------+--------------------------------------------------------------------------------------------------+-----------------------------------------------------------+-----|
|   6 +                 +                                                                                                  +                                                           + (4) |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     | 5               |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|   5 +                 +                                                                                                  +                                                           +     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|   4 +                 +                                                                                                  +                                                           +     |
|     | 33              |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|   3 + 20              +                                                                                                  +                                                           +     |
|     |                 |                                                                                                  |                                                           | --- |
|     | 10  66          |                                                                                                  |                                                           |     |
|     | 29  45  59  8   |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|   2 + 13  16  71  75  +                                                                                                  +                                                           +     |
|     | 40  52          |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |  3  |
|     | 37  67          | Familiar                                                                                         |                                                           |     |
|     |                 | Unfamiliar                                                                                       |                                                           |     |
|   1 +                 + Familiar                                                                                         +                                                           +     |
|     |                 | Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar  Unfamiliar  Unfamiliar               |                                                           |     |
|     | 42              | Familiar    Unfamiliar                                                                           |                                                           |     |
|     | 69              | Familiar    Familiar                                                                             |                                                           |     |
|     |                 | Familiar    Familiar    Unfamiliar  Unfamiliar  Unfamiliar                                       | Overall                                                   | --- |
*   0 *                 * Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar  Unfamiliar  Unfamiliar  Unfamiliar   * Delivery            Language Use        Topic Development *     *
|     |                 | Familiar    Familiar    Familiar    Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar   |                                                           |     |
|     |                 | Familiar                                                                                         |                                                           |     |
|     |                 | Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar                                       |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|  -1 + 23  25          + Unfamiliar                                                                                       +                                                           +     |
|     |                 | Unfamiliar  Unfamiliar                                                                           |                                                           |     |
|     | 34  51  55      |                                                                                                  |                                                           |  2  |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|  -2 +                 +                                                                                                  +                                                           +     |
|     | 1               | Unfamiliar  Unfamiliar                                                                           |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|     |                 |                                                                                                  |                                                           |     |
|  -3 +                 +                                                                                                  +                                                           + (1) |
|-----+-----------------+--------------------------------------------------------------------------------------------------+-----------------------------------------------------------+-----|
|Measr|+Examinee        |-Rater                                                                                            |-Criteria                                                  |Scale|
+----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Arabic L1, 2 groups



```
+---------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee   |-Rater                                                                  |-Criteria          |Scale|
|-----+------------+------------------------------------------------------------------------+-------------------+-----|
|   5 +            +                                                                        +                   + (4) |
|     |            |                                                                        |                   |     |
|     | 14         |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|   4 +            +                                                                        +                   +     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     | 50         |                                                                        |                   |     |
|   3 + 56         +                                                                        +                   + --- |
|     | 65         |                                                                        |                   |     |
|     | 28  53     |                                                                        |                   |     |
|     | 43         |                                                                        |                   |     |
|     | 63         |                                                                        |                   |     |
|   2 + 22  32  72 +                                                                        +                   +     |
|     | 2          | Familiar    Unfamiliar                                                 |                   |     |
|     |            |                                                                        |                   |     |
|     |            | Familiar                                                               |                   |  3  |
|     | 26  44     | Unfamiliar                                                             |                   |     |
|   1 + 64         + Familiar                                                               +                   +     |
|     |            | Familiar    Familiar    Unfamiliar                                     |                   |     |
|     |            | Familiar    Familiar    Unfamiliar  Unfamiliar  Unfamiliar  Unfamiliar |                   |     |
|     | 70         | Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar  Unfamiliar | Delivery          |     |
|     | 17  35     | Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar  Unfamiliar | Overall           |     |
*   0 *            * Familiar                                                               * Language Use      * --- *
|     |            | Familiar    Familiar    Familiar    Unfamiliar                         |                   |     |
|     |            | Familiar    Unfamiliar                                                 |                   |     |
|     | 41  58     | Familiar    Familiar    Unfamiliar                                     | Topic Development |     |
|     | 12  9      | Unfamiliar  Unfamiliar                                                 |                   |     |
|  -1 + 4          +                                                                        +                   +     |
|     | 39         | Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar             |                   |     |
|     |            |                                                                        |                   |  2  |
|     |            | Familiar                                                               |                   |     |
|     |            |                                                                        |                   |     |
|  -2 +            + Familiar    Unfamiliar                                                 +                   +     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|  -3 +            +                                                                        +                   + --- |
|     | 19         |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|     |            |                                                                        |                   |     |
|  -4 +            +                                                                        +                   + (1) |
|-----+------------+------------------------------------------------------------------------+-------------------+-----|
|Measr|+Examinee   |-Rater                                                                  |-Criteria          |Scale|
+---------------------------------------------------------------------------------------------------------------------+
```

Chinese L1, 2 groups



```
+-------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Examinee|-Rater                                                                              |-Criteria                             |Scale|
|-----+---------+------------------------------------------------------------------------------------+--------------------------------------+-----|
|   6 + 57   68 +                                                                                    +                                      + (4) |
|     | 15   24 |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|     | 30      |                                                                                    |                                      |     |
|   5 +         +                                                                                    +                                      +     |
|     |         |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|     | 54      |                                                                                    |                                      |     |
|   4 +         +                                                                                    +                                      +     |
|     |         |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|     | 7   74  |                                                                                    |                                      |     |
|   3 +         +                                                                                    +                                      +     |
|     | 62      |                                                                                    |                                      | --- |
|     | 60      |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|   2 +         +                                                                                    +                                      +     |
|     | 3       |                                                                                    |                                      |     |
|     | 46      | Familiar                                                                           |                                      |     |
|     |         | Familiar    Unfamiliar  Unfamiliar  Unfamiliar                                     |                                      |  3  |
|   1 + 18      + Familiar                                                                           +                                      +     |
|     |         | Familiar    Familiar    Familiar    Unfamiliar                                     |                                      |     |
|     | 73      | Familiar    Familiar    Familiar    Familiar    Familiar    Familiar               |                                      |     |
|     |         | Familiar    Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar             | Overall                              |     |
*   0 *         * Familiar    Familiar    Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar * Delivery           Topic Development * --- *
|     | 11   21 | Familiar    Familiar    Familiar    Unfamiliar  Unfamiliar                         | Language Use                         |     |
|     | 38      | Familiar    Familiar    Familiar    Unfamiliar                                     |                                      |     |
|     | 47   61 | Familiar    Unfamiliar                                                             |                                      |     |
|  -1 + 31      +                                                                                    +                                      +     |
|     |         | Familiar    Familiar                                                               |                                      |  2  |
|     | 49      | Familiar                                                                           |                                      |     |
|     |         | Familiar                                                                           |                                      |     |
|  -2 +         + Familiar                                                                           +                                      +     |
|     | 36  6   |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|     |         | Unfamiliar                                                                         |                                      | --- |
|  -3 +         +                                                                                    +                                      +     |
|     |         |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|     | 48      |                                                                                    |                                      |     |
|  -4 +         +                                                                                    +                                      +     |
|     |         |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|     |         |                                                                                    |                                      |     |
|  -5 + 27      +                                                                                    +                                      + (1) |
|-----+---------+------------------------------------------------------------------------------------+--------------------------------------+-----|
|Measr|+Examinee|-Rater                                                                              |-Criteria                             |Scale|
+-------------------------------------------------------------------------------------------------------------------------------------------------+
```

Russian L1, 2 groups

Appendix N

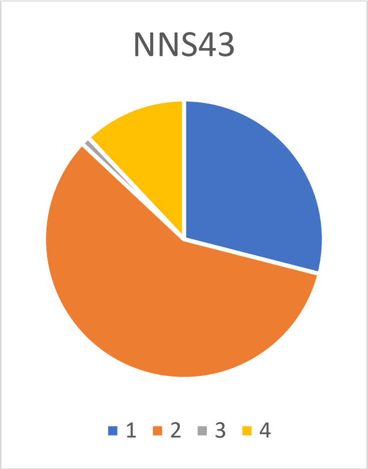
Familiar and unfamiliar Raters per Examinee L1

	Arabic L1	Composite familiarity score	Chinese L1	Composite familiarity score	Russian L1	Composite familiarity score
Unfamiliar	NNS06	16	NNS06	12	NS05	25
	NNS10	18	NNS39	19	NS16	28
	NNS31	24	NNS10	23	NS04	32
	NNS39	24	NNS42	25	NS11	32
	NNS22	29	NNS22	27	NS36	35
	NNS42	30	NNS24	28	NS41	36
	NNS45	31	NNS02	29	NS29	37
	NNS02	32	NNS28	30	NS25	38
	NNS38	33	NNS43	31	NNS10	41
	NNS43	34	NNS45	31	NS12	41
	NNS24	35	NNS08	32	NS21	43
	NNS28	36	NNS46	32	NNS31	43
	NNS08	37	NNS38	33	NS33	44
	NS11	37	NNS23	34		
	NNS13	37	NNS13	37		
	NNS23	38	NS33	39		
	NNS46	38	NNS15	40		
	NNS15	39	NNS01	41		
	NS04	42	NNS19	42		
	NS12	42	NNS35	42		
	NS16	42	NNS27	44		
	NNS01	43				
	NNS36	44				
Total	23		21		13	
Familiar	NS05	45	NS05	46	NS17	45
	NS41	45	NS37	48	NNS43	45
	NNS19	46	NS04	51	NS09	47
	NS21	47	NS11	51	NS34	47
	NS29	47	NS12	51	NNS39	47
	NNS32	49	NS25	51	NS14	48
	NNS35	49	NS34	52	NS30	50
	NS17	51	NS09	53	NS03	51
	NS34	51	NS16	53	NS07	51
	NS37	52	NS21	53	NNS02	53
	NNS27	53	NNS32	53	NNS22	53
	NS07	54	NS07	55	NNS27	53
	NS09	54	NS20	55	NS37	54
	NS20	55	NNS31	56	NNS38	54
	NS25	56	NS03	57	NS20	55
	NS14	58	NS41	57	NS26	55
	NS03	60	NS36	58	NNS13	56
	NS44	60	NS29	60	NNS28	56
	NS30	62	NS14	62	NNS23	57
	NS33	62	NS17	62	NNS35	58
	NNS18	64	NS30	62	NS40	58
	NS40	65	NNS18	64	NNS06	59
	NS26	66	NS44	64	NNS45	59
			NS26	66	NNS46	59
			NS40	66	NNS08	60
					NNS19	60
					NNS24	60
					NS44	62
					NNS01	63
					NNS32	63
					NNS15	64
					NNS18	66
					NNS42	66
Total	23		25		33	

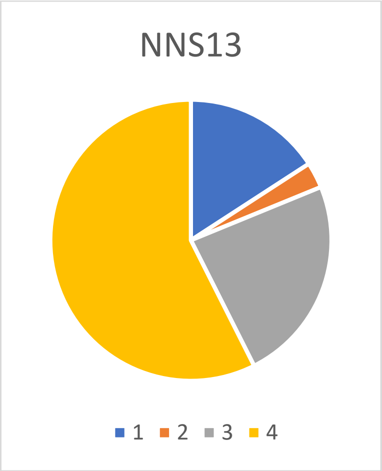
Appendix O

Rater Pie-Charts Based on Criteria Attention in Comments

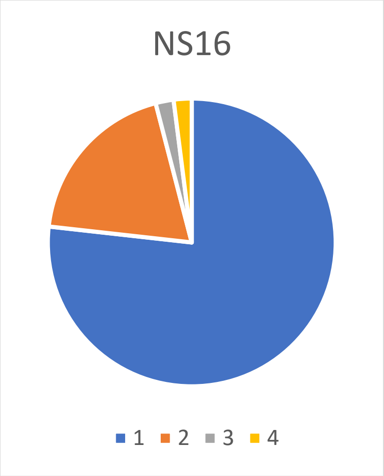
Language Use-oriented

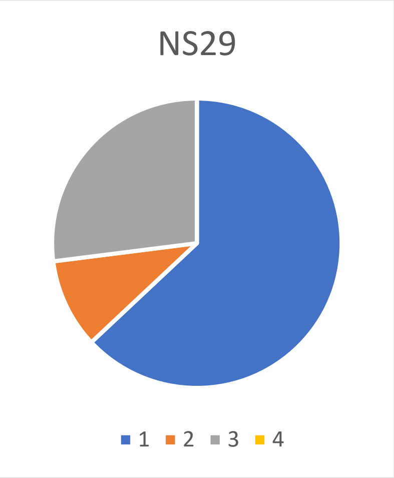


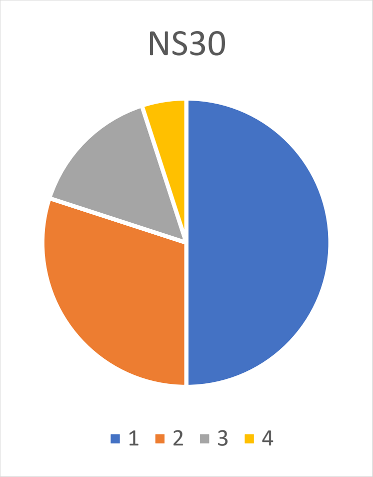
General-oriented

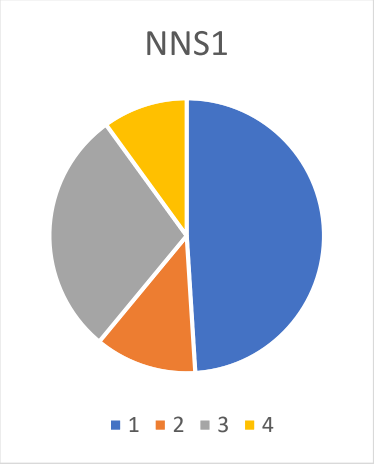


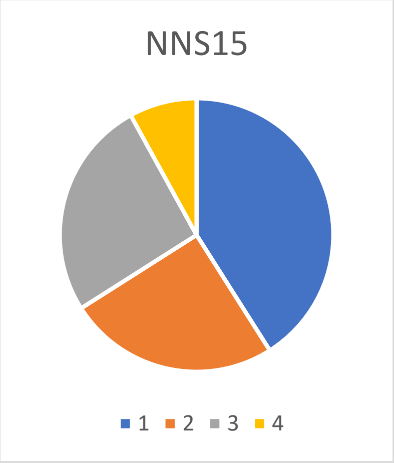
Delivery-oriented

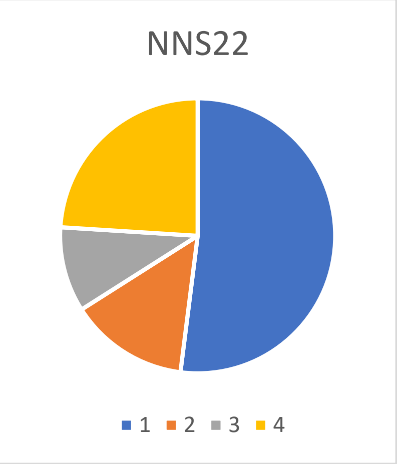

NS16
1 2 3 4

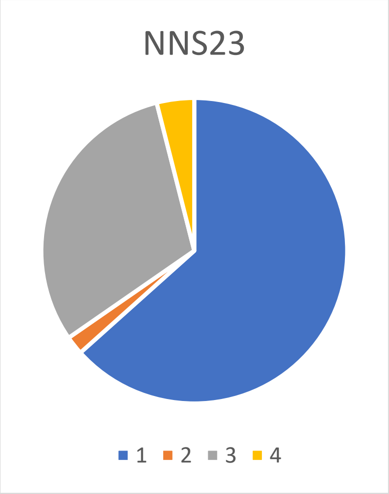

NS29
1 2 3 4

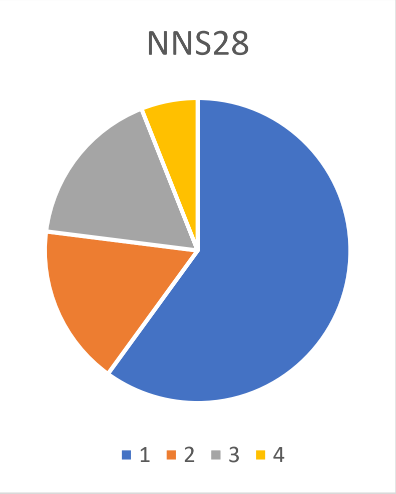

NS30
1 2 3 4

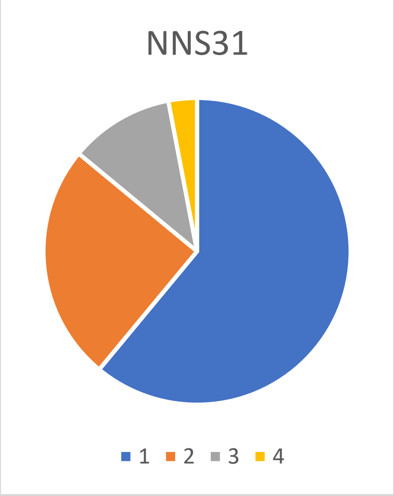

NNS1
1 2 3 4

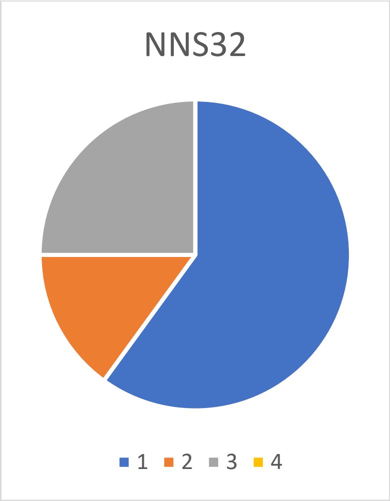

NNS15
1 2 3 4

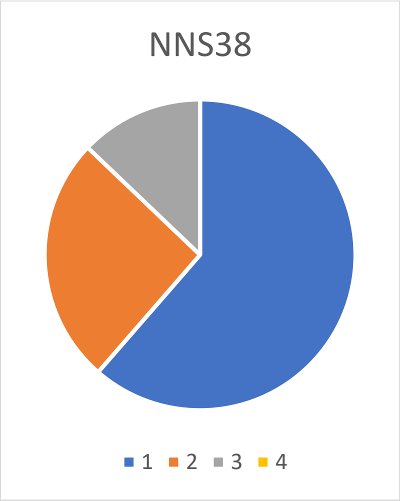

NNS22
1 2 3 4

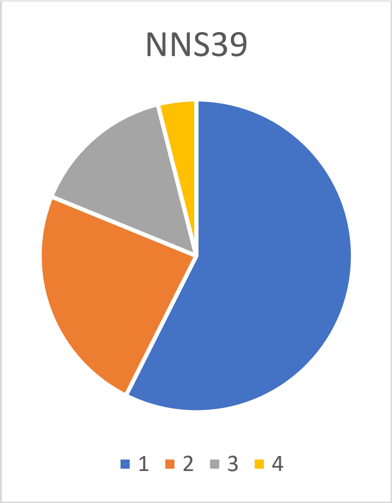

NNS23
1 2 3 4

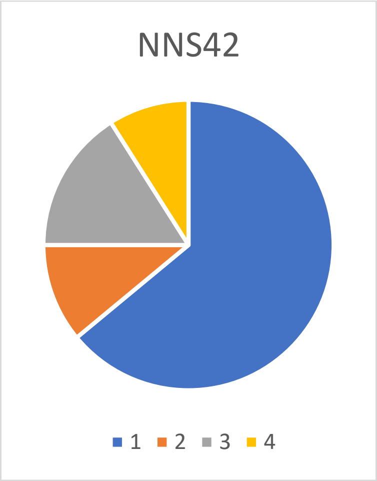

NNS28
1 2 3 4


NNS31
1 2 3 4

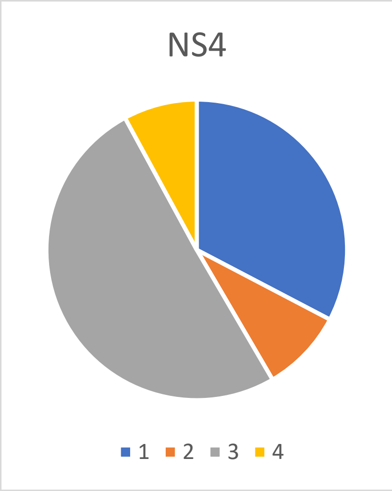


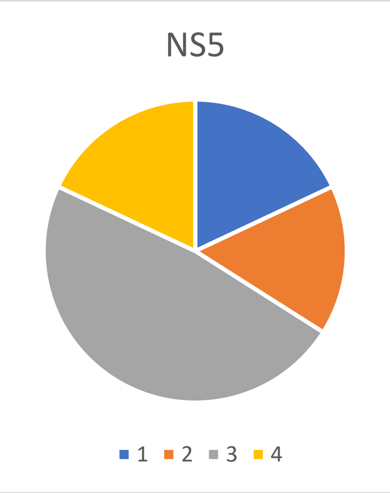


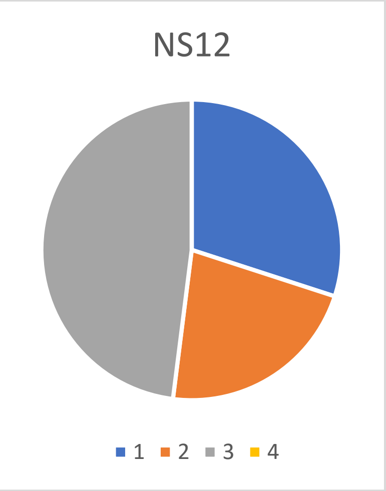


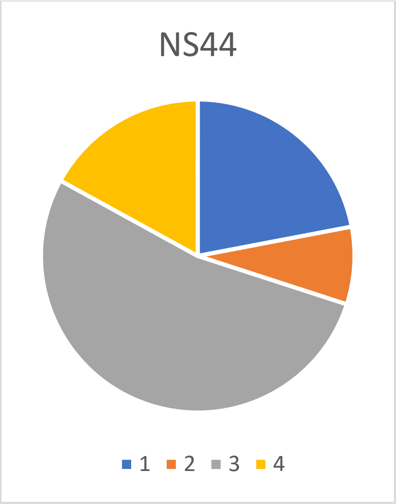


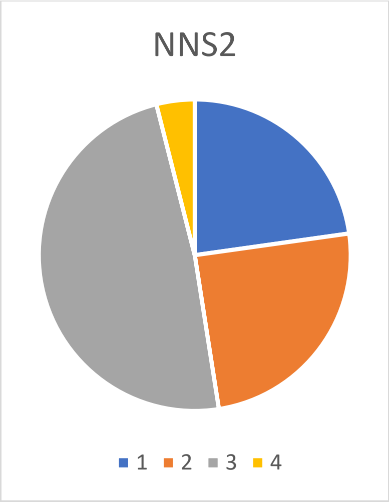
Topic Development-oriented:

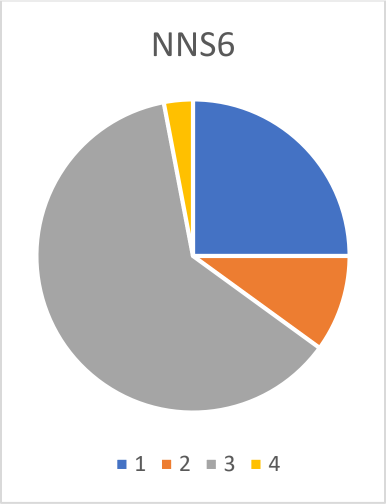


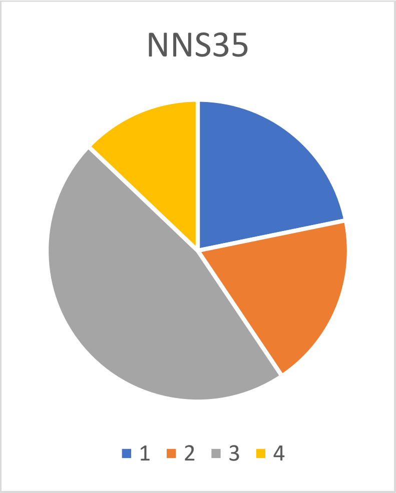


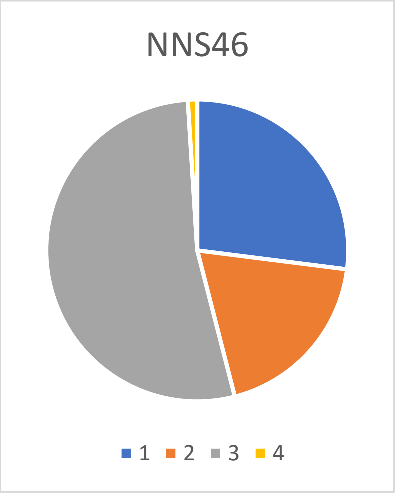




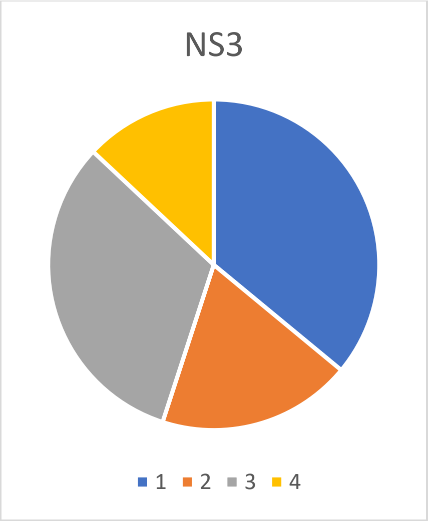


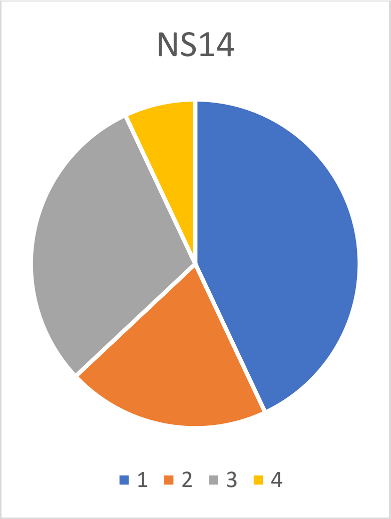


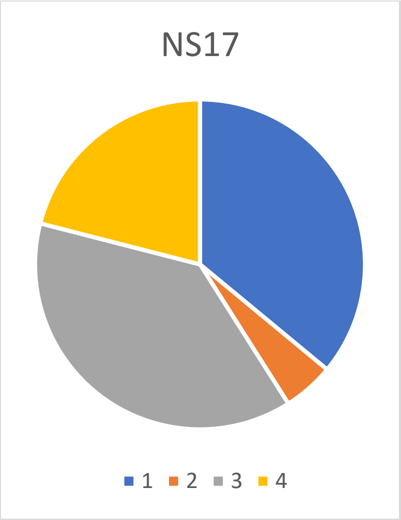


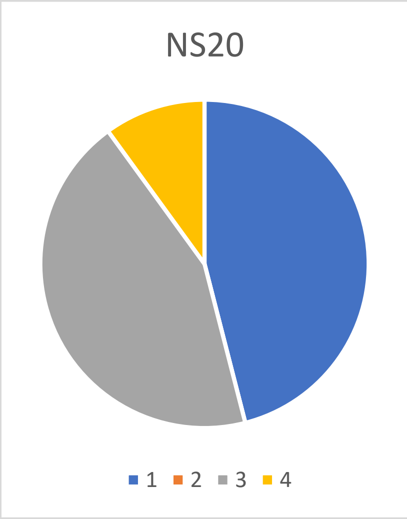


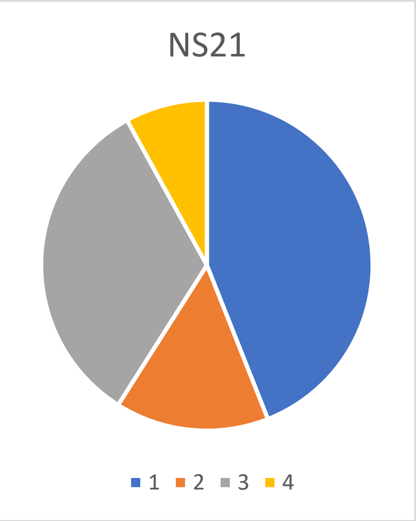
Two categories-oriented

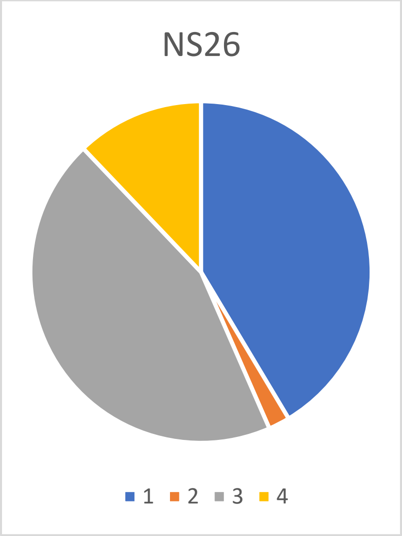

NS3
1
2
3
4

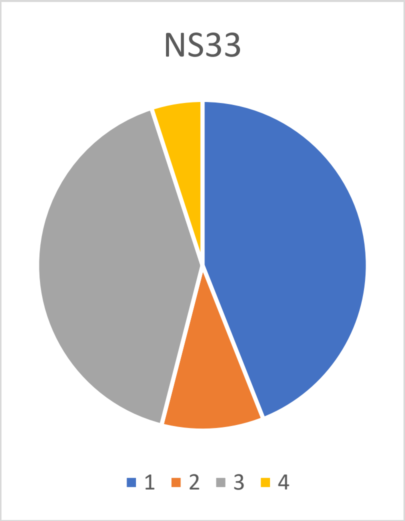

NS14
1
2
3
4

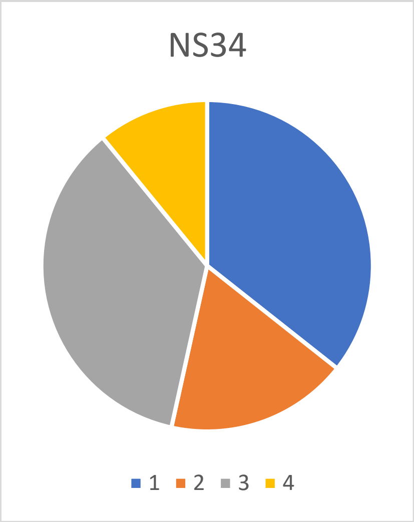

NS17
1
2
3
4

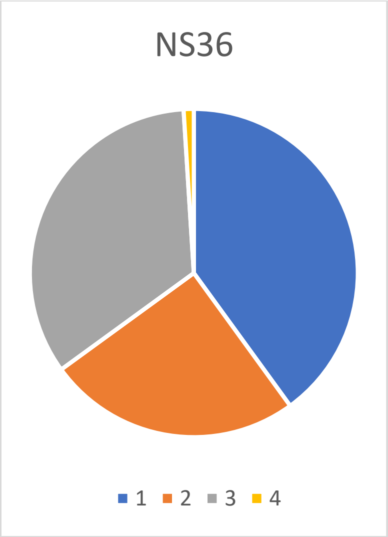

NS20
1
2
3
4

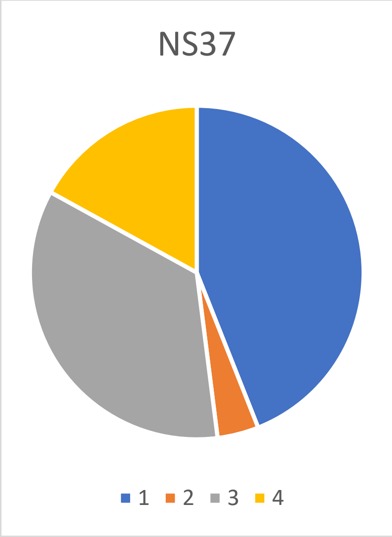

NS21
1
2
3
4

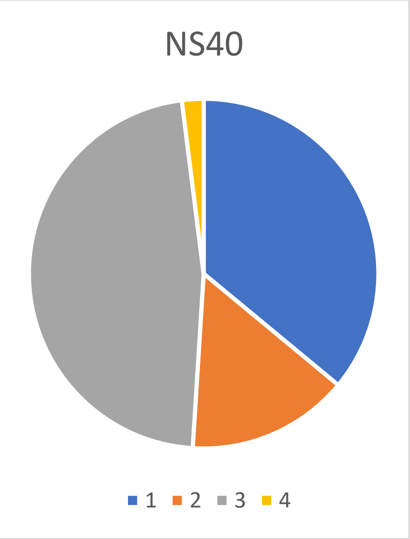

NS26
1
2
3
4

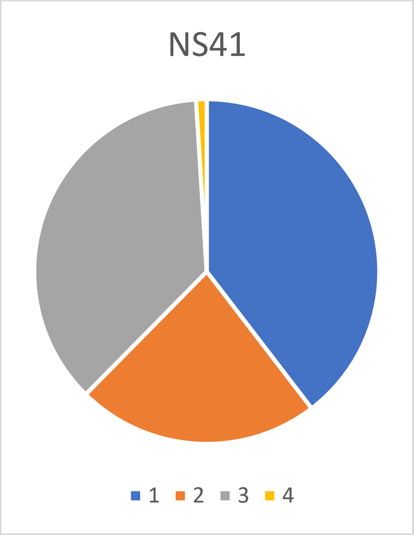

NS33
■ 1 ■ 2 ■ 3 ■ 4

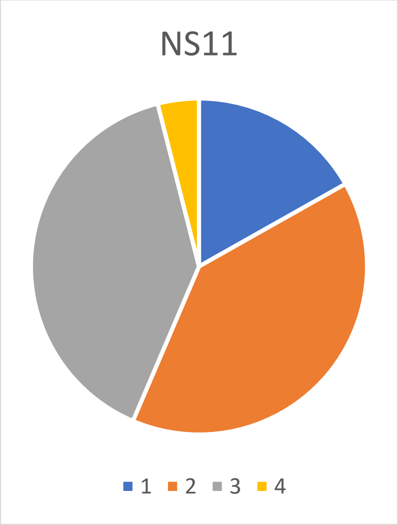

NS34
■ 1 ■ 2 ■ 3 ■ 4

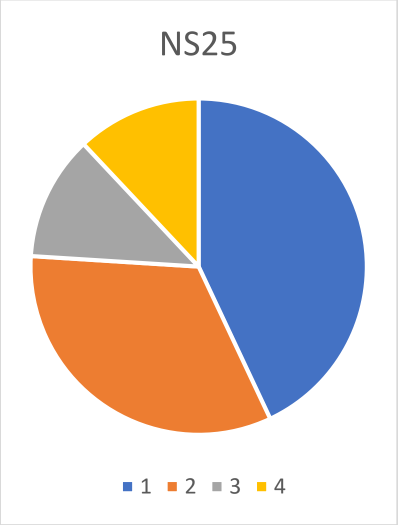

NS36
■ 1 ■ 2 ■ 3 ■ 4

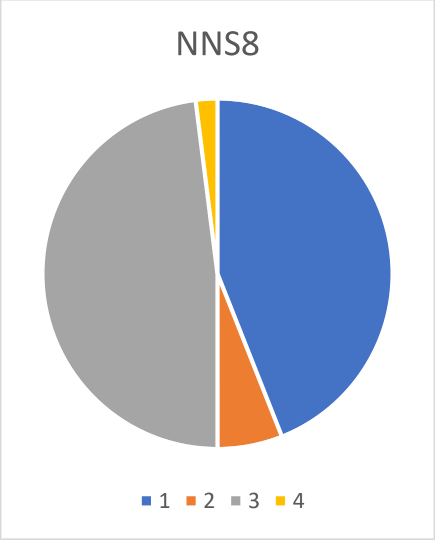

NS37
■ 1 ■ 2 ■ 3 ■ 4

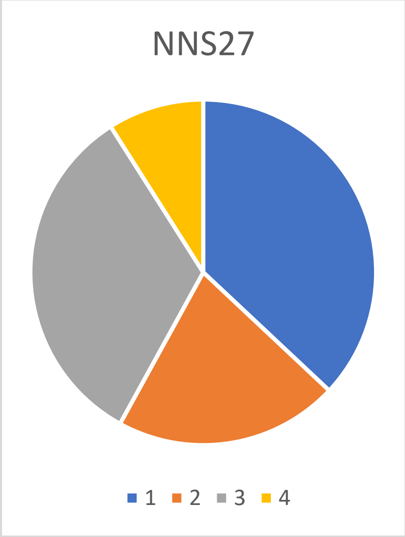

NS40
■ 1 ■ 2 ■ 3 ■ 4

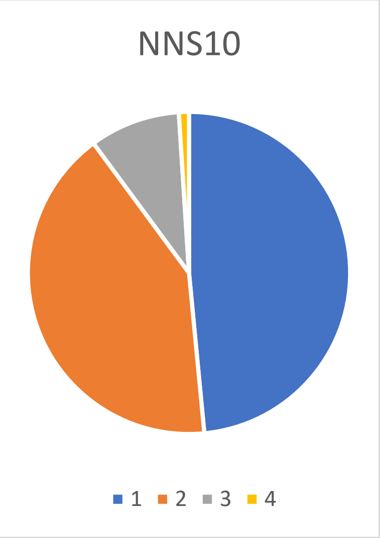

NS41
■ 1 ■ 2 ■ 3 ■ 4


NS11
■ 1 ■ 2 ■ 3 ■ 4

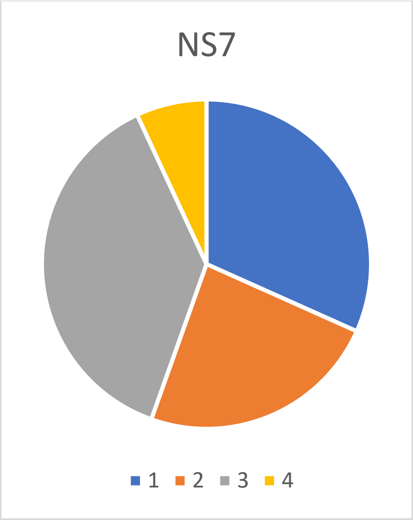

NS25
■ 1 ■ 2 ■ 3 ■ 4

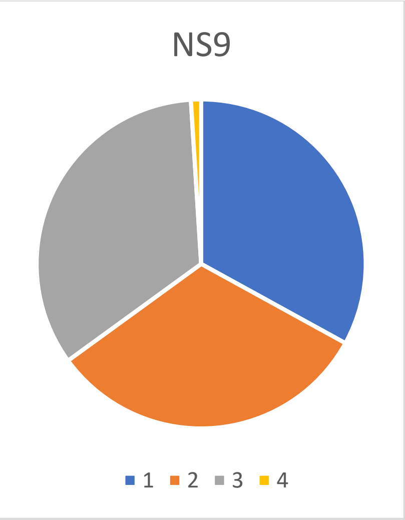


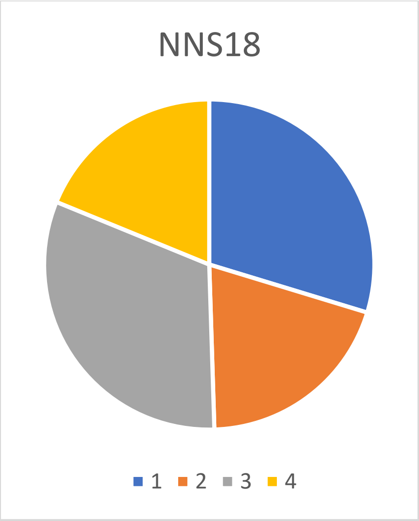


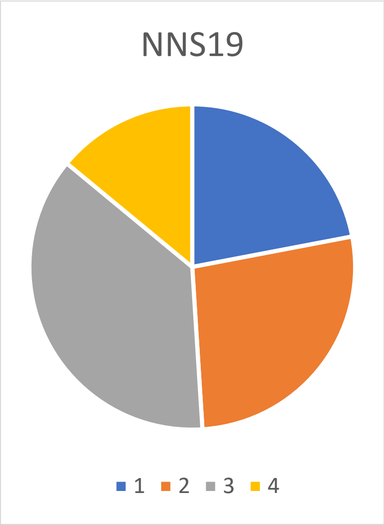


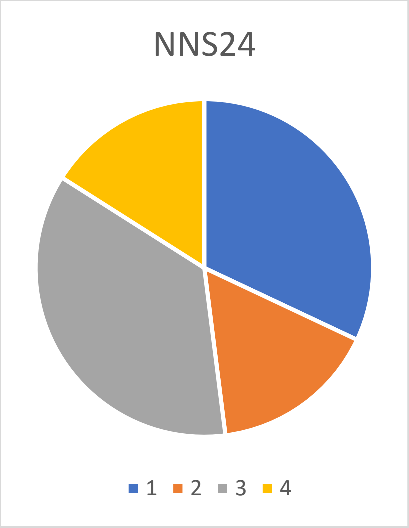
Balanced

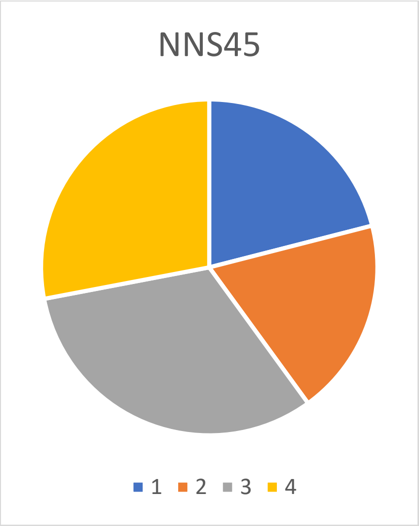







NNS19
1
2
3
4


NNS24
1
2
3
4


NNS45
1
2
3
4