RATER EXPERTISE IN A SECOND LANGUAGE SPEAKING ASSESSMENT:
THE INFLUENCE OF TRAINING AND EXPERIENCE

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

SECOND LANGUAGE STUDIES

DECEMBER 2012

By

Lawrence Edward Davis

Dissertation Committee:

John Norris, Chairperson
James Dean Brown
Thom Hudson
Luca Onnis
Thomas Hilgers
For Donna Seto Davis
ACKNOWLEDGMENTS

Many individuals made important contributions to my dissertation project. First, I thank my dissertation committee for many useful comments and suggestions made during the process. I particularly thank Professor Luca Onnis who introduced me to statistical learning and behavioral economics (domains which informed the final section of the study) and Professor J.D. Brown for providing publication opportunities outside of the dissertation process, and for his gentle corrections in the face of my many transgressions against APA style. In addition, John Davis, Hanbyul Jung, Aleksandra Malicka, Kyae-Sung Park, John Norris, and Veronika Timpe assisted in the piloting of instruments used in the study; their comments and observations did much to improve the data collection procedures.

I am grateful to Xiaoming Xi at Educational Testing Service (ETS) for her help in obtaining access to the TOEFL iBT Public Use Dataset and to Pam Mollaun at ETS for her help in recruiting TOEFL scoring leaders to provide reference scores. I also thank the eleven ETS scoring leaders who provided additional scores for responses from the TOEFL iBT Public Use Dataset. A number of individuals assisted in recruiting participants for the study, including Barbara Davidson, Jennifer Hickman, Aurora Tsai, and Mark Wolfersberger. Although they must remain anonymous, I extend my heartfelt thanks to those individuals who participated as raters in the study. It goes without saying that there would be no dissertation without them, and I appreciate their efforts in completing a long list of tedious tasks, with minimal compensation.

My dissertation advisor, Professor John Norris, played a key role in all aspects of the project. This dissertation has benefitted greatly from his broad knowledge and sage
advice. Even more, John has been an immensely patient and attentive mentor, and provided me with opportunities to learn, work, and publish that have been crucial to my development as a professional.

The study would not have been possible without financial support. I thank the Department of Second Language Studies and the College of Languages, Linguistics, and Literatures at the University of Hawai‘i for providing graduate assistantships throughout my Ph.D program. Data collection was directly supported by a TOEFL Small Dissertation Grant from Educational Testing Service and a dissertation grant from the International Research Foundation for English Language Education (TIRF). Collection of data and writing of the dissertation was also greatly facilitated by a dissertation completion fellowship from the Bilinski Educational Foundation. I thank ETS, TIRF, and the Bilinski Foundation for their very generous support.

Finally, I thank the many friends, colleagues, and family who have supported me throughout the Ph.D program. It is not possible to mention everyone, but special thanks go to John Davis, a colleague, friend, and fellow traveler always willing to lend a keen intellect or a sympathetic ear as the situation required. Finally, completion of this dissertation would have been impossible without the support of my wife, Donna Seto Davis; last in mention, but first in my heart.
ABSTRACT

Speaking performance tests typically employ raters to produce scores; accordingly, variability in raters' scoring decisions has important consequences for test reliability and validity. One such source of variability is the rater's level of expertise in scoring. Therefore, it is important to understand how raters' performance is influenced by training and experience, as well as the features that distinguish more proficient raters from their less proficient counterparts. This dissertation examined the nature of rater expertise within a speaking test, and how training and increasing experience influenced raters’ scoring patterns, cognition, and behavior.

Experienced teachers of English (N=20) scored recorded examinee responses from the TOEFL iBT speaking test prior to training and in three sessions following training (100 responses for each session). For an additional 20 responses, raters verbally reported (via stimulated recall) what they were thinking as they listened to the examinee response and made a scoring decision, with the resulting data coded for language features mentioned. Scores were analyzed using many-facet Rasch analysis, with scoring phenomena including consistency, severity, and use of the rating scale compared across dates. Various aspects of raters' interaction with the scoring instrument were also recorded to determine if certain behaviors, such as the time taken to reach a scoring decision, were associated with the reliability and accuracy of scores.

Prior to training, rater severity and internal consistency (measured via Rasch analysis) were already of a standard typical for operational language performance tests, but training resulted in increased inter-rater correlation and agreement and improved correlation and agreement with established reference scores, although little change was
seen in rater severity. Additional experience gained after training appeared to have little effect on rater scoring patterns, although agreement with reference scores continued to increase. More proficient raters reviewed benchmark responses more often and took longer to make scoring decisions, suggesting that rater behavior while scoring may influence the accuracy and reliability of scores. On the other hand, no obvious relationship was seen between raters' comments and their scoring patterns, with considerable individual variation seen in the frequency with which raters mentioned various language features.
# TABLE OF CONTENTS

Acknowledgments ........................................................................................................ iv
Abstract .......................................................................................................................... vi
List of tables .................................................................................................................. xiii
List of figures .................................................................................................................. xiv

Chapter 1: Introduction ............................................................................................... 1
  1.1 Effects of training and experience on raters ................................................... 2
  1.2 Potential approaches to rater behavior from the domain of psychophysics 5
  1.3 Purpose of the study and research questions ................................................ 7
  1.4 Definitions ........................................................................................................... 8
    1.4.1 Rater expertise/proficiency ................................................................. 8
    1.4.2 Rater training ......................................................................................... 9
    1.4.3 Experience .............................................................................................. 9
    1.4.4 Rater cognition ...................................................................................... 9
    1.4.5 Rater behavior ....................................................................................... 9
  1.5 Organization of the dissertation ........................................................................ 10

Chapter 2: Variability in rater judgments and effects of rater background ............... 11
  2.1 Variability in scoring patterns ........................................................................ 11
    2.1.1 Consistency ........................................................................................... 12
    2.1.2 Severity ................................................................................................. 12
    2.1.3 Bias ....................................................................................................... 13
    2.1.4 Use of the rating scale ......................................................................... 13
    2.1.5 Halo ....................................................................................................... 14
    2.1.6 Sequence effect ................................................................................... 14
    2.1.7 Impact of rater variability on scores .................................................... 15
  2.2 Variability in rater decision-making ................................................................. 18
    2.2.1 Writing assessment ............................................................................... 18
    2.2.2 Speaking assessment ............................................................................ 23
    2.2.3 Summary ............................................................................................... 24
2.3 Effect of rater professional background on scores and decision making ..... 25
  2.3.1 Writing assessment ................................................................................ 25
  2.3.2 Speaking assessment ............................................................................. 28
  2.3.3 Summary ............................................................................................... 32
2.4 Effect of rater language background on scores and decision making ........ 33
  2.4.1 Writing assessment ................................................................................ 33
  2.4.2 Speaking assessment ............................................................................. 37
  2.4.3 Summary ............................................................................................... 40
Chapter 3: Expertise in scoring language performance ................................. 41
  3.1 Characteristics of expertise ....................................................................... 41
    3.1.1 Who is an expert? What is expertise? ................................................... 42
    3.1.2 Psychological characteristics of experts ................................................. 43
  3.2. Influence of experience on scoring patterns and decision making .......... 48
    3.2.1 Writing assessment ................................................................................ 49
    3.2.2 Speaking assessment ............................................................................. 53
    3.2.3 Summary of experience effects ............................................................. 55
  3.3 Influence of training on scoring patterns and decision making .............. 55
    3.3.1 Writing assessment ................................................................................ 57
    3.3.2 Speaking assessment ............................................................................. 63
    3.3.3 Summary of training effects .................................................................. 66
Chapter 4: The nature of scoring judgments in language performance tests ...... 68
  4.1 Models of the decision making process from writing assessment .......... 68
  4.2 Potential Approaches from Experimental Psychology ............................ 76
Chapter 5: Research questions and method ................................................... 85
  5.1 Research questions .................................................................................... 86
  5.2 Context of scoring ..................................................................................... 88
  5.3 Participants ............................................................................................... 90
  5.4 Materials .................................................................................................. 94
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.1 TOEFL Speaking Test responses</td>
<td>94</td>
</tr>
<tr>
<td>5.4.2 TOEFL Speaking Test scoring rubric</td>
<td>95</td>
</tr>
<tr>
<td>5.4.3 Reference scores</td>
<td>96</td>
</tr>
<tr>
<td>5.4.4 Online data collection setup</td>
<td>98</td>
</tr>
<tr>
<td>5.4.5 Language and Professional background questionnaires</td>
<td>99</td>
</tr>
<tr>
<td>5.4.6 Interactive data collection instruments</td>
<td>99</td>
</tr>
<tr>
<td>5.5 Research question 1: Effects of training and experience</td>
<td>101</td>
</tr>
<tr>
<td>5.5.1 Overview</td>
<td>101</td>
</tr>
<tr>
<td>5.5.2 Instruments and procedures</td>
<td>103</td>
</tr>
<tr>
<td>5.5.2.1 Orientation</td>
<td>103</td>
</tr>
<tr>
<td>5.5.2.2 Scoring sessions</td>
<td>104</td>
</tr>
<tr>
<td>5.5.2.3 Rater training</td>
<td>109</td>
</tr>
<tr>
<td>5.5.6 Analyses</td>
<td>111</td>
</tr>
<tr>
<td>5.5.6.1 Severity and consistency of scores</td>
<td>111</td>
</tr>
<tr>
<td>5.5.6.2 Accuracy of scores</td>
<td>115</td>
</tr>
<tr>
<td>5.5.6.3 Scoring behavior</td>
<td>115</td>
</tr>
<tr>
<td>5.6 Research question 2: Differences between more-proficient and less-</td>
<td>116</td>
</tr>
<tr>
<td>proficient raters</td>
<td>116</td>
</tr>
<tr>
<td>5.6.1 Overview</td>
<td>119</td>
</tr>
<tr>
<td>5.6.2 Instruments and procedures</td>
<td>119</td>
</tr>
<tr>
<td>5.6.2.1 Rater scoring behavior</td>
<td>119</td>
</tr>
<tr>
<td>5.6.2.2 Rater cognition</td>
<td>119</td>
</tr>
<tr>
<td>5.6.3 Analyses</td>
<td>121</td>
</tr>
<tr>
<td>5.6.3.1 Use of the rating scale</td>
<td>121</td>
</tr>
<tr>
<td>5.6.3.2 Scoring behavior</td>
<td>122</td>
</tr>
<tr>
<td>5.6.3.3 Rater cognition</td>
<td>122</td>
</tr>
<tr>
<td>5.7 Research question 3: Evaluation of the relative view of magnitude</td>
<td>124</td>
</tr>
<tr>
<td>5.7.1 Overview</td>
<td>124</td>
</tr>
<tr>
<td>5.7.2 Instruments and procedures</td>
<td>125</td>
</tr>
</tbody>
</table>
5.7.2.1 Sequence effect ................................................................. 125
5.7.2.2 Comparison of pairwise and isolated judgment ............... 125
5.7.2.3 Isolated judgment ............................................................. 128
5.7.2.4 Pairwise judgment ............................................................ 129
5.7.3 Analysis .............................................................................. 130
5.7.3.1 Sequence effect ............................................................. 130
5.7.3.2 Comparison of pairwise and isolated judgment ............... 132

Chapter 6: Results ........................................................................ 134
6.1 Effects of training and experience on scoring patterns and behavior .... 135
   6.1.1 Many-facet Rasch analysis .................................................. 135
   6.1.2 Between-rater consistency in scoring ...................................... 139
   6.1.3 Within-rater consistency in scoring ....................................... 143
   6.1.4 Accuracy of scores ............................................................. 148
   6.1.5 Rater behavior while scoring ............................................... 152
6.2 Scoring Behavior and Cognition of More-Proficient and Less-Proficient Raters .......................................................... 156
   6.2.1 Scoring patterns ............................................................... 156
   6.2.2 Behavior while scoring ........................................................ 160
   6.2.3 Cognition while scoring ..................................................... 168
6.3 Usefulness of a relative view of magnitude judgment ............... 174
   6.3.1 Sequence effect ............................................................... 174
   6.3.2 Comparison of pairwise and isolated judgments ................ 177

Chapter 7: Summary and Discussion .............................................. 183
7.1 Effects of training and experience .......................................... 183
   7.1.1 Summary of results .......................................................... 183
   7.1.2 Discussion ....................................................................... 186
7.2 Differences between more- and less-proficient raters ............... 193
   7.2.1 Summary of results .......................................................... 193
   7.2.2 Discussion ....................................................................... 195
7.3 Relevance of the relative judgment view to judgments of language ability...

7.3.1 Summary of results ........................................................................................................... 199

7.3.2 Discussion ..................................................................................................................... 200

Chapter 8: Conclusion ............................................................................................................ 205

8.1 Limitations of the study .................................................................................................. 205

8.2 Theoretical implications of the study ............................................................................. 208

8.3 Practical implications of the study .................................................................................. 212

8.4. Possibilities for future research ...................................................................................... 214

8.5 Conclusion ....................................................................................................................... 218

Appendix A: Description of the tasks used in the TOEFL iBT Speaking Test ........ 219

Appendix B: Scoring rubric used in the study ................................................................. 220

Appendix C: Academic and professional background of TOEFL scoring leaders... 221

Appendix D: FACETS results for ETS scoring leaders ....................................................... 224

Appendix E: Professional background ETS questionnaire .............................................. 224

Appendix F: Modified LEAP-Q Questionnaire ................................................................. 226

Appendix G: Scoring instrument ............................................................................................. 228

Appendix H: Scoring session wrap-up questionnaire ......................................................... 230

Appendix I: Common person linking plots ........................................................................ 231

Appendix J: Instrument for collecting verbal report data .................................................. 233

Appendix K: Categories used to code verbal recall data .................................................... 234

Appendix L: Questions asked after isolated and pairwise scoring sessions ................. 235

Appendix M: Average raw scores responses repeated across scoring sessions ......... 236

Appendix N: Time required to make scoring decisions ..................................................... 237

Appendix O: Scoring patterns of more-proficient, less-proficient, and improving raters ............................................................................................................................ 238

Appendix P: Raw score scale boundaries ........................................................................... 240

Appendix Q: Raters comments during stimulated recall .................................................. 241

References ............................................................................................................................. 242
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. The stages of the scoring process (from Lumley, 2002)</td>
<td>74</td>
</tr>
<tr>
<td>5.1. Academic, professional, and testing backgrounds of participants</td>
<td>93</td>
</tr>
<tr>
<td>5.2. Score distributions of speaking responses in the TOEFL Public Use Dataset</td>
<td>94</td>
</tr>
<tr>
<td>5.3. Selection criteria for raters used to answer research question 2</td>
<td>117</td>
</tr>
<tr>
<td>5.4. Rater background information</td>
<td>118</td>
</tr>
<tr>
<td>5.5. Descriptive statistics for responses used in relative judgment experiments</td>
<td>128</td>
</tr>
<tr>
<td>6.1. Rater severity measures from multifaceted Rasch analysis</td>
<td>138</td>
</tr>
<tr>
<td>6.2. Mean pairwise inter-rater correlations within scoring sessions</td>
<td>142</td>
</tr>
<tr>
<td>6.3. Effect sizes for changes across sessions in inter-rater correlations</td>
<td>142</td>
</tr>
<tr>
<td>6.4. Agreement indices within scoring sessions</td>
<td>143</td>
</tr>
<tr>
<td>6.5. Infit mean square values from multifaceted Rasch analysis</td>
<td>145</td>
</tr>
<tr>
<td>6.6. Frequency of scores showing rater-by-examinee bias</td>
<td>148</td>
</tr>
<tr>
<td>6.7. Rater accuracy in terms of Pearson correlation with reference scores</td>
<td>150</td>
</tr>
<tr>
<td>6.8. Rater accuracy in terms of agreement (Cohen's Kappa) with reference scores</td>
<td>151</td>
</tr>
<tr>
<td>6.9. Effect sizes for changes across sessions in agreement (Cohen's kappa) with reference scores</td>
<td>152</td>
</tr>
<tr>
<td>6.10. The number of times the exemplars were checked while scoring</td>
<td>154</td>
</tr>
<tr>
<td>6.11. Average Pearson product-moment correlations for successive pairs of scores</td>
<td>176</td>
</tr>
<tr>
<td>6.12. Responses to questionnaires administered immediately after the pairwise and isolated judgment data collection sessions</td>
<td>182</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Model of the holistic marking of compositions (from Milanovic, Saville, &amp; Shen, 1996)</td>
<td>72</td>
</tr>
<tr>
<td>4.2 The stages of the scoring process (from Lumley, 2002)</td>
<td>74</td>
</tr>
<tr>
<td>4.3 Debits made to individual accounts at a major UK bank over one year (from Stewart, Brown &amp; Chater, 2006)</td>
<td>79</td>
</tr>
<tr>
<td>4.4 Distribution and cumulative ranking of scores for the TOEFL iBT speaking test</td>
<td>80</td>
</tr>
<tr>
<td>5.1 Sampling design for the study</td>
<td>102</td>
</tr>
<tr>
<td>5.2 The first page of the scoring instrument showing the review of exemplars in progress</td>
<td>106</td>
</tr>
<tr>
<td>5.3 The first page of the scoring instrument showing scoring in progress</td>
<td>107</td>
</tr>
<tr>
<td>5.4 Summary of the rater training</td>
<td>111</td>
</tr>
<tr>
<td>6.1 Measurement rulers for many-facet Rasch analysis of scores</td>
<td>137</td>
</tr>
<tr>
<td>6.2 Rater severity (in logits) across scoring sessions</td>
<td>140</td>
</tr>
<tr>
<td>6.3 Rater scoring consistency in terms of infit mean square</td>
<td>146</td>
</tr>
<tr>
<td>6.4 Rater accuracy in terms of Pearson correlations with reference scores</td>
<td>149</td>
</tr>
<tr>
<td>6.5 The number of times the exemplars were reviewed during a scoring session</td>
<td>154</td>
</tr>
<tr>
<td>6.6 The average time required to make a scoring decision</td>
<td>156</td>
</tr>
<tr>
<td>6.7 The percentage of within-rater exact agreement across scoring sessions</td>
<td>158</td>
</tr>
<tr>
<td>6.8 Scale use of more-proficient, less-proficient, and improving raters</td>
<td>159</td>
</tr>
<tr>
<td>6.9 Use of exemplars while scoring for proficient, non-proficient, and developing raters</td>
<td>162</td>
</tr>
<tr>
<td>6.10 The average time required to make a scoring decision for proficient, non-proficient, and developing raters</td>
<td>164</td>
</tr>
<tr>
<td>6.11 Distribution of decision times for more-proficient, less-proficient, and improving raters</td>
<td>165</td>
</tr>
<tr>
<td>6.12 The average number of breaks (5 minutes or longer) taken while scoring by proficient, non-proficient, and developing raters</td>
<td>167</td>
</tr>
</tbody>
</table>
6.13. The frequency of raters' comments on various aspects of language performance, collected during stimulated recall .................................................. 170

6.14. The frequency of raters' comments on various aspects of language performance, collected during stimulated recall .................................................. 171

6.15. The frequency of raters' comments regarding scoring or the scoring process, collected during stimulated recall .................................................. 172

6.16. The average ratio of rater comments per utterance ............................................. 172

6.17. The proportion of correct discriminations in pairwise and isolated judgment conditions ........................................................................................................ 178

6.18. D-prime values for discrimination of responses with reference scores of three vs. four and four vs. five ................................................................. 178

6.19. Proportion of hits and false alarms for discrimination of responses with reference scores of three vs. four ................................................................. 180

6.20. Proportion of hits and false alarms for discrimination of responses with reference scores of four vs. five ................................................................. 180
CHAPTER 1

Introduction

In tests of speaking performance it is the rater who performs the act of measurement; that is, the assignment of a number to an observation through the application of a rule (Stevens, 1946). How and why raters make scoring decisions, therefore, has important consequences for test reliability and validity, and indeed, the nature of rater-associated variability in scores has been a persistent source of concern for language performance tests overall (Norris, Brown, Hudson, & Yoshioka, 1998) and speaking tests in particular (Fulcher, 2003). Understanding the nature of such variability is no simple matter, however, given the complexity of the scoring task and the many factors that may influence raters' judgments. The apparent simplicity of a score belies the fact that language performance tests often measure complex phenomena and that raters must make judgments on the basis of brief scoring rubrics where the fit between performance and score is not always obvious (Cumming, Kantor, & Powers, 2001; Lumley, 2005). Yet, despite the challenges raters face, a variety of studies have observed that raters are indeed capable of producing consistent scores (e.g., Huot, 1990; McNamara, 1996).

How raters achieve consistent scores is of considerable practical and theoretical interest. In practical terms, understanding rater characteristics that contribute to consistency in scoring has obvious application for ensuring the reliability of test scores. Moreover, a better understanding of how scoring expertise develops has equally obvious application to the design of rater training procedures. Beyond issues of reliability, knowledge of the decision-making processes and the language features raters actually use
to make scoring decisions should lead to a better understanding of the construct(s) that the scores actually represent. In theoretical terms, models of rater judgment allow for clearer thinking regarding how the various aspects of the scoring context contribute to more (or less) consistent and valid scores, and may suggest directions for research to clarify how different scoring procedures or scoring aids (such as rubrics) actually promote desirable scoring patterns. In addition, a theoretical understanding of how scores are produced may give insight into the ways in which human scoring compares to machine scoring. Machine-generated speaking tests scores do not necessarily capture the same construct that human raters do (Bridgeman, Powers, Stone, & Mollaun, 2012), and an understanding of human scoring is needed if claims are to be made that machine-generated scores are equivalent to human scores.

1.1 Effects of Training and Experience on Raters

Despite the potential value of understanding rater expertise, work in this area represents a relatively modest portion of the research on raters and rating, and the majority of that work has been conducted largely in writing assessment contexts. Studies examining the effect of experience have suggested that experienced raters are more internally consistent in their scores, but do not necessarily show close agreement with each other (Barkaoui, 2010a; Weigle, 1994). Rather, differences in severity among raters are typical in language performance tests (Eckes, 2011; McNamara, 1996) and longitudinal studies of raters suggest that some degree of difference between raters may persist over periods as long as one to two years (Lim, 2011; Lumley, 2005). Systematic differences in severity between novice and expert raters have also been observed,
although findings conflict regarding the nature of this difference. Weigle (1998) found that novice essay raters were more severe, while Barkaoui (2010a) found novices to be more lenient. Lim (2011) reported that only a portion of novice raters were excessively severe or lenient, but those raters whose scores were most different from the group tended to adjust their scoring over time to be similar, but not necessarily identical, to more experienced raters. Comparatively few studies of rater expertise have been conducted in speaking test contexts, though in a recent study, Kim (2011) found that experienced raters were slightly more severe, but that there was no difference between experienced and novice raters in within-rater scoring consistency.

Beyond analyses of scores, qualitative differences have also been observed in how individuals with more or less professional experience approached the scoring task. Again focusing on writing assessment, Cumming (1990) reported differences in scoring criteria used by novice and experienced teachers, with experienced teachers making more mention of specific linguistic features. Experienced teachers were also more directed in their evaluation behavior, as evidenced by such actions as consciously focusing their attention on specific scoring criteria. Similarly, Wolfe, Kao, and Ranney (1998) found that expert raters tended to mention specific rather than general language features in their comments; they also made fewer comments regarding items not mentioned in the scoring rubric and were more likely to wait until after reading to make an evaluation of the text. A similar tendency to read first and then evaluate has also been reported among experienced raters scoring L1 English compositions (Huot, 1993). In a more recent study, Barkaoui (2010b) found that essay raters with less experience referred to the scoring rubric more often and made more comments to justify their scores. Kim (2011) also
observed that inexperienced raters in a speaking test relied more on the scoring rubric, but they also occasionally misinterpreted the rating scale descriptors and sometimes mentioned construct-irrelevant features of performance.

Other studies have reported various impacts of training on raters' scores and scoring processes. In an experimental pre/post study, Weigle (1994, 1998) found that training resulted in improved within-rater scoring consistency but did not eliminate scoring differences between raters, much like the results seen for the comparison of experienced/novice raters made in the same study. Think-aloud protocols from novice raters also suggested that training helped them to better understand the rating criteria and increased their awareness of, and concern for, the standards used by other raters. In a similar vein, Barnwell (1989) suggested that trained raters are socialized into an understanding of the rating scale. In a comparison of trained ACTFL Spanish Speaking Test raters versus untrained native speakers of Spanish, Barnwell found that the ranking of examinees was similar between the groups but the actual scores were different, suggesting that the ACTFL rater training had adjusted raters' perceptions. In her research on an ESL speaking test, Wigglesworth (1993) reported that targeted feedback and retraining also reduced the prevalence of individual rater bias towards different items and language features. On the other hand, more recent studies from writing assessment have seen little effect of feedback on scoring patterns of experienced raters (Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, 2011), suggesting that training may have little impact once raters have become familiarized with the testing context.
To summarize, some general observations regarding existing research on rater expertise are:

1. The majority of studies have been done in writing assessment contexts.
2. The majority of studies have examined differences between experienced and inexperienced raters; studies specifically addressing the effects of rater training are relatively few.
3. The majority of studies have focused on analyses of scores. A few studies have examined rater decision-making of experienced and novice raters using verbal report techniques (e.g., Wolfe, Kao, & Ranney, 1998; Barkaoui, 2010a,b; Kim, 2011), and one study (Weigle, 1998) has examined the effect of training.
4. While some findings appear to be relatively consistent, such as an increase in within-rater consistency with training and/or experience, other findings are less consistent, such as the effect of experience and training on rater severity, or the degree to which training influences scores.

In light of these issues, this dissertation aims to extend our understanding of rater expertise in a speaking test context, with a focus on both the effects of training as well as experience. In addition, the study examines effects of training and experience on both scores and decision making processes.

1.2 Potential Approaches to Rater Behavior from the Domain of Psychophysics

A variety of models of rater decision-making have been proposed (e.g., Freedman & Calfee, 1983; Lumley, 2005; Milanovic, Saville, & Shen, 1996) and are useful for
thinking about rater expertise, in the sense that they provide a description of the scoring
task that raters must master. Conversely, such models do not necessarily specify what it is
that changes as raters become better at scoring, so their usefulness for understanding the
nature of expertise is limited. Models of magnitude judgment from the domain of
psychophysics have the potential to provide a more specific account of the cognitive
basis of expertise, and in particular Laming’s (1997, 2004) argument that magnitude
judgments are fundamentally relative in nature may be useful in this regard. This relative
view of human judgment is based on the claim that, at a basic level, all magnitude
judgments are made through comparisons of the item being judged with other similar
items in the immediate environment, or that have been encountered in the past. This
assumption stands in contrast to the commonly-used definition of measurement given at
the beginning of this introduction, which implies that measurement is a process of
applying some sort of standard “measuring stick” to an item, resulting in an absolute
determination of magnitude. In terms of rater expertise, the relative judgment view would
posit that, as raters encounter more examinee performances, they develop a greater
collection of examples that can be used for comparison when making judgments,
allowing more precise comparisons to be made and ultimately resulting in more
consistent scoring (Laming, 2003).

Although Laming’s ideas were originally developed to explain magnitude
judgments of physical phenomena (i.e., psychophysics), the notion that human judgment
is essentially relative has more recently been used to explain economic decision making
(e.g., Stewart, Chater, & Brown, 2006), and Laming (2003) has argued that judgments
made in educational testing contexts are relative as well. The relative judgment view has
been supported with a variety of empirical findings, including the observation that
differences in magnitude are easier to distinguish when physical stimuli are presented
side-by-side (which allows for direct comparison), as well as the observation that
magnitude judgments of successive stimuli tend to be related; that is, the current stimulus
is compared to the previously judged stimulus (Laming, 1997). A similar such sequence
effect for scores has been observed in studies of English L1 writing assessment (Daly &
Dickson-Markman, 1982; Hughes, Keeling & Tuck, 1980), as well as in think-aloud
studies of English L2 writing tests (Cumming et al., 2002; Milanovic et al., 1996; Sakyi;
2000). Nonetheless, support for the relative judgment view comes mainly from the
domain of psychophysics and its relevance to magnitude judgments of complex social
phenomena, such as judgments of language ability, remains unclear.

1.3 Purpose of the Study and Research Questions

The purpose of this dissertation is to extend our understanding of rater expertise
within a speaking test context. The primary focus includes an investigation of the ways in
which rater scoring patterns and behavior change with experience, as well as the
characteristics of scoring behavior and cognition associated with raters showing more or
less desirable patterns in scoring. In addition, changes in cognition and behavior resulting
from training will be examined, and an initial effort will be made to examine the claim
that judgments of language ability are relative in nature. The study focuses on the
following research questions:
1. What effects do training and experience have on scoring patterns and scoring behavior of raters?

2. What features of scoring behavior and cognition distinguish more-proficient and less-proficient raters?

3. Can rater decision making and expertise be understood in terms of a relative view of magnitude judgment?

1.4 Definitions

Terminological usages in the published literature on rater decision making are not always consistent. For example, activities intended to familiarize raters with the rating scale and/or ensure consistent scoring have been referred to as both norming and training (Kim, 2011). So, for the sake of clarity, definitions of a few key terms, as used in this dissertation, are given below.

1.4.1 Rater expertise/proficiency. The definition of rater expertise (or proficiency) used here starts with that of Lim (2011) who defined expertise in terms of desirable scoring patterns. That is, raters who produce good scores are said to have expertise. Going further, the term expertise is used to refer to the knowledge, skills, and behaviors that enable a rater to produce consistent, valid, and accurate scores within a particular scoring context. Consistency in scoring refers to the extent to which a rater's scores agree with scores obtained from (a) other raters, (b) the same rater on other occasions, or (c) statistical models of the test context. Validity in scoring is defined as the extent to which scores incorporate the language features targeted by the test, and exclude construct-irrelevant aspects of an examinee's performance. Accuracy in scoring refers to
the extent to which a rater's scores align with the intended scale of measurement.
Accuracy differs from scoring consistency in that between-rater or within-rater agreement
does not necessarily prove agreement with the scale as established by the test developers
or through previous use (Eckes, 2011, p. 29).

1.4.2 Rater training. Rater training (or simply training) refers to activities
undertaken by raters that are intended to improve or ensure the consistency, validity, or
accuracy of scoring. Training refers to activities occurring outside of operational scoring,
and therefore excludes the use of scoring aids to make specific scoring judgments.

1.4.3 Experience. As used in this dissertation, experience refers to the extent to
which the rater has previously participated in scoring activities within the specific testing
context (Lim, 2011). It does not refer to aspects of the rater's background or the rater's
previous experiences outside of the testing context (e.g., Cumming, 1990).

1.4.4 Rater cognition. Rater cognition refers to the mental processes occurring
during scoring, at either a conscious or unconscious level. This includes both the features
of examinee performance attended to while scoring, as well as the mental actions taken to
reach a scoring decision. However, the data reported in this dissertation primary address
the language features consciously attended to while scoring.

1.4.5 Rater behavior. Rater behavior refers to the observable behaviors produced
by raters while scoring, such as interactions with the scoring instrument (e.g., button
clicks) or the time taken to reach a scoring decision.
1.5 Organization of the Dissertation

This dissertation is organized as follows. In chapter 2 the nature of variability in rater scoring patterns and decision making processes is discussed, along with a review of the ways in which the professional and language backgrounds of raters may influence such variability. Chapter 3 explores the nature of expertise, including the general features of expertise as described in the psychological literature as well as effects of experience and training on scoring patterns and decision making behavior of raters in language tests. Chapter 4 explores models of the decision making process within language performance tests, from the view that an understanding of the scoring task is necessary as a basis for understanding what it is that expert raters do. In chapter 5 the study is described, starting with the research questions that guided the study, then followed by a detailed description of the study context, participants, materials, and analyses. Chapter 6 reports the results of the analyses, organized by research question, while chapter 7 synthesizes the findings reported in chapter 6 in light of previous research and suggests possible explanations for trends seen in the data. Finally, chapter 8 summarizes the theoretical and practical insights gained, the limitations of the study, and suggests avenues for future research.
CHAPTER 2

Variability in Rater Judgments and Effects of Rater Background

In language performance tests, raters are a potential source of unwanted variability in scores as well as the final arbiters of what aspects of human performance the scores actually represent. Therefore, variability in raters’ scoring patterns and the factors that contribute to such variability have received considerable attention within the literature on language performance assessment. This chapter reviews issues surrounding the variability of rater judgments in language tests and how aspects of raters’ backgrounds may influence such variability. Section 2.1 of the chapter discusses the types of variability in scoring patterns that have been reported for language tests and how this might impact test results. Section 2.2 follows with a summary of variability in scoring criteria and decision-making strategies used by raters. Sections 2.3 and 2.4 then take up the issue of how rater background may influence scores and decision making, with section 2.3 examining the influence of professional background, and section 2.4 discussing the effects of language background.

2.1 Variability in Scoring Patterns

Human judgments are seldom perfectly accurate or consistent, nor are two people ever exactly the same in their perceptions; therefore some degree of variability in rater judgments is an unavoidable fact of human scoring. Eckes (2011) defines rater variability as the variability in scores that can be attributed to raters rather than examinees, and since this variability is unrelated to examinee language ability it is therefore considered to be a source of construct-irrelevant variance (Messick, 1989).
Rater variability comes in a number of different forms, including both random variation as well as several different types of systematic scoring patterns (Eckes, 2011; Englehard, 1994; Linacre, 2007a; Knoch, Read, & von Randow, 2007; Leckie & Baird, 2011; McNamara, 1996; Meadows & Billington, 2005; Saal, Downey, & Lahey, 1980). Types of rater variation identified in the language testing literature are summarized below.

**2.1.1 Consistency.** Firstly, a rater's scores may simply vary randomly (McNamara, 1996). This variation may be apparent in both erratic patterns of scores compared to scores awarded by other raters, or random instability in scores awarded from one occasion to the next. Rater consistency is commonly examined using reliability measures such as inter-rater correlations, or model fit indices when many-facet Rasch measurement (MFRM) is used (e.g., McNamara, 1996; Myford & Wolfe, 2000). This type of variation is arguably the most problematic in that there is little that can be done after the fact to correct for inconsistency and make the scores usable (McNamara, 1996).

**2.1.2 Severity.** Different raters may be more or less strict in their judgments. For example, one rater may award consistently higher scores while another awards lower scores for the same performances. In addition, the severity of a rater's own judgments may vary within scoring sessions (rater drift; Quellmalz, 1980) or across scoring sessions (instability; Linacre, 2007a). Differences in severity between raters appear to be exceedingly common (McNamara, 1996; Eckes, 2011) and also durable over time (Lim, 2009; Lumley, 2005; Lumley & McNamara, 1995; Lunz & Stahl, 1990). The prevalence of severity differences between raters has also been a justification for the use of many-facet Rasch measurement, which has the ability to model such differences and take them into account when estimating examinee ability (Linacre, 2007a).
2.1.3 Bias. *Bias* refers to the tendency to award unusually high or low scores when scoring a particular test task, language performance feature, examinee or group of examinees, or other aspect of the testing situation (McNamara, 1996). For example, a rater may tend to give higher scores when reading essays on a particular topic. Rater bias has also received considerable attention in language tests; studies have addressed such issues as bias related to certain types of examinees (Eckes, 2005; Kondo-Brown, 2002; Winke, Gass, & Myford, 2011), language features (Schaefer, 2008), scoring sessions (Lumley & McNamara, 1995), and interlocutor skill (McNamara & Lumley, 1997). In addition, the effects of feedback and training on rater bias have been of interest as well (Fahim & Bijani, 2011; Knoch, 2011; Knoch, Read, & von Randow, 2007; O’Sullivan, & Rignall, 2007; Wigglesworth, 1993).

2.1.4 Use of the rating scale. Raters may vary in their understanding of the underlying structure of the rating scale (Eckes, 2011; McNamara, 1996). Raters may tend to overuse certain score levels and underuse others, or the intervals of the scale may be uneven (McNamara, 1996). That is, increasing scores may not reflect evenly spaced increases in ability (like the rungs of a ladder) but may rather be irregular, where, for example, an increase in score from 2 to 3 represents a small increase in language ability while a change in score from 3 to 4 represents a much larger increase in ability. Such irregularity in the way raters apply the rating scale can be random or systematic. A common systematic distortion of the rating scale is *restriction of range*, where scores cluster in one part of the scale (Eckes, 2011; Myford & Wolfe, 2003). One form of this phenomenon is *central tendency* (Eckes, 2011; Linacre, 2007a), which occurs when the rater overuses the middle categories of the rating scale and under-uses the extremes,
systematically giving lower-ability examinees scores that are too high, and higher-ability examinees scores that are too low.

2.1.5 Halo. Halo effect is possible when different aspects of performance are judged separately (i.e., analytic scoring) and refers to the tendency to award the same score across distinct performance categories (Fulcher, 2003). Possible causes of halo effect include (a) allowing a general impression to dominate scores (i.e., judging the performance holistically rather than analytically), (b) failure to distinguish different elements of performance, or (c) one aspect of performance dominates the rater's perception of other scoring criteria (Eckes, 2011; Engelhard, 1994).

2.1.6 Sequence effect. Sequence effect occurs when the score awarded to one test taker influences the score awarded to a subsequent test taker (Attali, 2011). This may be considered an instance of the broader psychological phenomenon of anchor and adjustment bias (Plous, 1993) where a value experienced earlier serves as an anchor or point of reference for adjusting a later estimate of magnitude. For example, an essay of average quality may be judged more harshly when preceded by an excellent paper because the average essay suffers in comparison. Conversely, when preceded by a poor quality essay, an average essay may seem stronger in comparison and receive a higher score. This type of comparison effect is referred to as contrast and has been observed in several studies of essay marking (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes & Keeling, 1984; Hughes, Keeling & Tuck, 1980; Spear, 1997). The opposite pattern, where latter values are adjusted to be more similar to the anchor value, is termed assimilation and has also been observed in essay scoring (Attali, 2011).
2.1.7 Impact of rater variability on scores. The impact of rater variability on scores seems to vary widely, but can be considerable. An example of this latter situation is the early study by Diederich, French, and Carlton (1961) who analyzed ratings produced by 53 readers scoring short essays written by U.S. college freshman. Raters were selected from a variety of professional backgrounds, including English, social science, natural science, and law. No scoring rubric was provided; rather, raters were asked to award scores based on their "own judgment as to what constitutes 'writing ability'" (p. 11). A total of 94% of the essays received scores in at least seven of the nine possible scoring categories, and all papers were awarded scores in at least five score categories, indicating considerable disagreement among raters. But, Diederich, French, and Carlton's study is also an extreme case, given the variety of rater backgrounds and the absence of guidance for scoring, and was in fact designed to study the range of variation in rater judgments.

On the other hand, when scoring criteria, training, and other scoring aids are provided, a variety of studies have observed that relatively high inter-rater reliability values may be obtained, with correlation coefficients often equaling .80 or more (Fulcher, 2003). But, reliability measures only describe the extent to which raters agree on the relative ranking of examinees (Hatch & Lazaraton, 1991) and do not capture differences in agreement (i.e., severity). As mentioned earlier, differences in severity between raters appear to be both common and durable over time, and so reliability measures provide an incomplete picture of rater variability. One approach that provides a more complete view of rater variability in scoring is many-facet Rasch measurement, which allows estimation of the magnitude of rater variation in severity, thus giving some indication of the degree
to which an “easy” or “hard” rater might influence an examinee’s score. In studies of language performance assessment, rater severity values commonly vary over a range of approximately two logits (e.g., Bonk & Ockey; 2003; Cho; 1999; Cogndon & McQueen, 2000; Lim, 2011; Lumley, 2005; Lumley & McNamara, 1995; McNamara, 1996). Moreover, in large groups of raters variation in severity can be much greater. For example, Myford, Marr, and Linacre (1996) analyzed ratings from 142 operational raters of the TOEFL Test of Written English (TWE) and found rater severity extended over a range of nearly 7 logits. In a similar study of the TOEFL Test of Spoken English (TSE), Myford and Wolfe (2000) observed rater severity values extending across a range of 3.5 logits in one administration (66 raters) and 3.2 logits in another administration (74 raters).

What does such variability in severity, measured in logits, actually mean for examinees? Logit values (short for log odds) are a measure of probability, in this case the log of the odds that an examinee will receive a given score (Bond & Fox, 2007). A difference in rater severity of 2 logits can be illustrated with an example. If a particular examinee had a 50% chance of getting a score of “3” from a given rater, they would have only a 12% chance of getting a “3” from a rater who is stricter by an amount of 2 logits. If the second rater is stricter by 3 logits, then the probability drops to 5%, and for 7 logits the probability is below 1% (Linacre, 2010; McNamara, 1996).

Another way to get a sense of the impact of rater variability on scores can be gained from studies which have used generalizeability theory (G-Theory; Shavelson & Webb, 1991) to directly quantify the contribution of rater variability to the overall variability in scores. Under a G-theory approach, the goal is to quantify the degree to which various aspects of the testing situation influence scores; specifically, techniques
from ANOVA are used to quantify the contribution of various factors to the overall score variance. Ideally, most of the variance in scores is attributable to differences between examinees. Variance contributed by raters can be viewed as construct-irrelevant, and includes both rater variability included as a main effect in the ANOVA model (i.e., generalized inter- or intra-rater variability) as well as rater variability modeled as interaction with examinees, prompts, or other features of the test (i.e., rater bias). Such G-theory studies of language performance tests have produced a range of results. Several studies have reported rater main effects as being less than 10% of total score variance (e.g., Brown & Ahn, 2011; Lynch & McNamara, 1998; Xi, 2007), and some have reported values below 1% (Bachman, Lynch, & Mason, 1995; Kim, 2009a; Lee, 2006; Lee & Kantor, 2005). In contrast, other studies have reported rater main effects on the order of 40% of total score variance (Yoshida, 2004) and in one case as much as 61% (Yamanaka, 2005). Variability associated with the interaction between raters and other aspects of the test context is usually rather modest and typically accounts for less than 10% of the variation in scores (e.g., Brown & Ahn, 2011; Kim, 2009a; Xi, 2007) although values as high as 16-17% have also been reported (Brown & Bailey, 1984; Yamanaka, 2005).

Although G-theory analyses tend to suggest less of a rater effect on scores than many-facet Rasch measurement might indicate, in fact the two approaches are complementary, with G-theory providing a broader view of the overall influence of raters on scores, and Rasch analysis providing a means of examining the influence of specific raters (Bachman, Lynch, & Mason, 1995; Lynch & McNamara, 1998). In any case, both
approaches suggest that the potential exists for rater variability to have a substantial impact on scores.

2.2 Variability in Rater Decision-Making

Rater variability in scores must ultimately derive from the decision making that produced the scores, and such decision making has been observed to vary in different ways, particularly in terms of the criteria used to evaluate examinee performances and the processes by which these criteria are applied. A brief review of variability in scoring criteria and decision making processes follows. Findings from writing assessment are discussed first followed by studies of speaking assessment. The two modalities are treated separately here and in subsequent sections because they differ in terms of both the language features to be judged as well as the ways in which raters interact with examinee responses; it therefore seems possible that perception and judgment may be qualitatively different in the scoring of writing versus speaking.

2.2.1 Writing assessment. A basic finding from studies of rater cognition is that use of scoring criteria differs between raters. In the study of inter-rater variability by Diederich, French, and Carlton (1961) mentioned earlier, factor analysis of scores suggested that raters could be classified into five groups; rater comments indicated that each group focused on a different set of language features while scoring. For example, one group focused on “ideas” and another focused on “mechanics”. However, as mentioned earlier, Diederich et al. did not provide raters with a scoring rubric or training, so variability in scoring criteria is to be expected. Nonetheless, a more recent study by Eckes (2008) also identified several different types of raters on the basis of differences in scoring criteria, this time within an operational test of German as foreign language
where raters were trained and experienced in applying the scoring rubric to writing samples.

A number of studies within English L2 writing assessment have also observed diversity in the features of performance attended to by raters. Vaughan (1991) used think-aloud protocols to examine the perceptions of nine experienced raters during holistic scoring of university English essays written by both ESL and English L1 students. Vaughan noted 14 categories of comments, with the frequency of different types of comments varying between raters. For example, one rater seemed to focus on introductions, another focused on mechanics, and a third commented most often on content issues. Like Vaughan, Sakyi (2000) also found differences in preferences between six experienced raters scoring ESL essays, this time written by first year university students in architecture, engineering, and pharmacy. Think-aloud data showed that one rater appeared to focus almost exclusively on errors, another focused on ideas, a third focused on personal reactions to the text, and the remaining three attempted to closely follow the scoring guide. Similarly, Milanovic, Saville, and Shen (1996) distinguished 11 different criteria mentioned by raters who were holistically scoring essays from the First Certificate in English (FCE) and Certificate of Proficiency in English (CPE) examinations, with different raters weighting various criteria in different ways.

Holistic scoring leaves considerable latitude in the hands of the rater regarding the precise features of performance used to make a scoring decision, but even when analytic scoring is used, differences remain in rater interpretations. Smith (2000) found that although scores for specific aspects of performance were fairly consistent among six
raters, the actual criteria used to produce the scores were less consistent. Raters scored three adult ESL essays on six specific criteria (such as “can use past tenses and other past markers”, p. 167); agreement was high as to whether a particular essay demonstrated competency for a given criterion, but think-aloud reports suggested that raters varied in the way they interpreted and/or applied the criteria. In addition, in one case general agreement that the essay was a “fail” was accompanied by disagreement regarding which of the criteria had actually gone unmet.

In addition to differences in scoring criteria, variability in the scoring processes of raters has also been reported. Using think-aloud data, Cumming (1990) documented 28 decision-making behaviors among readers of a university ESL writing placement test. These included seven “interpretation strategies” which were used to read and understand the essays and 21 “judgment strategies” which were involved in the appraisal of quality. Cumming (1990) observed that the frequency with which specific strategies were used varied dramatically from individual to individual; he suggested this might be the result of individual differences in the large number of specific decisions that must be made when generating scores (which in this case included scores for “language use,” “rhetorical organization,” and “substantive content”). On the other hand, only two of the strategies (“classifying errors” and “editing phrases”) accounted for the majority of strategies raters reported using.

Cumming’s classification scheme was subsequently used to analyze rater think-alouds collected in several studies of the TOEFL Test of Written English (Connor & Carrell, 1993; Cumming, Kantor, & Power, 2002; Erdosy, 2004). In contrast to earlier findings, Cumming et al. (2002) observed that no particular strategy dominated the data;
“read or re-read the composition” was the most frequent action, appearing in 15-20% of comments, and all other behaviors made up 10% or less of the dataset, suggesting considerable variability in rater decision making. Nonetheless, Cumming et al. also noted what they termed a prototypical scoring sequence, consisting of three steps

1. Conduct a surface-level scan of the essay, noting features such as length, format, and paragraphing,

2. Read the composition, engaging in both interpretation and judgment strategies, such as “classifying error types (lexis, syntax, morphology, spelling), leading to an assessment about the command of language” (p. 21),

3. Decide on a score, combining and reviewing judgments made earlier.

Cumming et al. noted that even this prototypical decision-making process was far from monolithic. Rather, raters “integrated their decision making in complex, interactive episodes of information gathering, reasoning, and hypothesizing, prior to making, then evaluating, their assessment judgments” (p. 22).

A small number of other studies have described a diversity of decision-making sequences used in the assessment of writing. Vaughan (1991) identified the following five different reading styles. (Strictly speaking, only items 1-3 might be considered descriptions of the decision making process, with items 4 and 5 indicating scoring criteria used.)

1. A “single focus” approach, where a specific indicator or rule was used in decision making, such as looking for features that would automatically cause the essay to receive a “fail” grade,

2. A “first impression dominates” approach where initial impressions proved
especially important,

3. A “two category” approach where performance in two major areas of
performance was considered (e.g., either content or grammar must be good to
receive a passing score,

4. A “laughing rater” approach, where personal reactions to the text were
decisive, and

5. A “grammar oriented” approach where attention was focused strongly on
grammatical accuracy.

A “first impression dominates” approach was also seen in Smith’s (2000) study,
where six raters each scored three texts written by adult migrant ESL students. In Smith’s
study, the term “first-impression-dominates” was used to describe a strategy where raters
first developed an impression and then looked for evidence to support their judgment.
Smith also observed a nearly opposite “read-through-once-then-scan” strategy where the
text was first read through without comment and then reread to look for evidence that it
met a “pass” standard for each of the six subscales used in grading. A third “performance
criteria-focused” strategy was also observed, where raters scanned the text for specific
indicators to demonstrate achievement of each of the subscales, without reading the text
through. Like Vaughan, Smith also noted that individual raters tended to rely on a
particular style of reading.

Milanovic, Saville and Shen (1996) also observed a first-impression-dominates
scoring approach, although they termed this behavior a “provisional mark” strategy. Like
Smith, they observed strategies involving multiple reading: one in which all essays were
read twice as a matter of principle (termed “principled two-read”) and another in which a
second reading was conducted only when there was difficulty in assigning a score during the initial read (termed “pragmatic two-read”). The focus of each reading varied; for example, the first reading might consist of a general read-through, or a simple scan of the layout of the text, or a scan for specific textual features (much like Smith’s “performance criteria-focused” strategy). Additional information would then be collected in the second reading. Finally, Milanovic et al. also observed a fourth strategy, where readers simply read through the essay once and picked out strong and weak points.

2.2.2 Speaking assessment. Studies of criteria used by raters in speaking assessments are less common, but also suggest that raters may attend to different language features when scoring. Eckes (2009), in an extension of his work on rater types described earlier, identified several rater types within the speaking portion of the same large-scale test of German as a foreign language, again based on the relative importance raters placed on different scoring criteria. Once again it is worth noting that Eckes’ participants were trained and operational raters, and might normally be assumed to share a common understanding of the rating scale. Studies in ESL/EFL speaking tests have also observed variability in scoring criteria mentioned by trained, and in some cases operational, speaking test raters (Harding & Ryan, 2009; May, 2009; Meiron, 1998; Orr, 2002; Papajohn, 2002; Winke, Gass, & Myford, 2011). These studies are discussed in detail in section 3.3.2 below, but to summarize, such studies have found that raters may focus on different scoring criteria (Meiron, 1998), or place different emphasis on the same criteria (Winke, Gass, & Myford, 2011). Moreover, raters awarding the same score for a performance may actually have opposite perceptions of specific aspects of the response (Orr, 2002).
A small number of studies have also examined decision making processes in speaking assessments. Pollitt and Murray (1996) observed that untrained raters appeared to use two types of decision strategies when judging video-recorded examinee interviews: (a) make a quick initial impression followed by collection of more information to support/revise the initial judgment, and (b) make separate judgments of various aspects of performance and then combine them to make a final overall appraisal. These approaches appear superficially similar to the “first impression dominates” and “performance criteria-focused” approaches noted in writing assessment studies. Brown (2007), working with operational IELTS speaking test raters, similarly reported a “feature focus” where an initial judgment was followed by further evaluation and then acceptance or revision, as well as a “section focus,” where each of the sections of the IELTS interview were considered separately and weighed together to reach a final score. Meiron (1998) also described two approaches to scoring among raters trained to score the SPEAK test. One was described as a “feature focus/multiple listening” approach where audio recordings of performance were replayed multiple times during which raters listened for specific language features. This approach was used by five of the six raters in Merion’s study. A second approach was used by a single rater who listened once, then based a score on an overall impression.

2.2.3 Summary. There is considerable evidence that rater's perceptions of examinee performance may vary; variability in perceptions may be associated with differences in scoring patterns (e.g., Diederich, French, and Carlton, 1961) or conversely, different raters may assign the same score for opposite reasons (e.g., Orr, 2002). It is also clear that training and/or experience do not necessarily eliminate differences in rater's
internal scoring criteria. Raters also make use of a variety of strategies to reach scoring decisions, although two types of strategies appear most common, in both writing and speaking assessment contexts: (a) forming an initial impression followed by additional evaluation, and (b) dividing attention across discrete aspects of the performance. What these differences mean in terms of the scores awarded remains unclear, however. The question remains as to whether differing strategies systematically lead to differing scores, or simply represent two different ways of reaching the same result.

2.3 Effect of Rater Professional Background on Scores and Decision Making

Given the ways in which raters may vary in their scores or cognition, an obvious next question is how such variability comes about. That is, what rater characteristics are associated with differences in scoring patterns or cognition? In a recent review of the writing assessment literature, Barkaoui (2007) observed that studies of ESL/EFL essay scoring have largely focused on raters' academic or professional background, first language background, or teaching or scoring experience. Examination of the speaking assessment literature suggests a similar set of concerns. The attention given to rater background seems to follow from the view that the experiences raters bring to the rating task are likely to influence scoring judgments (Weigle, 2002). This section examines the empirical support for this view from studies of the effects of rater professional background on scoring and decision making.

2.3.1 Writing assessment. Within writing assessment, empirical investigations have often focused on raters’ academic domain, comparing readers from an ESL or EFL teaching background to readers from other fields (Barkaoui, 2007). These include studies
where raters from an ESL background were compared to those with a background in L1 English teaching; in a number of cases the purpose was to determine whether rater groups varied in their evaluation of essays written by ESL or English L1 writers, an issue of interest in college composition programs which must fairly evaluate both types of students. For example, Brown (1991) compared holistic scores awarded by eight ESL and eight English faculty to 56 essays obtained from an undergraduate first-year English composition course and another 56 essays obtained from an ESL version of the same course. No difference was seen between rater groups in terms of overall scores, nor was there any difference in groups for scores given to essays produced by English NS or NNS students. On the other hand, when asked to rank the negative features of the essays, ESL faculty most commonly selected content while English faculty chose syntax.

In a study of the TOEFL TWE employing rater think-aloud reports, Cumming, Kantor, and Powers (2002) similarly found that raters from an English L1 teaching background made more mention of language use issues while ESL teachers focused more on content features. O'Loughlin (1992) observed that English L1 writing teachers were more severe overall when scoring essays produced by ESL students and also appeared to differently weight the importance of particular language features. While subscores for argument and organization were highly correlated with overall scores for both groups, for English L1 teachers scores for language use features (spelling/punctuation, grammar, and appropriateness) were more related to overall scores compared to ESL teachers. In contrast, Song and Caruso (1996) obtained nearly opposite results, finding that English faculty gave significantly higher holistic scores than ESL faculty, for essays by both ESL
and English L1 writers. Moreover, when using an analytic scale covering ten language features, scoring patterns were similar between the two groups of faculty.

In a broader examination of the effects of rater academic background, Mendelssohn and Cumming (1987) collected ratings from college faculty with backgrounds in engineering, English literature, and ESL. Faculty were asked to rank eight essays written by ESL students; these essays had also been modified to exhibit different combinations of effective/ineffective language use and effective/ineffective rhetorical organization (four combinations in total). All faculty consistently gave top rankings to essays with both good language use and organization and low rankings to essays that were ineffective in both categories, but when an essay showed mixed features, engineering professors gave higher rankings to essays with good language use, while English and ESL faculty preferred essays with better organization. Similarly, Santos (1988) found that college faculty in the physical sciences awarded lower scores for “language acceptability” compared to humanities/social sciences faculty, although scores for five other language features were similar for across faculty groups. Later, when asked to judge the seriousness of specific errors in the essays, faculty judgments were again largely in agreement, but instances of disagreement typically involved physical sciences faculty indicating more irritation with a specific error. Differences in the perception of error gravity was also examined by Vann, Meyer, and Lorenz (1984) who presented faculty with 12 typical ESL sentence-level errors and asked them to judge the seriousness of each for academic writing. A total of 319 faculty from the domains of humanities/social sciences, biological sciences, and math/physics participated, and while all three faculty groups were similar in how they ranked the seriousness of the error
types, in absolute terms math/physics faculty were least tolerant of errors (i.e., they scored the errors as more serious) while faculty in the humanities/social sciences were generally most tolerant.

Finally, a few studies have examined the judgment of language teachers versus native-speaking laypeople. Shohamy, Gordon, and Kraemer (1992) compared holistic scores produced by trained ESL teachers with untrained English L1 laypeople. Shohamy et al. found no difference between the groups in scores awarded to letters written by secondary level EFL learners. Furthermore, additional groups of laypeople and ESL teachers underwent rater training, and while training resulted in a significant improvement in inter-rater reliability, the degree of improvement was the same for both groups. In contrast, Schoonen, Vergeer, and Eiting (1997) compared the performance of teachers and laypeople in scoring a writing test for secondary-level Dutch students, and saw differences in the ways rater groups interacted with the tasks and language features being scored. When grading essays, teachers were more reliable than laypeople in scoring language usage, but when scoring revision tasks, where examinees were required to edit a text, no differences in scoring reliability were seen between the groups. Schoonen et al. commented that the relatively narrow range of potential responses to the revision tasks made it possible to write more precisely focused scoring instructions, which may have helped the laypeople to score language use more reliably.

2.3.2 Speaking assessment. While many of the studies described in the previous section were concerned with how university faculty from various backgrounds evaluated academic writing, parallel studies in speaking assessment have focused on the judgments of ESL teachers and content domain experts within the context of testing for specific
purposes. One example is Elder (1993), who examined the reactions of seven ESL and eight subject specialists towards the classroom performance of student math/science teachers who did not speak English as a first language. Raters from both ESL and math/science backgrounds gave similar ratings for overall communicative competence, but in terms of analytic scores awarded for specific language features, subject specialists seemed to place the most weight on interaction. Specifically, in a stepwise multiple regression, scores for interaction were the only significant predictor of scores for overall communicative effectiveness. In contrast, for ESL specialists the best predictors were scores for subject-specific language and comprehension of students' input, suggesting ESL teachers were attending more to these aspects. While it seems odd that ESL specialists would focus on subject-specific language, Elder suggested that in fact the ESL teachers were focused on language generally, while math/science specialists were attending more to classroom teaching strategies which were captured by the “interaction” subscale, such as checking students understanding, or using a proper level of formality.

A similar pattern of overall agreement but with differences in specific perceptions was found by Brown (1995), who compared Japanese foreign language teachers with tour guides who worked with Japanese tourists. In a language test for tour guides, Brown found that teachers and tour guides were similar in their marks for task fulfillment, with the exception of a task where the examinee must deal with an upset customer. Tour guides were much more severe in scoring this task, which Brown suggested was due to tour guides viewing the task as representing a crucial skill (dealing with unhappy customers) and so being unwilling to accept substandard performance. In addition, while both groups were similar in their overall severity, the language teachers were more severe
in their appraisal of grammar, vocabulary, and fluency, and less severe in grading pronunciation, perhaps because teachers were more habituated to nonnative Japanese pronunciation.

In a similar vein, Lumley (1998) examined the contention that ESL teachers would score more leniently than doctors in an ESL language test for health professionals (the Occupational English Test, OET). It had been implied that the ESL teachers who served as operational raters were unduly generous in their scores due to sympathy for immigrants. Lumley found that, in fact, groups of ESL teachers and doctors were in broad agreement regarding pass/fail decisions, although there was considerable individual variation in both groups. That is, in only a handful of exceptionally good or poor cases was the pass/fail decision unanimous among all judges, whether doctors or teachers. Moreover, examination of subscale scores suggested that, on average, ESL teachers were actually more severe than doctors. Interestingly, a later study of raters scoring open-ended written responses to the OET listening test observed that one rater, trained as an ESL teacher but with a background in health sciences research, was more critical of examinees' errors in using medical terminology than were two other raters with backgrounds primarily in ESL teaching (Harding & Ryan, 2009).

Finally, Douglas and Myers (2000) observed that veterinarians and applied linguists made similar comments when asked to evaluate videotaped examples of international veterinary students conducting doctor-client interviews with livestock owners. In unguided group discussions, both groups picked out similar aspects of the performances, but veterinarian school faculty focused somewhat more on professional issues such as establishing rapport and coverage of medical content, while the applied
linguists placed rather more emphasis on language features such as asking clarification
questions.

In addition to the comparisons with domain experts described so far, other studies
have examined differences in the judgments of language teachers and naive native
speakers. Such studies are often motivated by the desire to understand how speaking
ability might be evaluated in “the real world,” or to validate scores for oral proficiency
interviews where language proficiency is defined in terms the ability of a native speaker
to understand the examinee's output (c.f. Lantolf & Frawley, 1985). In one early study,
Galloway (1980) reported relatively little difference in scores awarded by U.S. Spanish
teachers and residents of Spain who were asked to evaluate videotaped performances of
university-level Spanish students. On the other hand, untrained Spanish speakers living in
the United States were generally more lenient than either teachers or Spanish nationals,
suggesting that exposure to the students' L1 (English) may have influenced their
impressions. In contrast, Barnwell (1989) observed that Spanish speakers living in Spain
were relatively critical in their evaluations of taped Spanish oral proficiency interviews,
compared to the judgments of a trained ACTFL rater. For three of the four Spanish
students evaluated, the majority of the 14 Spanish laypeople awarded a score lower than
that given by the ACTFL rater. A comparison of native speaker and official ACTFL
ratings was also conducted by Norris (1996) within the context of a German language
simulated oral proficiency interview (SOPI). Norris observed that the correlation between
ratings produced by German university students and official ACTFL ratings was only .52,
suggesting only moderate agreement in the ranking of examinees.
Language teachers and native-speaking laypeople have also been observed to differ in their appraisal of specific language features. Chalhoub-Deville (1995) found that Arabic speakers from different backgrounds emphasized distinct language features when holistically scoring speech produced by university-level American students of Arabic. Three dimensions of language ability were identified as underlying raters' holistic judgments, with each of the three rater groups in the study appearing to emphasize a different dimension. Untrained Arabic speakers living in Lebanon emphasized grammar and pronunciation, while Arabic teachers living in the United States relied more heavily on a dimension associated with creativity and adequacy of information, and naive Arabic speakers living in the U.S. focused most on length and the amount of detail provided. Hsieh (2011) also noted differences in the perceptions of ESL teachers and naive English L1 undergraduates within an English speaking test used at a U.S university for certifying the language abilities of international teaching assistants. While overall judgments of oral proficiency were similar for ESL teachers and undergraduates, written comments suggested that ESL teachers considered a variety of linguistic features separately, while undergraduates were more holistic in their approach to decision-making.

2.3.3 Summary. To summarize, a common finding from both writing and speaking assessment seems to be that raters' overall perceptions of language ability are often rather similar, but raters from different academic or professional backgrounds may also systematically vary in how they perceive specific aspects of performance. Differences in perceptions in some cases seem related to differences in professional values (e.g., Brown, 1995; Douglas & Myers, 2000; Elder, 1993). On the other hand, domain-associated differences in scoring do not always show a consistent pattern. For
example, while faculty in the natural sciences were generally found to be more critical of errors in writing (Santos, 1988; Vann, Meyer, & Lorenz, 1984), ESL teachers were actually found to be stricter than doctors in grading the oral performance of health professionals (Lumley, 1998). While it seems likely that professional background may influence raters' perceptions, the nature of this impact appears variable and may depend on the specific features of the scoring task (e.g., Schoonen et al., 1997).

### 2.4 Effect of Rater Language Background on Scores and Decision Making

As mentioned at the start of section 2.3, studies of factors contributing to rater variability have often focused on rater professional background, language background, or experience. This section considers differences between raters from native- (NS) or non-native (NNS) language backgrounds. A common theme in these studies has been to investigate whether NS and NNS language professionals are similar in their judgments of language ability, and so can be used interchangeably in language performance tests. Also of interest have been the reactions of raters to examinees with a common language background, where shared cultural or linguistic factors may influence scoring, as well as situations where raters are otherwise familiar with examinees' first language or interlanguage, which may influence raters' reactions to features such as pronunciation.

#### 2.4.1 Writing assessment

Comparisons of the scoring judgments of NS and NNS language teachers have produced mixed results, with some researchers reporting little difference between the two groups while others have observed more substantial divergence. In the former category is Shi (2001), who found no difference in the holistic scores awarded by Chinese and expatriate EFL teachers to ten essays written by Chinese
university students. No rubric was provided, and written comments were collected to determine the criteria raters used in their judgments. The frequency with which various language features were mentioned also showed similar overall patterns for both groups, but differed in a few specific cases such as ideas and organization, which was more often mentioned by NNS teachers, and intelligibility and accuracy, which was more often mentioned by English NS teachers.

Several studies have examined differences between Japanese L1 and English L1 speakers. Connor-Linton (1995) found that American and Japanese English teachers were similar in the scores they awarded to essays written by adult Japanese students, but that the two groups differed in the specific language features attended to, with Japanese teachers focusing on accuracy and Americans focusing on discourse and specific grammar issues. Kobayashi (1992) observed differences in the severity with which English L1 and Japanese L1 speakers scored different language features. Kobayashi’s study involved a total of 269 raters, including professors, graduate students, and undergraduates, who scored two English essays written by Japanese university students. Essays were scored on four scales, and Japanese L1 raters generally gave higher scores for grammaticality and naturalness, while English L1 raters gave higher scores for clarity of meaning and organization (with the exception of English L1 undergraduates, who awarded relatively low scores across the board).

Japanese raters’ degree of familiarity with English has also been found to influence their judgment of English language writing. Rinnert and Kobayashi (2001) asked Japanese undergraduates, Japanese EFL instructors, and American EFL instructors to score Japanese EFL essays that had been altered to exhibit different kinds of language
problems along with rhetorical features characteristic of either American or Japanese writing. Essays were scored for general quality as well as six specific language features. For Japanese undergraduate students with little training in English writing, scores awarded for content showed by far the highest correlation with scores for overall quality, suggesting these students were mostly focusing on content. In contrast, Japanese students with a stronger English background, as well as Japanese and American EFL teachers, appeared to have a more balanced focus on all of the language features. Written comments were also collected from the raters and appeared to follow a progression, where comment patterns for the Japanese speakers became more and more like those of American teachers as the Japanese speakers gained more experience with English. Rinnert and Kobayashi suggested that with increasing exposure to English, the perceptions of the Japanese had adapted to be more in line with English language norms. This conclusion was supported in a follow-up study by Kobayashi and Rinnert (1996), who found a pattern of adaptation in readers' reactions to essays with either English or Japanese rhetorical patterns. When asked to compare an essay with a Japanese rhetorical pattern and one employing a more American organizational style, English L1 teachers gave higher holistic scores to essays with an American organizational pattern, Japanese EFL teachers and experienced students gave similar scores to both, and inexperienced students favored the Japanese-style essay.

Erdosy (2004) also observed that cultural and educational background appeared to play a role in the scoring criteria used by two English NS and two NNS raters as they scored essays taken from the TOEFL Test of Writing. Comments made during think-aloud protocols suggested that individual raters focused on different aspects of
performance, and that these differences seemed to coincide with pedagogical approaches used in the rater’s home country. For example, a focus on grammar/structure for one NNS rater was tied to pedagogical approach which focused on explicit grammar instruction, which the individual had experienced as an EFL language learner. Erdosy ultimately proposed that differences in criteria arose because raters “differed in their perceptions of language proficiency, in their assumptions of how language could be acquired, and in their definitions of the endpoints of a learning curve” (p. 55).

The studies described so far all suggest that rater background may influence perception of the specific features that are important for quality writing. But, it should be noted that in these studies raters were given neither rubrics nor training, so some degree of difference in perception is hardly surprising and raises the question of the degree to which raters might differ in more structured testing situations. One such study in a more structured context is Johnson and Lim (2009), which examined trained essay raters in an operational university English test (the MELAB) and found no difference in the severity of holistic scores awarded by NS raters and NNS raters from a variety of language backgrounds. Moreover, the NNS raters did not show any systematic bias towards examinees from their own or other language backgrounds. On the other hand, Hill (1996) found quite different results in an operational test used to assess the English ability of Indonesian classroom teachers. In this case, English L1 raters from Australia were more severe than Indonesian raters in scoring essays written by Indonesian examinees. Interestingly, this particular test targeted the local Indonesian variety of English, and the Australian raters, unfamiliar with this variety, were asked to base their decisions on “a level of performance which would be considered acceptable for a foreign language
teacher in Australian classrooms” (p. 35). It was made clear that the test targeted the local variety of English and raters were also given training in the use of the scoring rubric, but it is possible that the two groups of raters were still applying different standards, with the judgments of Australians based on an Australian context and the judgments of Indonesians based on an Indonesian context.

2.4.2 Speaking assessment. Like much of the work mentioned in the previous section, several studies in speaking assessment have investigated the criteria used by NS and NNS raters in unstructured scoring situations, with the result generally being that raters from both groups are similar in their overall severity but differ in their perceptions of specific details of performance. One such example is Zhang and Elder (2011) who found no difference in holistic scores awarded to Chinese undergraduates by Chinese L1 and English L1 EFL instructors. Raters scored videotaped oral speech samples collected in the speaking portion of the Chinese College English Test (CET-SET), and although the raters were untrained and no scoring rubric was provided, no difference in severity or intra-rater consistency was seen between the rater groups. However, an analysis of written comments provided by the raters showed statistically significant differences in the number of times different language features were mentioned, although on a percentage basis the frequency of comments for different features were similar, with the single exception of “linguistic resources”, which was more often mentioned by Chinese L1 teachers. More substantial differences in rater perceptions were reported by Gui (2012), who examined scores and rater comments made during an English speech contest at a Chinese university. Gui found that, while English L1 and Chinese L1 teachers of EFL showed high agreement regarding contestants’ scores and the competition winners, they
exhibited opposite patterns in their comments on examinees' speech delivery, usage of expressions, and pronunciation. The comments of Chinese teachers in these areas were unanimously positive, while the comments of American English L1 judges were much more negative.

Kim (2009b) also observed little difference in the scores of Canadian English L1 and Korean L1 English teachers who evaluated a variety of English language speaking tasks performed by Korean ESL students. No rubric or training was provided, and for each audio-recorded response raters were asked to award a holistic score and write a brief scoring justification in their L1. No difference was seen in the rater groups in either overall severity or internal rater consistency as measured using many-facet Rasch analysis. In addition, a separate G-theory analysis found little difference in their relative contributions to overall variability in scores (Kim, 2009a), with rater effect accounting for less than 1% of the variability in scores for both English NS and NNS raters. Interaction effects between raters and examinees were a bit higher but still similar for both NS and NNS raters, accounting for 5.5% and 4.0% of overall score variance, respectively. A total of 19 language features were mentioned in the written comments, with general patterns in the frequency of comments being roughly similar for both groups. However, English L1 raters made more comments overall, and for a few specific language features the difference was considerable. Compared to the Korean L1 teachers, English L1 teachers made three times as many comments (or more) regarding overall language use, fluency, and specific grammar use, suggesting a relatively greater focus on these features.
As with the studies from writing assessment, the question arises as to whether differences between NS and NNS raters might be minimized in operational testing contexts where scoring rubrics and training have been provided to promote a common understanding of the rating scale and scoring criteria. In terms of scores, Xi and Mollaun (2009) found high levels of agreement between the scores of trained TOEFL iBT Speaking Test raters working in India and operational raters working in the United States. Moreover, no difference in agreement was seen in scores awarded by Indian raters to Indian examinees. On the other hand, a recent study suggests that familiarity with the examinees' L1 may influence the perceptions of trained native-speaking raters, particularly for judgments of pronunciation or accent. Winke, Gass, and Myford (2011) found that English L1 undergraduates who were trained to score the TOEFL iBT Speaking Test were more lenient in their judgments when they had studied the L1 of the examinee. Specifically, raters who had studied Spanish or Chinese gave higher holistic scores to examinees from Spanish or Chinese language backgrounds. In retrospective verbal reports, some raters also admitted to a conscious preference or dislike of particular accents or of strong accents generally, which influenced their scoring. It should be noted here that the undergraduates participating in the study were not employed in the operational scoring of the TOEFL and so were perhaps less concerned about the consequences of bias.

Nonetheless, similar bias has been reported among operational raters in a large-scale English test, the IELTS. Carrey and Mannell (2009) found that IELTS examiners working in different test centers in five different countries tended to give higher pronunciation scores to examinees with familiar accents. Examiners analytically scored
audio recorded IELTS interviews with test takers from Chinese (Cantonese), Korean, and Indian language backgrounds. Examiners working at test center in Hong Kong gave higher pronunciation scores to Chinese examinees, those working at test center in India awarded the highest scores to Indian candidates, and examiners working in Korea awarded the highest scores to Korean examinees. All of the examiners at the Korean location and 95% of examiners in Hong Kong were native English speakers, while 90% of the examiners in India reported that English was their second language. Carrey and Mannell attributed the findings to a phenomenon they termed “interlanguage phonology accommodation,” where exposure to individuals from a particular language background leads to a psychoacoustic process that changes the basic phonetic perceptions of examiners. In other words, differences in the scoring of pronunciation may not have been the result of an explicit preference for a particular accent, but rather the product of a more basic psycholinguistic change in perception.

2.4.3 Summary. It appears that raters from different language backgrounds may show a considerable degree of agreement in gross judgments of language ability, even when no rubric or training has been supplied to standardize raters' perceptions. Nonetheless, individuals from different language backgrounds may show differences in their reactions to specific language features. Such variability may be related to differences in cultural values (e.g., Kobayashi & Rinnert, 1996; Rinnert & Kobayashi, 2001), pedagogical methods (e.g., Erdosy, 2004), or exposure to learners' L1 or interlanguage (Winke, Gass, & Myford, 2011; Carrey & Mannell, 2009).
CHAPTER 3

Expertise in Scoring Language Performance Tests

The previous chapter considered the ways in which raters may vary in scoring patterns as well as the aspects of rater background that may influence such rater variability. A third rater variable that may influence scoring patterns or decision making is rater expertise; this chapter focuses on this issue. First, to provide a basis for understanding what expertise in scoring language performance tests might look like, section 3.1 briefly describes how the cognition of experts differs from non-experts, as reported in the general psychological literature on expertise. Section 3.2 then examines differences between experts and novice language test raters in terms of their scores and decision making behaviors. Finally, Section 3.3 considers the development of expertise through training, summarizing studies of the effect of training on raters' scoring patterns and perceptions.

3.1 Characteristics of Expertise

In addition to professional and language background, other variables that have been of interest in studies of raters are the effects of scoring experience and training. The fundamental issue underlying this domain of research is the nature of rater expertise, and a full understanding of this issue requires answers to a number of questions, including:

- Who can (or should) be considered to be an expert rater?
- How do experts and novices differ in their scoring, behavior, or cognition? (i.e., What do expert raters do that novices do not, or cannot, do?)
- How does expertise in scoring develop?
How do various interventions (e.g., training) contribute to the development of rater expertise?

Given that such questions are the focus of this dissertation, the remainder of this chapter explores issues related to rater expertise. As a starting point for considering these questions, this section briefly explores the characteristics of experts as they have been conceptualized in the domain of expertise research.

3.1.1 Who is an expert? What is expertise? The Random House Webster's College Dictionary (1997) defines an expert as “a person who has special skill or knowledge in a particular field” (p. 459). Similarly, in the preface to a major reference work on expertise, Ericsson (2006) defines expertise as “the characteristics, skills, and knowledge that distinguish experts from novices and less experienced people” (p. 3). Despite these simple definitions, it is not always so straightforward to tell who is an expert and who is not, and a variety of criteria have been used to identify experts within the expertise literature. Criteria include social recognition of expert status, the amount of relevant experience possessed, and capability for reliably excellent performance (Ericsson, 2006). The social criteria approach defines experts as those individuals are viewed as having expertise by their peers; this approach has been used as a necessity in some domains where objective indicators of performance are unavailable and it is difficult for non-experts (such as researchers) to identify top performers. Similarly, professional credentials have also been used as an indicator of expertise (e.g., Cammerer & Johnson, 1991). Definitions based on social criteria or length of experience have been criticized, however, based on the observation that such “experts” may exhibit
performance that is little better than novices, even within the area of supposed expertise; a notable recent example might be the performance of professional investors in judging the risk associated with mortgage-backed securities. In response, Shanteau (1992) has pointed out that the literature suggests certain domains are more amenable to accurate decisions or judgments than others. Specifically, experts' decision making will be more accurate in domains where phenomena are more constant, concrete, or predictable, such as chess, livestock judging, or weather forecasting. Decision performance of experts will be poorer in less predictable domains involving human behavior, such as clinical psychology, student admissions, and the stock market. Language testing would seem to fall into the latter category, highlighting the challenges that even expert language raters must face.

More commonly, experts have been defined in terms of outstanding performance, as in the definition used by Phillips, Klein, and Sieck, (2004): “When we speak of expertise we refer to individuals who have achieved exceptional skill in one particular domain” (p. 299). Exceptionality is a key feature of this definition: experts are top performers. This view of expertise underlies the discussion of the psychological characteristics of experts in the next section.

3.1.2 Psychological characteristics of experts. If experts are seen as those capable of superior performance, then one must wonder how experts are able to perform at such a high level. As a starting point, it might be useful to consider the limitations to performance that experts somehow overcome. Mislevy (2010) divides such limitations into processing limitations and knowledge limitations. Processing limitations include such things as limitations to working memory, unconscious biases in reasoning, and
imperfections in our sensory systems. Knowledge limitations include the inability to
determine which information is relevant and which is not, difficulty in integrating
information, being unable to predict events, and not knowing how or when to take action.
Presumably, experts have characteristics that allow them to overcome such limitations.

But the question remains: what is it that makes experts exceptional? A variety of
sources have summarized the psychological characteristics of experts and provide insight
into this question (e.g., Feltovich, Pietula, & Ericsson, 2006; Glaser & Chi, 1988;
Mislevy, 2010; Phillips, Klein, & Sieck, 2004; Shanteau, 1992). Such lists of
characteristics overlap to a considerable degree. Feltovich et al. (2006) provide a
particularly helpful review, which will be used as the basis for this discussion. Feltovich
et al. (2006) describe the following as characteristics of expertise:

1. Expertise is limited in its scope and elite performance does not transfer
2. Knowledge and content matter are important to expertise
3. Expertise involves larger and more integrated cognitive units
4. Expertise involves functional, abstracted representations of presented information
5. Expertise involves automated basic strokes
6. Expertise involves selective access of relevant information
7. Expertise involves reflection
8. Expertise is an adaptation
9. Simple experience is not sufficient for the development of expertise (pp. 47-60)

Each of these characteristics will now be briefly discussed, followed by a brief
discussion of how these characteristics might apply to the task of making judgments of
language ability.

*Expertise is limited in its scope and elite performance does not transfer*. Feltovich
et al. calls this “one of the most enduring findings in the study of expertise” (p. 47).
While individuals may be competent in a variety of areas, people seldom achieve elite
performance in more than one domain. Expertise is also highly context-specific, and expertise in one domain does not ensure competence in other, even closely related, domains. For example, physicians can vary widely in the quality of their specific clinical judgments depending on their previous experiences (Feltovich et al., p. 47).

**Knowledge and content matter are important to expertise.** It is common sense to say that experts know more than non-experts. Aside from being a database of information, however, robust knowledge of a domain also leads to cognitive changes that allow experts to process information more effectively, minimizing the processing limitations mentioned by Mislevy (2010).

**Expertise involves larger and more integrated cognitive units.** One consequence of having great knowledge about a subject is an improved ability to recognize patterns and use these patterns to package information into larger chunks. Larger chunks then make it possible to hold more information in memory. With increasing knowledge, experts are better able to remember information in terms of patterns rather than attempting to remember every detail.

**Expertise involves functional, abstracted representations of presented information.** Experts' ability to organize information into bigger chunks in large part relies on the structure of their previous knowledge: familiarity with patterns in the domain makes it easier to organize new information in a deeper and more principled way. In contrast, non-experts' efforts to organize information are more haphazard, more superficial, and are less effective for the content or task at hand.

**Expertise involves automated basic strokes.** Expertise typically entails large amounts of practice with specific tasks; as a result, actions become automatic and can be
carried out more quickly and with less attention. Moreover, the automation of lower-order cognitive processes seems to be a prerequisite for proficient higher-order reasoning, monitoring, or integration.

*Expertise involves selective access of relevant information.* Experts are able to focus their attention on relevant information and notice details that non-experts might miss. These abilities require an understanding of what information is important and what is not.

*Expertise involves reflection.* Metacognition is important for expert performance because it helps experts to monitor their own understanding and progress, avoiding pitfalls and mistakes that then require backtracking and waste time. Experts may also be more likely to withhold judgment in favor of collecting more information, which also helps avoid the need to revise decisions (Glaser & Chi, 1988). Finally, metacognition helps experts to recognize atypical situations and adjust accordingly.

*Expertise is an adaptation.* Expertise is not simply a matter of degree but also a matter of kind; in the words of Feltovich et al. “experts certainly know more, but they also know differently” (p. 57). Expertise can be viewed as set of adaptations involving knowledge, skills, and cognitive mechanisms that allow efficient and effective performance of specific tasks.

*Simple experience is not sufficient for the development of expertise.* Feltovich et al. note that reaching exceptional levels of performance requires long periods of dedicated effort, often 10 years or more depending on the domain. But, once general competence is achieved, usually in a relatively short period of time such as 50 hours, there is little additional benefit from additional unstructured experience. Rather,
advancement to an elite level of performance requires directed practice targeting specific aspects of performance, as well as feedback and reflection.

It should be noted that the characteristics of expertise summarized above focus on the internal processes of individuals. Mislevy (2010) makes the point that expertise is also socially mediated. Knowledge representations used by experts exist not just inside the expert's head, but also in external artifacts that help experts to organize and manipulate information. Mislevy lists items such as maps, mathematical formulas, and insurance forms as examples of such artifacts, which perform important functions such as supplementing working memory, providing a common framework for understanding things, and encouraging people to think in non-intuitive ways previously discovered to be useful by others.

So, how might the characteristics of experts identified in the expertise literature apply to the human scoring of language performance tests? Several of the characteristics in Feltovich et al.'s list arguably provide useful starting points for thinking about rater expertise. For example, the finding that expertise is limited in scope might suggest that rater experiences in teaching or even other testing or scoring contexts do not necessarily translate into expertise in scoring a particular test. In another example, the finding that simple experience is not enough to develop elite-level skill might suggest that considerable training or focused experience is needed to develop high levels of scoring consistency and accuracy. Also, professional or rater training may provide knowledge structures around which perceptions of examinee performance can be organized, allowing raters to be more focused on relevant language features as well as better able to process and monitor the totality of the examinee's performance. Finally, Mislevy's point regarding
the social dimension of expertise might suggest that the ability to make effective use of artifacts such as rubrics or other scoring aids is also an element of rater expertise. While these points are speculative, specific findings from the expertise literature may prove useful in interpreting the results of empirical studies of language raters, and in the following sections will be referenced as appropriate.

3.2. Influence of Experience on Scoring Patterns and Decision Making

This section now turns to empirical studies of rater expertise in language performance assessment. Like the general expertise literature, rater expertise has been operationalized in a variety of ways within the domain of language testing. Most typically, expertise has been investigated in terms of experience, rater training, or a combination of both. Of the studies described in this section, the most common approach has been to operationalize expertise as a combination of both language teaching background and training in scoring (e.g., Barkaoui, 2010a, b; Erdosy, 2004; Huot, 1993; Kim, 2011; Sakyi, 2000, 2003; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994). On the other hand, a few studies have operationalized expertise purely in terms of teaching (e.g., Cumming, 1990; Delaruelle, 1997) or scoring (e.g., Huot, 1988; Lim, 2011), and in one case, expertise was operationalized as agreement with other scorers (Wolfe, 1997; Wolfe, 2006; Wolfe, Kao, & Ranney, 1998). In evaluating the literature on rater experience, it should be kept in mind that different studies have operationalized expertise in different ways, which may very well influence findings.

Of the relatively modest number of studies that have explicitly examined the effects of experience on raters' scoring or decision making in language testing, most have
been conducted in writing assessment contexts. These are considered first, followed by studies of rater experience in speaking assessment.

3.2.1 Writing assessment. One area of interest within writing assessment has been to investigate differences in the native perceptions of individuals with more or less language teaching experience, who are asked to make scoring decisions without the aid of a scoring rubric or training. For example, Cumming (1990) examined the scores, scoring criteria and scoring processes of seven student teachers and six experienced teachers scoring ESL university placement essays. A total of twelve essays from writers of varying English proficiency and writing experience were scored on three scales; the student teachers gave significantly higher scores on two of these scales. An analysis of rater verbal comments made while scoring indentified a total of 29 different decision making strategies, and while there was considerable individual variation, significant differences between groups were seen in the prevalence of eleven of the strategies. Overall, experienced teachers spread their attention more evenly across a variety of language features while novice teachers relied on two major strategies: “classify errors” and “edit phrases.” In addition, expert teachers also made more use of reflective strategies, consistent with findings from the general expertise literature that reflective practice is an important element of expertise (Feltovich et al., 2006).

In a contrasting example, Delaruelle (1997) reported little difference in the language features mentioned by experienced and inexperienced teachers. Three novice teachers with less than 1 year of ESL teaching experienced were compared to three experienced teachers who were enrolled in or had finished a graduate teaching program; both groups had no prior experience with the test in question. Responses to two writing
tasks produced by adult ESL examinees were scored holistically and rater comments were collected using a think aloud protocol. Delaruelle identified 14 different features mentioned by raters, with organization, grammar, and task fulfillment being the three most commonly mentioned items for both novice and experienced teachers. While overall patterns in the frequency of comments were similar, differences were seen in the style of commenting: novices tended to consistently follow a particular sequence while making comments, often starting by mentioning a possible score or commenting on the overall organization of the response. In contrast, experienced teachers addressed a wider variety of issues in their opening comments, adjusting more to the specific characteristics of a given piece of writing. This finding suggests the experienced teachers were more adaptable in their response, and fits well with Cumming's (1990) observation that experienced teachers focused on a wider variety of language features. Cumming suggested that experienced teachers had a broader understanding of the “problem” of scoring essays and thus made better use of a variety of strategies, a view quite similar to the finding from the general expertise literature that experts conceptualize problems in deeper, more principled, and more effective ways than do novices (Glaser & Chi, 1988; Feltovich et al., 2006, see also item 4 in section 2.5.2 above).

Unlike Cumming (1990) and Delaruelle (1997), most studies of experienced vs. novice teachers have provided scoring instructions and training, with the focus being on how the two types of raters apply the scoring system. For example, Sakyi (2003) provided scoring rubrics and training to five novice ESL teachers (recent B.Ed graduates) and five teachers with five or more years of teaching experience. Following training the teachers then holistically scored six essays from the TOEFL Test of Written English, with
concurrent think aloud protocols and retrospective comments collected to document
decision making processes. Sakyi found that experienced teachers were faster in scoring,
reading through the essay quickly to evaluate the structure and content of the text and
then summarizing the essay's qualities at the end. Novice teachers were slower and more
focused on errors, in some cases editing errors as they read. In addition, novice teachers
spent more time simply trying to understand what the writer was trying to communicate.
Both experienced and novice teachers employed a variety of scoring strategies, although
novices were more likely to refer to the scoring rubric, trying to match the essay to
particular rubric descriptors. As with Cumming's study, experienced raters seemed to
have a broader or more intuitive understanding of the scoring task and were faster as
well, suggesting higher efficiency in carrying out the task, again another characteristic of
expertise generally.

Differences in reading style were also associated with better rating performance
by Wolfe and colleagues (Wolfe, 1997, 2006; Wolfe, Kao, & Ranney, 1998), who
investigated the cognition of trained essay raters in a large-scale English L1 writing test.
In Wolfe's studies all raters were trained and experienced in scoring the test, but were
divided into three groups labeled “competent,” “intermediate,” and “proficient” on the
basis of correlations with the scores of other raters. Verbal report data showed that
proficient raters (who had the highest correlations) made fewer early decisions and were
more likely to review the essay, following a read-review-decide decision process, while
raters in the other groups were more likely to make early decisions and then
monitor/revise their scores. The strategy of withholding judgment in favor of collecting
additional information has also been reported in the expertise literature (see item 7,
section 2.5.2 above). Unlike Sakyi's (2003) study, both groups showed a similar degree of focus on different language features, perhaps not surprising given that all raters were operational scorers.

It is also possible that rater experience may interact with different features of the testing context such as the type of rating scale used. Barkaoui (2010b) reported differences in the decision making strategies of 14 experienced and 11 novice raters who differed in professional qualifications, teaching experience, and experience scoring EFL writing tests. The participants scored a set of essays from the TOEFL Test of Written English using both holistic and analytic rating scales, concurrently thinking out loud while scoring. Barkaoui found that the type of rubric used (holistic or analytic) had a larger influence on raters' comments than did experience. On the other hand, the impact of rubric type seemed to be greater on novices, who shifted from a focus on language use features when scoring holistically to a broader focus on different features when scoring analytically. Barkaoui attributed this change to the fact that the analytic rubric covered a range of language features, which may a broadened novice raters' scoring focus. In addition, novice raters tended to refer to the scoring rubric more and revise their scores more often regardless of scoring system used, while experts were more likely to read or re-read the essay, apparently withholding judgment until later in the scoring process.

Finally, in addition to studies of raters' scoring criteria or decision making, other studies have focused on scores. Myford, Marr, and Linacre (1996) examined severity and consistency of scores within the TOEFL Test of Written English (TWE) and found no correlation between scoring performance and the number of years an individual had worked as a TWE rater, the number of ratings produced, or the percentage of the rater's
time spent around nonnative English speakers. It should be noted that Myford, Marr, and Linacre's study examined scores from established raters, however, who all had scored at least 100 TWE essays. On the other hand, Lim (2011) examined raters that were new to an operational adult ESL test (the Michigan English Language Battery, MELAB), where holistic ratings were collected at roughly one-month intervals over periods ranging from 12-21 months. Using many-facet Rasch analysis, Lim found that some novice raters showed acceptable severity and consistency from the start. In other cases where raters started out being too lenient/severe or showed inconsistent scoring patterns, such problems disappeared by the second or third scoring session.

3.2.2 Speaking assessment. Relatively few studies have specifically examined the effects of rater experience on scoring in the contexts of speaking tests, but it appears that inexperienced raters do not necessarily differ from experienced raters in their scoring patterns, and when they do, their performance becomes more like that of experienced raters after a few scoring sessions. For example, Bonk and Ockey (2003), in a general validation study of a group discussion EFL test administered annually at a Japanese university, found that new raters were more likely to show inconsistent scoring patterns (Rasch model misfit) than were raters experienced in scoring the test. But, of the four (out of a total of seven) new raters showing misfit in scoring their first test administration, none were misfitting in the next administration the following year. While all raters were trained to score the test and were EFL teachers with master’s degrees in applied linguistics or related fields, new raters were also new employees of the university, and Bonk and Ockey suggested that the correction of misfit may have been the result of increasing familiarity with the student population of examinees.
In what appears to be the only study to date to place primary focus on the effects of rater experience in the scoring of a speaking test, Kim (2011) examined the scoring patterns and decision making of novice, developing, and expert raters engaged in scoring a computer-mediated test of oral English, used for placement purposes in a community adult ESL program. Raters were divided into groups of three raters each on the basis of features including teaching experience, relevant teacher preparation coursework, general experience scoring speaking tests, and previous experience scoring the community ESL test. Novice raters were completely new to the test, while developing raters had scored from 1-3 administrations and experts had scored from 4-7 administrations. Raters participated in three scoring sessions, each a month apart; in each session raters scored 30 videotaped responses on five scales addressing different language features. All raters participated in a training exercise before each scoring session, and prior to sessions 2 and 3, raters were given individual feedback based on their scoring performance in the previous session. Using many-facet Rasch analysis, Kim observed minimal difference between rater groups in terms of severity. Rater internal consistency as measured by Rasch model fit was largely within acceptable bounds for all raters; however, novice and developing raters were more likely to show significant bias interactions with other aspects of the testing situation (e.g., examinee, task, language feature) although the absolute number of such interactions were consistently low across all sessions. Raters' verbal reports suggested that expert raters gave more balanced attention to all five of the language features scored and were less likely to confuse features belonging to different scales. In sessions 2 and 3, however, novice and developing raters became better at giving equal attention to different language features and showed less confusion in
applying the different scales. Overall, increasing experience, either in terms of professional experience or experience scoring, seemed to have minimal influence on scoring patterns, perhaps because inexperienced raters were already doing a fairly good job of scoring from the start. On the other hand, scoring experience (along with feedback) did seem to have a positive effect in helping novice and developing raters apply the scoring criteria in an appropriate way.

3.2.3 Summary of experience effects. To summarize this section, studies on the influence of rater experience generally seem to suggest that prior experience, either in teaching or with the testing context, influences raters' decision making processes. Moreover, these influences seem to be consistent with findings from the general expertise literature, such as the observation that experienced raters are faster (Sakyi, 2003), consider a broader range of language features (Cumming, 1990; Kim, 2011; Sakyi, 2003), and are more likely to delay making a judgment in favor of collecting more information (Barkaoui, 2010b; Wolfe, 1997). In terms of scores, novice raters who are overly severe or lenient, or show poor internal consistency, seem to adjust to be more like their more experienced peers after a few scoring sessions (Bonk & Ockey, 2003; Lim, 2011). It is also interesting to note that novice raters may show perfectly acceptable scoring patterns right from the start, perhaps a result of rater training procedures. Such effects of training are considered in the next section.

3.3 Influence of Training on Scoring Patterns and Decision Making

Focused practice or training has been identified in the general expertise literature as a factor leading to expert performance, and it is widely accepted in the language
testing community that rater training is necessary to ensure the reliability and validity of scores produced in language performance tests (Fulcher, 2003). From a reliability standpoint, the goal of rater training is to reduce the differences in scores from different raters, minimizing rater-associated variability in scores and promoting fairness by ensuring that examinees assigned to different raters are judged on an equal basis. From a validity standpoint, the goal is to lead raters to a common understanding and application of the scoring criteria, thereby ensuring that different raters are operationalizing the construct of language ability in the same way. Rater training is not without controversy, however; while some have viewed it as a socialization into a community of raters (e.g., Barnwell, 1989; Pula & Huot, 1993) others have seen it as an indoctrination that extinguishes the very diversity in viewpoint needed to achieve a rich understanding of performance and make the best possible decisions (e.g., Moss, 1994, and to a lesser degree, McNamara, 1996). Regardless, rater training in one form or another appears to be ubiquitous, at least for standardized language performance tests (Luoma, 2004; Weigle, 2002).

The precise effects of rater training are no doubt tied to the particular activities used to train raters, and so common training procedures are summarized here. Although rater-training procedures for many large-scale language performance tests are not publicly available, procedures for several smaller tests have been described in the literature (e.g., Elder, Barkhuizen, Knoch, & von Randow, 2007; Lim, 2011), and for some of the larger tests brief descriptions of actual or parallel training protocols are available as well (e.g., Furneaux & Rignall, 2007; Shaw, 2002; Xi & Mollaun, 2009). Rater training commonly begins with an introduction to the test as well as the procedures
and criteria to be used when scoring. This is usually followed by practice scoring of a set of examinee responses, with practice scores compared to previously established references scores for the same responses. In face-to-face training sessions practice scoring is often followed by discussion of the results, and may include specific instruction in how to deal with difficult cases (e.g., Furneaux & Rignall, 2007). Where online training is used, individual written feedback may be provided which notifies raters of their specific scoring patterns and areas to target for improvement (e.g., Elder, Barkhuizen, Knoch, & von Randow, 2007). Training activities usually conclude with a certification exercise where another set of responses are scored, and if the rater's scores meet a certain level of agreement with established reference scores then the rater is allowed to continue on to operational scoring.

Within empirical studies of rater training, the basic question has generally been some version of “does it work?” That is, do training procedures actually lead to changes in rater scoring patterns or cognition such that raters are more closely in agreement in their scores, and/or their criteria or approaches for awarding scores. Such empirical studies are described next, with studies from writing assessment contexts described first, followed by studies from speaking assessment.

3.3.1 Writing assessment. A number of early studies of rater training did indeed support the conclusion that training was effective in reducing variability in scores. Shohamy, Gordon, and Kraemer (1992), who as described earlier compared ESL teachers and laypeople, also examined scoring patterns of trained and untrained English L1 speakers in a fully crossed design including four rater groups: trained EFL teachers, untrained EFL teachers, trained laypeople, and untrained laypeople (five individuals per
Training included explanation of the rating scale, practice scoring, and group discussion of the practice scores. Trained and untrained raters then scored a set of 50 written responses (letters written by secondary-level EFL students) using both holistic and analytic rating scales. Trained raters showed higher inter-rater reliability regardless of professional background. In addition, an ANOVA detected a significant main effect for training but no effect for rater background. On the other hand, intraclass correlation coefficients were .80 or higher for all groups and for both types of rating scales, suggesting that untrained raters were also capable of achieving high levels of inter-rater agreement.

McIntyre (1993) also observed a decrease in rater variability in a pre/post study of rater training. Experienced ESL teachers \((N = 83)\) completed a rather novel training session in which the rating scale was discussed, followed by the individual scoring of 15 written samples without reference to the scoring descriptors. Next came a presentation of how to score seven benchmark samples, followed by a re-scoring of the same 15 samples scored earlier (consulting the rubric this time). McIntyre compared the original scores to those awarded immediately after review of the benchmarks and found that the second group of scores were more closely clustered together, with more scores falling on the mode of the scoring distribution for each text. It should also be noted that for seven of the 15 samples, initial scores (prior to training or viewing the rubric) for the 83 raters were all within one point of the median, indicating good agreement given that a 9-point scale was used, and suggesting that a primary effect of training might be to standardize raters' scoring of difficult cases.
Weigle (1994) also used a pre/post design to examine 16 ESL teachers (eight experienced with the test and eight novices) who scored university EAP placement essays before and after training. In this case, training consisted of individual practice scoring of 17 representative writing samples, followed by review of benchmark scores for each sample and group discussion. When scoring the same essays pre- and post-training, the scores of novices changed by a greater amount than experienced raters. In addition, multi-faceted Rasch analysis showed that in pre-training, novices were generally more severe and less consistent (i.e., showed more Rasch model misfit) than experienced raters, but following training the scores from novices generally become more consistent and similar in severity to the scores of experienced raters (Weigle, 1998). Weigle (1999) also observed that untrained novice raters scored one of two writing prompts more severely than the other, and that this difference was eliminated after training. Nonetheless, considerable individual differences in severity persisted after training for both novice and experienced raters, supporting the conclusion that training is more likely to reduce within-rater inconsistency rather than eliminate differences in severity between raters.

Also, despite improvement in scoring performance, Weigle (1994) observed little change in novice raters’ scoring strategies following training. One rater who engaged in a relatively quick, global assessment of texts retained this strategy after training, and a similar lack of change was seen in a rater who carefully compared performance features to rubric descriptors. Moreover, concurrent think aloud protocols collected for the same essay scored before and after training (several weeks later) were surprisingly similar in many cases, even including identical wording, although scores might change to be more
in alignment with other raters’ values. Nonetheless, think-aloud protocols from novice raters suggested that training helped them to (a) better understand the rating criteria, (b) revise their expectations regarding test taker performance, and (c) increase their awareness of, and concern for, the standards used by other raters.

In a more recent study, Fahim and Bijani (2011) also used a pre-post design to examine the effects of rater training. Twelve experienced ESL teachers scored 15 IELTS compositions written by university EFL students; approximately a month later the raters underwent training in the use of the IELTS analytic rating scale, including familiarization activities and practice scoring, and then shortly afterwards re-scored the same 15 essays. Difference scores were computed for each rater by subtracting the rater's scores from reference scores established by a certified IELTS examiner. Following training, raters whose initial scores were in greatest disagreement with the reference scores moved considerably closer to the reference values. In addition, scores overall tended to move closer to the reference scores, suggesting that both scoring accuracy as well as inter-rater agreement improved following training.

Furneaux and Rignall (2007) also examined the scoring accuracy (agreement with established reference scores) of 12 raters before and after operational rater training for the IELTS writing test. Raters underwent a two-day initial training session which included introduction to the scoring system, review of examinee responses across a range of levels, practice scoring, and discussion. Following the training raters scored additional responses at intervals of two weeks, two months, and six months after training. Furneaux and Rignall observed a steady increase over time in the percentage of scores that were in exact agreement with reference scores, except for the final session where accuracy was
lower. Very similar results were reported by Shaw (2002), who examined the effects of a training and feedback scheme used to train operational raters in the use of a new scoring system for the Certificate of Proficiency in English (CPE) writing test. Raters were provided with initial training in the use of the new marking scheme and then marked additional essays on four separate sessions over a period of eight weeks; prior to each session feedback on each rater’s performance in the previous session was provided. Shaw found that the percentage of scores showing exact agreement with previously established reference scores increased steadily over time, with the exception of the final session where agreement fell, perhaps due to interference with a concurrent live marking session which employed the old marking scheme.

Shaw's study included a feedback component as well as an initial training. The effect of such feedback has been an area of interest for a number of other researchers, who have investigated whether feedback can improve the scoring performance of established raters, particularly in terms of bias towards the scoring of specific language features. O'Sullivan and Rignall (2007) attempted to reduce scoring bias by providing operational IELTS writing test raters with graphical feedback showing the scoring categories for which the rater's scores were too high or low, along with explanatory comments and suggestions for improvement. Two groups of 10 raters each completed an initial scoring session consisting of 25 responses, and then one group received the written feedback two weeks following the initial scoring. Both groups then scored another 25 responses. No difference in the number of rater by language feature bias interactions was seen across scoring sessions for either group of raters, nor was there any reduction in
between-rater variability. Nonetheless, in a post-scoring questionnaire raters generally indicated that they felt the feedback was useful and impacted their scoring.

Like O'Sullivan and Rignall, a number of other studies of writing tests have also seen little effect of feedback on bias interactions. On such study was reported by Elder, Barkhuizen, Knoch, and von Randow (2007), who investigated the usefulness of an online self-access training program for re-training existing raters to score a university diagnostic English test. The training consisted of scoring a minimum of 10 writing samples on nine different subscales; immediately after scoring each response, raters received a report showing the discrepancies between their scores and previously established reference scores. The training had little consistent effect on bias patterns or within-rater consistency (which was within acceptable bounds before training), but did somewhat reduce differences in overall severity between raters. Like O'Sullivan and Rignall's study, raters generally viewed the training as helpful, even though in reality the feedback had only a modest effect on scoring. In a subsequent study of the same test, Knoch, Read, and von Randow (2007) compared the online training procedure to a face-to-face session where raters were presented with information regarding their severity, internal consistency, and bias towards specific language features. Once again, no consistent effect on bias was seen for either training method, but both were equally effective in reducing between-rater variation in severity. This latter effect was most pronounced for raters who were most severe or lenient; scores for these raters adjusted towards average values following training, similar to the results of Fahim and Bijani's (2011) study described earlier. Also, raters generally thought the training was effective, although some who received the online training commented that they would have
preferred to receive face-to-face training, and vice versa. Knoch et al. noted that, of the
nine raters who showed the highest level of improvement in scoring performance, none
felt they would have preferred the other training format, suggesting that raters'
preferences may also influence the effectiveness of training.

Finally, Knoch (2011) evaluated face-to-face training for operational raters in an
ESL test for healthcare professionals, and found no effect of training on rater severity,
internal consistency, or bias, for raters of both writing and speaking tests. Knoch's
protocol was similar to the face-to-face protocol used by Knoch et al. (2007), and 84% of
raters reported incorporating the severity/bias feedback into their ratings, an interesting
finding given the lack of any actual impact on scores.

To summarize, attempts to use feedback to reduce scoring bias in established
raters have largely proved unsuccessful in writing test contexts. On the other hand, such
feedback may improve inter-rater agreement and raters have generally viewed feedback
positively, even when there was little or no effect on actual scoring.

3.3.2 Speaking assessment. Although studies that have explicitly addressed rater
training in speaking tests are sparse, they also suggest that training can be effective in
moderating raters' scoring patterns. Wigglesworth (1993) conducted a rater feedback
study much like those reported in the previous section, where eight established raters of
an adult ESL examination were given feedback on their performance in the form of
graphical representations of bias generated using Rasch analysis. Raters also participated
in re-training which consisted of a group scoring/discussion session followed by
individual practice scoring. Following feedback and training, the range of rater severity
was reduced, indicating that individual raters were scoring closer to average and were
therefore more in agreement. Wigglesworth also observed that rater bias was reduced as well, with instances of statistically significant bias interactions being substantially eliminated following training. Also, for at least three of the raters bias values were reduced across the board, indicating generally more consistent scoring across different tasks and rating scales. These latter results stand in contrast to the findings reported by similar studies within writing assessment, where feedback appeared to have no effect on the prevalence of bias of experienced raters.

In an examination of the effects of training on naive raters, Kang (2008) found that training influenced perceptions of accent by undergraduate raters who judged the speaking performances of international teaching assistants (ITAs) at a U.S. university. A total of 63 undergraduates rated 11 speech samples, after which 29 student raters participated in a one-hour session where they met informally with ITAs, with the object of reducing any negative stereotypes undergraduates might have regarding ITA pronunciation. Following training the same 11 speech samples were scored again, with trained undergraduates found to be less severe in their appraisal of accent.

Although not specifically addressing the effects of training, a number of studies have looked at variability in scoring patterns or cognition in trained raters, which provides some insight into the question of how well training actually serves to standardize scores or perceptions. One finding discussed earlier in section 2.1 is that raters continue to show differences in severity following training, and some degree of inter-rater disagreement persists over time (Brown, 1995; Lumley & McNamara, 1995; McNamara, 1996; Myford & Wolfe, 2000). In addition, studies of rater cognition suggest that trained speaking test raters also maintain differences in their perceptions of scoring
criteria and/or decision making processes. Meiron (1998) examined the cognition of both newly trained raters and re-trained experienced raters scoring audio recorded samples from the SPEAK test (Speaking Proficiency English Assessment Kit). Of the six raters examined in Meiron's study, each showed a unique pattern of attention to different language features while producing holistic scores, as indicated by the frequency with which different features were mentioned in verbal reports produced while scoring. Meiron found that raters could be divided into two general groups, one focused primarily on grammar and the other on pronunciation. Meiron also observed two different approaches to the scoring process, one involving multiple listenings to the recorded examinee responses, with a different language feature considered in each listening, and another where the recording was played once and a score awarded based on a holistic impression. Papajohn (2002) also examined the scoring criteria and processes used by newly trained SPEAK raters. Raters were being trained to score a new version of the test, and they produced written diagrams describing their decision making process as the final step of the training. A wide variety of different scoring procedures were observed, although it is worth noting that only three of the nine raters in the study subsequently passed a rater certification exercise the first time.

In a few cases, trained raters have differed quite dramatically in their perceptions. Orr (2002), in an examination of scoring criteria used by rater's within the First Certificate in English test (FCE), observed multiple cases where raters gave the same score to a particular oral response but made opposite comments regarding the quality of a particular language feature. For example, in one case two raters both awarded an examinee a score of three for “grammar and vocabulary” (on a scale of 1-5), with one
rater commenting “the vocabulary was pretty inappropriate” while the other rater commented “the vocabulary was appropriate” (p. 146). In other cases, such disagreement in perception was also accompanied by disagreement in score. Orr noted that such findings arose from both a differential focus on particular scoring criteria, as well as differences in how qualitative judgments were mapped to the rating scale; for example, a language feature characterized as “more than adequate” by two raters might receive a score of 3.5 from one rater and a 5 from the other.

Finally, in a more recent study of the TOEFL iBT Speaking Test, Winke, Gass, and Myford (2011) examined the perceptions of 15 trained undergraduate raters. In an analysis of raters’ verbal reports a total of eight major categories of comments were identified, but only three categories were mentioned by all 15 raters, while the others were mentioned by subsets of 4-12 raters. So, while there was a degree of commonality in the issues mentioned by raters, there was also a considerable degree of variability. In addition, two of the three issues noted by all raters (i.e., the examinee's accent and volume/quality of voice) had little relationship to the scoring rubric. Moreover, as described earlier in section 2.4, some individuals appeared to show a conscious bias towards certain accents, particularly when the examinee's L1 was a language that the rater had studied. It should be noted, however, that the participants in the study were not certified as operational raters and were younger, less rigorously trained, and had not completed a degree.

3.3.3 Summary of training effects. To summarize, training of naive raters seems to be helpful in making raters more internally consistent (Weigle, 1998), and may improve inter-rater agreement or accuracy as well (Shohamy, Gordon, & Kraemer, 1992;
Furneaux & Rignall, 2007; Shaw, 2002). However, it is very clear that training does not fully standardize raters' perceptions or scores. Considerable variation in rater severity persists after training (Brown, 1995; Lumley & McNamara, 1995; Myford & Wolfe, 2000) as well as variability in scoring criteria and decision making processes (Meiron, 1998; Orr, 2002; Papajohn, 2002; Winke, Gass & Myford, 2011). Moreover, re-training of established raters seems to have modest effects on scoring patterns. Attempts to use feedback to reduce the scoring bias of operational raters have produced mixed results, with an early study suggesting that bias could be reduced (Wigglesworth, 1993) and more recent studies observing little if any effect (Elder et al., 2007; Knoch, 2011; Knoch et al., 2007; O'Sullivan & Rignall, 2007).
CHAPTER 4

The Nature of Scoring Judgments in Language Performance Tests

The previous chapter considered the characteristics of experts from the perspective of both the general literature on expertise as well as studies of rater experience and training in language performance assessment. However, expertise is highly context dependent; therefore, a full understanding of rater expertise must extend beyond simple description of the characteristics of experts or non-experts and include an account of the scoring task itself. That is, an understanding of the nature of expertise also requires a description of what it is that experts are so good at doing.

A number of studies have identified various decision-making strategies and approaches as discussed in chapters 2 and 3. In this chapter, efforts to model the scoring process more generally are reviewed. In addition, this chapter describes another possible way of modeling for the scoring process taken from experimental psychology, and explores some of the ways such an approach might inform research into rater expertise.

4.1 Models of the Decision-Making Process from Writing Assessment

While a number of studies of rater cognition within writing assessment contexts have described strategies used by raters when making scoring decisions (e.g., Cumming, 1990; Cumming, Kantor, & Powers, 2002; Vaughan, 1991), relatively few attempts have been made to model the process as a whole. Such models are important for understanding rater decision making because they describe how different decision-making strategies interact and can therefore be used to make predictions regarding the effects of rater background or training on scoring, making it possible to design studies in a more principled way. This discussion focuses on models developed within writing assessment
because it appears that no attempts have yet been made to produce similar models of rater
decision making in speaking assessment. However, this approach is somewhat
problematic because the judgment of written and oral language performance represent
very different cognitive and practical challenges, so models of judgment developed
within writing assessment cannot be assumed to apply to the judgment of speaking as
well. On the other hand, such models are presented here in the view that they may
provide a useful starting point for thinking about how raters make scoring decisions in
speaking tests, and also represent how researchers in language assessment have
conceptualized the scoring process. This section will consider four published models of
rater decision making in writing assessment contexts.

An influential early model from L1 writing assessment was developed by
Freedman and Calfee (1983). This model was developed using the research of the time
(based on analysis of scores) as well as ideas from cognitive science, and viewed the
scoring process as consisting of three basic steps:

1. Read and comprehend the text to create a “text image.” The text image is made up of
   the rater’s perceptions of the text; each rater is expected to form a slightly
different text image.
2. Evaluate the text image and store impressions. This is where the rater makes various
   judgments about the text, and stores the results of these judgments as impressions.
3. Articulate an overall evaluation. This is the final step wherein a score is indicated and
   comments made, if any.

Freedman and Calfee suggested that the scoring process could be recursive in nature,
with cognition shifting back and forth across stages. In addition, the text image may be
stored in long-term memory, assumedly available to contribute to the evaluation of compositions encountered in the future.

Freedman and Calfee (1983) postulated that when forming a text image, similarities will outweigh differences for homogenous groups of experienced raters. In other words, one effect of rater background or experience would be to reduce the variability in text image created by raters while scoring. Although not stated explicitly, it follows that raters with similar internal representations of the text should produce similar scores. Also, it might be expected that experience or training could modify the evaluation of the text image by (a) influencing the selection of text features the raters attend to, and/or (b) changing the severity with which different features are judged, and/or (c) adjusting the relative weight placed on different language features.

Wolfe (Wolfe, 1995, Wolfe, 1997; Wolfe, 2006; Wolfe, Kao, & Ranney, 1998) proposed an adapted version of Freedman and Calfee's model. Wolfe's model specifically includes a *framework for scoring* (essentially the scoring processes described by Freedman & Calfee, 1983) as well as a *framework for writing*, which takes into account features of performance valued in the testing context that constitute the milieu for evaluative decisions. Based on his work with more- and less-proficient raters (with proficiency defined in terms of inter-rater agreement, see section 3.2.1), Wolfe suggested that less-proficient raters, who tended to employ a read-monitor-read sequence, may be distracted from identifying the overall organization of ideas in the text and thus construct a text image that fails to capture important qualities of the original document. In contrast, more-proficient raters tended to withhold judgment until having read and considered the text as a whole, allowing for the creation of a more robust text image. One would assume
that scores based on such a text image would be more representative of the overall qualities of the text and should therefore be less affected by idiosyncrasies in rater perceptions, leading to improved inter-rater agreement.

It should be noted that Freedman and Calfee’s (1983) model had a strong cognitive focus and was originally developed to frame questions for future research. A later model, focused on more pragmatic uses, was proposed by Milanovic, Saville, and Shen (1996; Figure 4.1). This model was based on the findings of Cumming (1990) and developed through a series of self-report studies involving the Cambridge main suite English language examinations (the FCE, CAE, and CPE). Milanovic et al.’s model was intended to capture both the decision-making process as well as the textual features attended to at each stage of the process. Interestingly, one purpose for developing the model was to guide rater training in the authors’ various research studies, so to some extent the model may have been prescriptive as well as descriptive.

In its rough outline, Milanovic et al.'s model follows a sequence similar to that proposed by Freedman and Calfee as well as Wolfe. In the earlier steps, the composition is read and a general impression developed. In later steps judgments are made regarding the text; this may involve a variety of behaviors such as re-reading the text, checking the scoring rubric, and the like. In the final phase, provisional scores are confirmed or revised. Where Milanovic et al.'s model diverges from that of Freedman and Calfee's is in the fact that it identifies the steps of the process in a more specific, concrete way. Although this model is based on empirical data, it is unclear whether it captures the full range of processes used by raters. For example, Wolfe's (1997) observation, cited above,
that less-proficient raters engage in a sequence of partial reading followed by an
evaluation, would fit poorly, if at all, in Milanovic et al.'s model. On the other hand, the
model was created in consultation with experienced raters and so can be viewed as
representing what proficient raters probably do when scoring. In this case, novice or less-
proficient raters might be expected to deviate from this process, and one measure of
expertise could then be the extent to which raters follow the “expert” sequence.

Experience or training might also be expected to influence specific aspects of the
process described by Milanovic et al. For example, training and/or experience will almost
certainly contribute to the internalization of the marking scheme and interpretation of test
tasks, something the model lists as a first step in the process (i.e., pre-marking). Another
place of potential influence is the “rate” step, where experience or training may influence
how different language features are perceived and evaluated. Finally, expert raters might
be expected to avoid the “modify” step as much as possible, and instead put more effort
into their initial evaluation, thus avoiding inefficient backtracking during the scoring
process (Feltovich, Pietula, & Ericsson, 2006).

Perhaps the most fully developed model to date of the process by which raters
evaluate writing was developed by Lumley (2002, 2005). Lumley worked with think-
aloud data produced by four experienced raters scoring an established Australian adult
migrant English test, so this model also depicts accomplished performance of the scoring
task. The model postulates a three-step scoring process, rather similar to the stages
proposed by earlier models (Table 4.1).
Table 4.1

The stages of the scoring process (Lumley, 2005)

1. First Reading
   - Read the text and develop an overall impression, which may be based on local or global features.

2. Scoring
   - Re-read the text or referring to the scoring rubric as needed.
   - Articulate and justify an initial scoring decision.

3. Conclusion
   - Confirm the initial score or revise.

---

Figure 4.2. The stages of the scoring process. From Assessing second language writing: The rater’s perspective, by T. Lumley, 2005, p. 291. Copyright 2005 by Peter Lang GmbH.
In addition to the steps of the scoring process listed in Table 4.1, Lumley’s model includes the scoring context as a second dimension, with three levels that overlay the scoring process: an institutional level, an interpretation level, and an instrumental level (Figure 4.2). The institutional level is the institutional context of the test, while the instrumental level is the interaction between the individual rater and institutional requirements. The third, instrumental level describes the actions taken by raters while scoring, and would be the part of the process where rater expertise would most likely be realized. At the instrumental level experience or training might be expected to influence the type of general impression formed from a particular text in stage 1 of scoring, the specific language features that are referenced in stage 2, and the process used to confirm or revise scores in stage 3.

It should be noted that Lumley’s model includes many factors that come into play at different stages of the scoring process and different levels of the scoring context. For example, for the activity “justification and articulation of scoring decisions” (the second stage of scoring within the instrumental level of context) Lumley suggests that there are at least seven different factors that may come into play, including the scoring rubric, the rater’s interpretation of the requirements of the task, other language features not listed in the rubric, fatigue or affect, and the possible consequences for test takers. Moreover, these factors may interact to result in indecision; this indecision may then be resolved through any of a number of different strategies to finally arrive at a provisional score. Perhaps the most useful lesson to be taken from Lumley’s model is that scoring within language performance tests is highly context-dependent, may involve a great many factors, and is likely to be a complex and perhaps messy process. Within such a process,
rater expertise might be realized in many different ways, and there may in fact be no single or simple description of what constitutes expertise in scoring language performance.

Finally, while the models of the scoring process described in the this section are useful tools for conceptualizing that process and suggesting avenues for research, they are ultimately limited to the specific contexts in which they were constructed, that is, writing assessment. As mentioned earlier, these models provide a helpful starting point for thinking about rater decision making in speaking assessment, but they cannot simply be assumed to apply to judgments of oral language given the considerable differences in how raters interact with examinee responses produced in the two modalities. In the hopes of extending current thinking regarding rater cognition, the following section discusses a different approach to the study of judgment taken from the field of experimental psychology, and considers the potential application of this approach to understanding rater expertise.

4.2 Potential Approaches from Experimental Psychology

Studies of rater decision-making within language assessment contexts, or indeed educational assessment in general, represent only a small portion of the research effort that has been directed at the issue of human judgment. In a recent review, Brooks (2012) noted that although there are established traditions of judgment research in fields such as medicine, business management, and economics, as well as a variety of established theoretical frameworks in psychology, these traditions have had little impact on educational assessment. Of particular importance for language assessment studies is the
potential for psychology and allied fields to offer new models for exploring rater
cognition, as well as methodologies that go beyond the limitation of conscious thoughts
accessible through verbal report. This discussion will focus on approaches tied to
psychological studies of magnitude judgments of natural phenomena (i.e.,
psychophysics), based on the view that the rater’s task in language assessment may be
usefully interpreted as a similar form of magnitude judgment (Laming, 2003, 2004; Suto
& Greatorex, 2008).

Human perception of the magnitude of physical stimuli (e.g., length, weight,
loudness, brightness) has been of considerable interest to psychologists for well over 100
years (see Laming, 1997 for a review). Recently, an important contribution to this area
has been Laming’s (1984, 1997) argument that such judgments are fundamentally relative
in nature, whereas previously, magnitude judgment had been seen in terms of an absolute
mapping of a physical stimulus to a resulting perception. More recently, the notion that
human judgment is essentially relative in nature has also been used as the basis for a
theory of decision-making proposed by Stewart, Brown, and Chater (2006), termed
Decision by Sampling (DbS). Decision by Sampling postulates that judgments of
magnitude are “constructed from a series of binary, ordinal comparisons to a sample of
attribute values drawn from memory and is its rank within the sample” (Stewart, et al.,
2006, p. 1). That is, judgments are made by making pairwise comparisons of the item at
hand with other items available in the immediate surroundings or stored in memory.
Moreover, magnitude is based on the cumulative rank of the item compared to the other
items available for comparison, not on some absolute scale:
We do not assume that people have stable, long-term internal scales along which they represent value, probability, temporal duration, or any other magnitudes. Instead, we assume that people can only sample items from memory and then judge whether a target value is larger or smaller than these items. (p. 2)

Cognitively, the theory assumes only two simple abilities: the ability to make ordinal comparisons and the ability to accumulate frequency distributions of items/phenomena. This position is quite different from the more intuitive view that judgments are made by applying some sort of internalized measuring stick, standing in contrast to Stevens' (1946) oft-cited definition of measurement as the application of a rule to an observation.

Decision by Sampling and other similar relative judgment models were developed to explain a variety of economic behaviors, but may provide insight into judgments of language ability within performance assessment as well. For example, DbS predicts that when there are many examples available in a certain region of a magnitude continuum (i.e., a “dense” part of the distribution) small changes in magnitude will be given greater weight (or be detectable with more precision) because the examples available for comparison are more closely spaced. Stewart et al. (2006) cite the example of debits from people’s checking accounts (i.e., outlays such as checks written or debit card purchases): small debits are highly frequent, while large debits are more rare (figure 4.3a). If debits are ranked cumulatively (Figure 4.3b) the relative rank of a debit for a given amount of money changes very quickly in the left-hand section of the curve, where a relatively small difference between debits, perhaps £20 in the example, is readily noticeable. On the other hand, the portion of the curve to the right represents relatively infrequent large debits, and the cumulative rank of a given debit changes only slowly with increasing value. Here, a difference of £20 has much less psychological impact. In concrete terms, a
difference in price between £10 and £30 (e.g., for a meal) is expected to have greater psychological impact than the difference between £1,480 and £1,500 (e.g., for a new computer).

Figure 4.3. Debits made to individual accounts at a major UK bank over one year: (a) raw frequency of debits of various amounts; (b) cumulative frequency (ranks) of the same debits. Adapted from “Decision by sampling,” by N. Stewart, N. Chater, & G. D. A. Brown, 2006, *Cognitive Psychology, 53*, pp. 4-5. Copyright 2006 by Elsevier, Inc. Used with permission.)
Figure 4.4. Distribution and cumulative ranking of scores for the TOEFL iBT speaking test: (a) distribution of scores; (b) rankings of scores (raw data from the TOEFL Public Use Dataset, $N = 1000$, Educational Testing Service, 2008).

Now consider a language performance testing situation where language ability (however defined) is normally distributed (Figure 4.4a). If this normal distribution of ability is converted to a distribution of cumulative rankings, then a sigmoid curve results (Figure 4.4b). Assuming this score distribution is mirrored in peoples’ internal distribution of language performance, one should expect that in the steep central section of the sigmoid curve, small differences in ability will carry more psychological weight (or be more noticeable or more detectable) than the same amount of difference in ability seen at the relatively flat ends of the curve. Within a performance test, this situation could be interpreted in terms of discrimination between test takers: discrimination is highest between individuals in the center of the distribution (or sigmoid curve) and poor for those at the ends. In a sense, none of this is new. Sigmoid probability curves are a staple of test analyses based in Item Response Theory, where such curves (termed item characteristic curves, or ICCs) are used to analyze item discrimination (Bachman, 2004). What the
theory of Decision by Sampling provides, however, is a proposed psychological mechanism that produces a similar effect in rater decision making.

Within the realm of psychophysics, a variety of experimental results have been cited to support the view that judgment is based on comparisons. Laming (1997) discusses this body of work in detail, and two examples are provided here. Firstly, if judgment is relative in nature, then it should be easier to discriminate between performances presented in pairs as opposed to judging the absolute magnitude of performances presented in isolation. Within the domain of psychophysics, a common finding is that pairwise comparisons are much more precise than absolute judgments; that is, when two stimuli are presented alongside each other or in quick succession, much smaller differences can be discriminated (Laming, 1997, 2004). For example, Pollack (1952) examined the perception of frequency of tones and found that participants could distinguish no more than about five categories when the tones, spread across a range of 100-8000hz, were judged in isolation (separated by a pause of 25 seconds). But, when the same range of tones are separated by only 0.5s, effectively allowing one tone to be contrasted with the previous one, there are about 1000 differences in frequency that can be detected at least 70% of the time (Wier, Jesteadt, & Green, 1977). The general finding that pairwise judgments are more precise has been one of the foundational pieces of evidence used in support of the position that psychophysical judgments are essentially relative in nature (Laming 1997, 2004; Stewart, Brown, & Chater, 2005).

Secondly, if judgment is relative, then when judgments are made in succession the earlier judgments may be expected to serve as a point of reference for the later judgments (anchor and adjustment bias; Plous, 1993). This results in each judgment being correlated
with the previous judgment (or several judgments) in the sequence. Within the psychophysics literature, a variety of studies have observed such sequence effects, where judgments have been found to be correlated with judgments made previously (e.g., Jesteadt, Luce, & Green, 1977; Ward, 1979). A similar effect has also been reported for the more complex social judgment required to estimate the price of an item, where price judgments were positively, though weakly, correlated with the previous judgment in approximately 40% of individuals (Matthews & Stewart, 2009).

As mentioned earlier, the view that magnitude judgments are ultimately relative was developed on the basis of experimental studies of the perception of physical phenomena. Although these ideas have recently been translated to the field of behavioral economics, it remains an open question as to whether the relative judgment approach is relevant to complex social judgments such as those made in the scoring of language performance tests. However, there is limited evidence that a relative judgment view may apply to language performance tests as well. For example, a number of studies of L1 writing assessment have observed sequence effects in scores awarded to compositions (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes, Keeling, & Tuck, 1980).

In addition, studies of raters' decision making while scoring L2 English have observed raters making explicit comparisons between one or more previous compositions (Milanovic et al., 1996; Sakyi, 2000). Cumming et al. (2002), in their taxonomy of rater decision-making strategies, even included the item “decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgments collectively” (p. 53, italics added), acknowledging explicit comparison with
other essays as a judging strategy. Similarly, within a speaking test context, Meiron (1998) identified a strategy she described as “comparison of examinee sample to other examinee samples / comparison of examinee sample to other non-native speakers” (p. 47). Also, Wolfe's model of essay rating discussed in the previous section includes a “compare” element in the “justification” step of scoring (Figure 4.2), although in Wolfe's model comparisons happen at the end of the scoring process, after a provisional scoring decision has already been made. Although the conscious comparisons described here do not necessarily prove that raters' cognitive processes are based on comparisons between examples, they at least demonstrate that raters can, and do, compare examinees while scoring.

At the start of this section it was suggested that approaches from experimental psychology might prove useful in the study of rater decision making within language tests. To close, I briefly mention a few such potential uses. First, Stewart, Brown, and Chater’s (2006) theory of Decision by Sampling, as well as the relative judgment view more generally, propose a mechanism for the basic psychology of the judgment process which can be used to generate testable hypotheses; examples of a few such hypotheses were given earlier. The current schematic models of language assessment rater cognition, while useful for conceptualizing the decision process, do not so easily lend themselves to testable hypotheses. Relative judgment models also suggest a particular definition of rater expertise. Specifically, expertise consists of thorough experience with instances of performance similar to the ones being judged (i.e., expert raters have a thick internal distribution of comparable items). This definition in turn leads to several predictions. First, “expert” raters should be more reliable in scoring then those with less experience.
As raters gain expertise by encountering more examples of examinee responses, each new example acts as an additional notch in a scale with increasingly finer gradations, leading to an increased ability to detect differences between performances and accordingly achieve more consistency in scores. Another prediction would be that rater training activities will have the most effect on scores when they contribute to the construction of a robust internal distribution of comparable examples; for instance, exposure to many examples of examinee responses (with feedback provided to confirm the relative quality of different performances) might be expected to be effective in developing scoring consistency. Similarly, scoring consistency and accuracy should be improved by the use of scoring aids that provide a standard against which examinee performance can be compared. While descriptors in a scoring rubric might fulfill this function, actual benchmark examples are more directly comparable to examinee responses and so might be even more effective.

These predictions are currently nothing more than speculation. Nevertheless, the point is that the relative judgment view is capable of generating a variety of testable hypotheses, which may guide research into rater cognition in a principled way. Much of the existing research in rater variability and expertise is largely atheoretical in nature, and the relative judgment view may provide a useful organizing framework for planning and understanding future work in this area.
CHAPTER 5

Research Questions and Method

As outlined in the introduction, the broad goal of this study was to better understand the nature of rater expertise in judging oral language performance. This effort was divided into three parts, with the first being an examination of the influence of training and experience on rater scoring patterns and scoring behavior. This part of the study employed a longitudinal design where participants scored audio-recorded oral responses on four separate dates, once prior to training and three times afterwards. The goal of this effort was to illuminate the contribution of training and experience to the development of accurate and consistent scoring, thus providing insight into how scoring expertise develops.

The second part of the study focused on a subset of raters identified as showing high accuracy and consistency in scoring (“proficient raters) or lower accuracy and consistency (“less-proficient” and “developing” raters). The goal of this section was to document differences in behavior and cognition associated with higher or lower expertise in scoring. Scoring behavior across time was compared between the groups, as well as rater cognition during scoring collected via verbal reports. These qualitative data were intended to clarify differences in behavior and cognition that might contribute to desirable scoring patterns.

The third and final part of the study was focused on the evaluation of a possible cognitive model for understanding the basic nature of magnitude judgments of language ability, as well as the underlying psychological nature of expertise in making such judgments. This work was informed by recent theory from the domain of psychophysics
(Laming, 1997) and behavioral economics (Stewart, Chater, & Brown, 2006) where it has been claimed that magnitude judgment is fundamentally relative; that is, judgments are made by comparing the current item to other similar items. Part three of the study examined two predictions that follow from this relative judgment view. The first prediction is that other examples in the immediate environment may inform scoring judgments, leading to a sequence effect (or more formally, an anchor and adjustment bias) where the current score is influenced by the previous score. The second prediction is that items can be discriminated more precisely when presented side-by-side, compared to when they are presented in isolation.

This chapter begins by stating the research questions examined in the study, followed by a description of the features common to the entire study: the study context, participants, and general materials. Finally, instruments, procedures, and analyses are described for each research question in turn.

5.1 Research Questions

Within the overall theme of describing the nature of rater expertise in a speaking test context, the study examined the following research questions.

1. What effects do training and experience have on scoring patterns and scoring behavior of raters?
   a. In what ways do rater severity and internal consistency change with training and experience?
   b. In what ways does rater accuracy, defined as agreement with previously established reference scores, change with training and experience?
c. In what ways does rater scoring behavior change with training and experience? Specifically, how does use of scoring aids and the time taken to reach a decision change?

2. What features of scoring behavior and cognition distinguish more-proficient and less-proficient raters?
   a. In what ways is scoring behavior different between more- and less-proficient raters? Specifically, how do use of scoring aids and the time taken to reach a decision differ?
   b. In what ways does rater cognition differ between more- and less-proficient raters? Specifically, what differences, if any, are seen in attention paid to distinct language features and in explicit attention given to the scoring process?

3. Can rater decision making and expertise be understood in terms of a relative view of magnitude judgment?
   a. Is a sequence effect observable in scores produced by novice raters in a speaking test context? If so, does the size or prevalence of the sequence effect decrease with training and experience?
   b. Is the ability of raters to discriminate between responses of differing quality improved when responses are presented together, compared to scoring responses in isolation?
5.2 Context of Scoring

The study was aimed at a general investigation of rater behavior and cognition and did not attempt to address a specific operational speaking test. However, for a number of practical reasons, materials based on the TOEFL iBT Speaking Test (TST) were used in the study, which included responses from the TOEFL Public Use Dataset (Educational Testing Service, 2007), the TST scoring rubric (adapted to a modified scale as described below), and a limited number of scoring rationales provided by ETS and used for rater training. The advantages of using materials from the TOEFL iBT Speaking Test were as follows:

1. The TOEFL Public Use Dataset was available to researchers, which provided a large number of recorded speaking assessment responses that could be presented to raters.

2. The spoken examinee responses are monologic and thus remove the interlocutor as a complicating factor (although potentially an important complicating factor for aspects of the speaking proficiency construct).

3. The TST materials were easily adaptable to data collection online, allowing recruitment of raters in any location. This factor was important because it was felt that recruitment of enough qualified participants in the local area would be difficult given the time demands required for participation in the study.

4. The presence of an existing network of ETS-trained and experienced raters made it possible to obtain additional scores for recorded responses contained in the Public Use Dataset. This was important for producing reference scores which were used as a baseline for comparison with the ratings of study participants, and which
were also needed for setting up the experimental study of pairwise versus isolated judgments in research question 3.

It should be noted that TST materials were used for the study because of the advantages stated above. The study was neither designed nor intended to make specific claims regarding the reliability or validity of TST scores.

The TST includes a speaking section with six items: two “independent” items where a response is given to a simple prompt, and four “integrated” items where the examinee first listens to a recorded conversation or lecture, or reads a text and then listens, and finally answers a question regarding the written and spoken materials (Appendix A). This study used only independent items because: (a) the recorded responses were somewhat shorter, reducing rater time commitment; (b) making judgments does not require reference to verbal and/or written academic materials included as part of the item prompt, simplifying the judgment process; and (c) different independent tasks tend to be very similar in character, which increases confidence that the two items are interchangeable for scoring purposes. When responding to an independent prompt, the examinee has 15 seconds to prepare and then 45 seconds to record an answer. Under operational testing, the recorded answers are later rated online by a dispersed network of ETS raters using a holistic 0-4 point scale that operationalizes the TOEFL speaking ability construct. While deciding on a holistic score, raters are also asked to consider performance in three domains: delivery (pronunciation, intonation, fluency), language use (grammar and vocabulary), and topic development (detail and coherence of content; Educational Testing Service, 2007).
5.3 Participants

The participants were 20 experienced teachers of English who had not previously worked as scorers for the TOEFL speaking test. Raters were recruited who had two years or more experience in teaching English (average 7.3 years). Fifteen of the raters had experience teaching at the college level, while the other five had taught high school students, adult classes, or a mixture of both. All raters reported having taught learners at intermediate proficiency level and above, similar to the TOEFL candidature.

Given that one focus of the study was to collect rater reflections on their scoring criteria and processes, it was felt that experienced teachers were more likely to be more consciously aware of the language features produced by test takers and therefore might be better able to verbalize their impressions and thought processes. Because the study examined how rater behavior and cognition change with experience, previous experience scoring oral language performance assessment was documented. Six people reported having occasionally scored local speaking tests, one person had worked one day as an interlocutor/scorer in a oral test administration as part of a research study, and one participant had previously scored the Cambridge FCE and CAE tests approximately one day a month, but had scored only 4-5 times in the previous year. So, in terms of the research questions, the participants were considered to be “inexperienced” in scoring speaking within the TST or similar context, although they did have English teaching or other relevant professional experience which may have contributed to expertise in scoring.
Recruitment was also limited to native/primary speakers of English and all raters reported having English as their first and/or strongest language. Existing research on native- versus non-native (NS vs. NNS) speaker raters has documented a variety of differences in scoring behavior and/or language features noticed (e.g., Barnwell, 1989; Hill, 1996), although these studies generally examined untrained raters. In studies of speaking tests where NNS speakers underwent rater training (e.g., Brown, 1995; Kim, 2009b; Xi & Mollaun, 2009), less difference was seen in scores, but in any case, conducting an NS/NNS comparison was not an objective of the study and it was felt that including NNS speakers might complicate the interpretation of findings.

As a practical matter, raters also needed to have the technological capability to access the materials, record their voices, and upload files to the online document management site used for the study. Participants’ ability to navigate computer-based tasks was monitored during an orientation session and training was given in how to carry out the specific tasks required for the study. In addition, technological requirements were kept as simple as possible by creating study materials that were usable across computer platforms and allowed participants to use whatever software they were most comfortable with for such tasks as listening to audio files or voice recording. Nonetheless, two people were excluded from the study after it became apparent that they did not have the computer skills to successfully carry out the study tasks; additional participants were recruited to replace them.

Recruitment of raters was initially through the personal contacts of the author, but also included announcements sent to college-level English for academic purposes (EAP)

---

1 One individual reported Hawai'i Creole English as dominant with English as next dominant.
programs and graduate programs in TESOL on the island of Oahu. Three individuals were also recruited who were living outside of Hawaii. During recruiting, the time requirement was made clear to each contact (approximately 32 hours over a period of 4-6 weeks); this time commitment made recruiting somewhat of a challenge, and a period of about 3 months was required to complete the recruiting process (September-December 2011). Ultimately a total of 25 people were recruited; of these, two people were excluded for difficulty with the technology as previously mentioned, and another three individuals dropped out early in the study due to the time commitment. To encourage participation and compensate people for their time $480 was paid to each participant, equivalent to $15/hr based on the estimated time to complete the work. A total of 10 women and 10 men completed the study; further information regarding the raters is given in Table 5.1.
Table 5.1

*Academic, professional, and testing backgrounds of participants (N=20)*

<table>
<thead>
<tr>
<th>Item</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational background</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TESL Certificate</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>M.A. TESOL or similar degree</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Other M.A. <em>a</em></td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Ph.D</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Are you currently a student? <em>b</em></td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>Teaching experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How many years have you taught ESL/EFL?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. = 7.3 (range 2-15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have you taught ESL/EFL learners in the last 12 months?</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Testing experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have you ever worked as a TOEFL iBT speaking test rater?</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Have you ever taught a TOEFL preparation class (of any kind)?</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Have you helped students prepare for the TOEFL Speaking test?</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Have you worked as a rater in a speaking test? <em>c</em></td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Have you taken any courses in language testing?</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

*a* Majors listed (number of responses)
Asian Studies (1)
Conflict Resolution (1)
English Composition (1)
International Education (1)

*b* Academic programs of current students
M.A. Second Language Studies (10)
Ph.D Second Language Studies (2)
M.A. Urban & Regional Planning (1)
Ph.D in Education (1)

*c* Testing contexts mentioned.
Local/placement test (6)
Cambridge FCE, CAE (1)
Speaking test research - 1 day as an interlocutor/rater for a mock EIKEN test administration (1)
5.4 Materials

5.4.1 TOEFL Speaking Test responses. As mentioned earlier, raters scored examinee responses taken from the TOEFL Public Use Dataset. The dataset contains scores, recordings, and test taker demographic data for 240 individuals from each of two forms of the test (480 test takers in all). For the current study, responses on items 1 and 2 (the two independent items) of Form 1 of the test were used, to make a total of 480 responses from 240 test takers. Responses from Form 1 were used because this data set includes a larger number of low-scoring responses compared to Form 2, providing a slightly more even distribution of responses across proficiency levels (Table 5.2). For item 1, the prompt was “Students work hard but they also need to relax. What do you think is the best way for a student to relax after working hard? Explain why.” (henceforth referred to as prompt A). The prompt for item 2 was “Some people think it is alright to stay up late at night and sleep late in the morning. Others think it is better to go to bed early at night and wake up early. Which view do you agree with? Explain why.” (henceforth referred to as prompt B).

Table 5.2

*Score distributions of speaking responses in the TOEFL Public Use Dataset*

<table>
<thead>
<tr>
<th>Score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>14</td>
<td>105</td>
<td>98</td>
<td>23</td>
<td>240</td>
</tr>
<tr>
<td>Item 2</td>
<td>11</td>
<td>91</td>
<td>105</td>
<td>33</td>
<td>240</td>
</tr>
</tbody>
</table>
Individual audio files containing each response were converted from the speex\(^2\) (.spx) format used in the TOEFL Public Use Dataset to mp3 format for increased compatibility with common audio players. Speex files were first converted to .wav format (48 kHz, 16 bit) using JetAudio Player (ver. 8; JetAudio 2007) and then further converted to .mp3 format (128 kbps) using Audacity (ver. 1.3). Single scores produced by ETS raters were available for each response, along with demographic data for each test taker, including native language, country of origin, age, and gender for all individuals. In addition, data regarding test-taker plans and English language background was available for some individuals.

5.4.2 TOEFL Speaking Test scoring rubric. Examinee responses were scored using the criteria listed on the TST rubric for independent items (Appendix B). The operational TST is scored using a four-point scale (1-4); this approach was adapted to a six-point scale (1-6) for use in the study. This measure was taken to make the scoring task more challenging, with the intention to make it less likely that raters would show high levels of agreement or consistency at the beginning of the study, leaving room for improvement following training and experience. The existing TOEFL scale was adapted by adding half points to make a seven-point scale (1, 1.5, 2, 2.5, 3, 3.5, 4), and then combining the lowest two scoring categories (due to very few responses in the lower part of the scale) and re-labeling the result to give a scale of 1-6. Accordingly, TOEFL scores of 2, 3, and 4 corresponded to 2, 4, and 6 in the new scale. The TST scoring criteria were

\(^2\)The speex format is an open-source compressed audio format specifically designed for voice recording, which produces telephone-quality audio (Xiph.org, 2006).
unchanged, with intermediate scores in the new scale assigned for responses that combined features of different levels or were otherwise intermediate in quality.

5.4.3 Reference scores. In addition to the single scores provided in the TOEFL Public Use Data Set, additional scores for the 480 responses used in the study were obtained from 11 operational TOEFL Speaking Test scoring leaders.\(^3\) These additional scores were used to produce reference scores that would both incorporate the revised rating scale used for the study as well as provide measures that were relatively accurate and precise. As mentioned earlier, such reference scores provided a standard for comparison with scores generated by raters in the study, and they also provided baseline values that were used to construct pairings of examinee responses for investigating the precision of isolated and relative judgments (see research question 3 below).

The eleven individuals recruited to provide reference scores each had several years of experience scoring the operational TST, and, in their role as scoring leaders, monitored and mentored the work of a team of 8-10 raters during operational TST scoring sessions (Pam Mollaun, personal communication, November 5, 2010). The group had an average of 3.7 years experience as scoring leaders (range 2.8-6.3 years) and at the time of scoring reported working an average of 80 hours per month on operational TOEFL Speaking Test scoring. In addition, the individuals had an average of 12.0 years of English teaching experience (range 5-20 years), and eight of the scoring leaders had earned M.A. degrees in TESOL or similar domains. Other details regarding the background of this group are given in Appendix C.

---

\(^3\)I thank Pam Mollaun at ETS for her assistance in recruiting these individuals.
Scoring by the reference group was done in a series of six sessions conducted during a scheduled break in TOEFL operational scoring in late December 2010 and early January 2011. Each scoring session consisted of 100 responses each, including 78 unique responses, with an additional 12 between-session repeats and 10 within-session repeats used to examine intra-rater consistency. The scoring materials were made available on a self-access basis (described below) and raters did one or two sessions a day, over a period of 3-10 days, according to their own schedules. Completing one or two sessions a day (100-200 responses) was similar to the workload required for TOEFL operational scoring (Pam Mollaun, personal communication, November 5, 2010).

Scoring was done using interactive Adobe Acrobat pdf forms as described in section 5.4.6 below. Scoring leaders were first directed to review the TOEFL scoring rubric and reminded to use half-scores whenever the response included features of performance from two different score levels or was otherwise intermediate between score levels. Aside from the rubric, no other rater calibration activities were used. Review of the rubric was followed by scoring, with the first 50 responses taken from item 1 of the TOEFL Public Use Dataset (Task A) and the second 50 responses taken from item 2 (Task B). Each rater received a different randomly selected set of responses in each session (except for the between-session repeats, which were the same for all sessions), and examinee responses were presented in random order, except for within-session repeats which were presented at the end.

Raw scores were processed using FACETS Rasch analysis software (ver. 3.62.0; Linacre, 2007b) to produce a reference score for each response. Specifically, fair-M averages (the average score, in original units, adjusted for differences in rater severity;
Linacre, 2007a) were used in the study. FACETS reports fair-M averages to one decimal point; the averages were used as-is without rounding to whole scores. The rulers and rater measures for the Rasch analysis are given in Appendix D. One rater (REF11) was misfitting (infit mean square = 1.44) indicating relatively poor internal scoring consistency, but a separate analysis excluding this rater found no discernible difference in fair-M scores.

5.4.4 Online data collection setup. It was anticipated that individuals located outside of the local area would need to be recruited for the study. Accordingly, a dedicated Google Apps domain was established for the purpose of providing raters with remote access to materials (instructions, recordings, data collection instruments, etc.), and giving raters a means to submit scores, recordings, and other data. Each participant was provided with a password-protected account through which individualized data collection instruments could be shared with the participant, and the participant could upload and share completed instruments with the researcher. This setup also made it possible to manage all outgoing and incoming instruments from a single administrator account, greatly simplifying data management. Audio files, which needed to be accessed by all participants, were located on a secure website within the same Google Apps domain. Routine communication was conducted via email, using whatever account the participant preferred, and the researcher was also available via phone as needed.

For data collection, participants used their own computers. All participants were required to have the Adobe reader software which was used to view the type of interactive pdfs employed in the study. For other aspects of data collection, raters were allowed to use whatever software they felt most comfortable with, but were required to
have a web browser, an audio player, and an audio recorder. Five of the participants had computers that were unable to play audio files and record voice at the same time (necessary for collecting verbal reports), so these individuals used separate digital recorders for voice recording and were loaned a digital recorder if needed. Audio files were generally played over the computer speakers; this step was necessary for the collection of verbal reports where it was felt important that the recording of the report contain both the rater's comments as well as the examinee response, so that the rater comments could be understood in context.

5.4.5 Language and Professional Background Questionnaires. A questionnaire was used to collect data on participants' educational, teaching, or other professional experiences that might be relevant to the scoring of speaking tests (Appendix E). The data collected included the degrees/qualifications held, information regarding teaching experience, and the participant's experience with the TOEFL and with speaking tests in general. In addition to professional background, a modified version of the LEAP-Q language background questionnaire was used to document rater language experiences (Marian, Blumenfeld, & Kaushanskaya, 2007; Appendix F). Both questionnaires were presented as interactive pdf documents, constructed in the same manner as described for the scoring instrument.

5.4.6 Interactive data collection instruments. All data in the study were collected using interactive pdf files. This approach was taken for several reasons:

1. It allowed documents to be transmitted and completed electronically, which made it possible to recruit participants outside of the local area.
2. Adobe Acrobat documents are specifically designed to preserve content and formatting across a wide variety of computer platforms, avoiding compatibility issues that might have been encountered if using spreadsheets, html files, or other formats.

3. Data from such interactive pdf forms can be easily downloaded into a spreadsheet; batch processing capability makes possible rapid data transfer even when multiple files are involved.

These forms were constructed using Adobe LiveCycle Designer (ver. 8.0, 2006) and then output into a standard Adobe Acrobat pdf format. Constructing the form in Adobe Designer allowed for a much greater degree of control over appearance and functionality of the resulting forms, compared to the more limited form authoring tools available in Adobe Acrobat. In particular, JavaScript commands were incorporated for collecting data on rater behavior while scoring, such as the number of times a button was clicked, the time taken to complete certain actions, and the frequency that scoring aids were used. In addition, Adobe Designer forms are based on an underlying XML file; once a form template had been constructed, code from this file could be directly manipulated to quickly make large-scale changes in form data, saving large amounts of time when customizing forms for each rater. This feature made it possible to quickly change the presentation order of examinee responses so that each rater experienced an independently randomized set of responses. Overall, a total of 420 unique pdf forms were produced for the study, so efficiency in form creation and data management was at a premium.
5.5 Research Question 1: Effects of Training and Experience

The remainder of this chapter discusses the specific materials and procedures used to answer each of the major research questions. Each research question is discussed in turn, starting with the first.

5.5.1 Overview. The experiment was designed to provide longitudinal data on rater development over time. Following an orientation session, participants completed one scoring session, then a rater training session, and finally three more scoring sessions; the design is shown in Figure 5.1.

The orientation was provided to make sure raters were familiar with the procedures used in the study and to see if they had the basic hardware and computer skills necessary to successfully access and use the study materials. This step was followed 2-4 days later by a scoring session where raters were asked to score without the use of a rubric (although examples of performance at each score level were provided). The purpose of this first session was to document the raters’ ‘native’ scoring patterns and behavior to serve as a basis for identifying changes resulting from training and experience. This step was followed 2-4 days later by a rater training session, which was followed the next day by a second scoring session to document any immediate effects of training. Two more scoring sessions occurred at 5-7 day intervals to document scoring patterns and behavior as raters gained increasing experience.

Flexible intervals were used between sessions to give raters room to schedule their participation. Such flexibility was felt important because scoring sessions typically required approximately five hours to complete, and most raters already had busy schedules as teachers and students. Participants were strongly encouraged and reminded...
to complete sessions within the window of time provided, however, and with a handful of minor exceptions, participants completed the sessions on schedule.

**Figure 5.1.** Sampling design for the study.
5.5.2 Instruments and procedures.

5.5.2.1 Orientation. Each rater completed an individual face-to-face orientation with the researcher at the start of the study, generally the only time raters met with the researcher during their participation. As described earlier, the purpose of the session was to familiarize participants with the details of the study, including use of their Google Docs account and navigation of the data collection instruments. The orientation started with a discussion of the tasks, schedule, time commitment, and compensation for the study, followed by the signing of a consent form. Participants then logged into their own Google Docs account and downloaded the instruments for the orientation session. These included a practice scoring document, as well as the language and professional background surveys.

The practice scoring instrument consisted of a scoring sheet (as described below) with twelve exemplars (two for each point of the scale) and ten responses to be scored. This exercise was intended to familiarize the participant with navigation of the type of interactive forms used in the study. After the participants listened to the exemplars and scored the responses on their own, they met with the researcher to discuss the procedure for doing stimulated recall. This was followed by another ten responses, where the participant first gave a score and then listened to the response again while recording a verbal report of their thought process. Finally, the participant was instructed in how to upload files to their Google Docs account and then uploaded the completed practice instrument and audio files. If time was available, then the language and professional background surveys were also completed and uploaded; if not, the participant was asked to fill out and upload the survey on their own. In addition to the training provided in the
orientation session, a short user-manual describing procedures for using Google Docs and the interactive forms was provided for the participant's reference.

The orientation was conducted using whatever computer the participant planned to use for the study, allowing for troubleshooting of any hardware-related issues and for the installation of software if needed. In most cases, the orientation was done in a laboratory on the University of Hawaiʻi campus with the rater bringing a laptop to the session, but in two cases the researcher met the raters at their place of work. Three other raters were located outside of Hawaii, so in these cases the orientation was done online via Skype. Raters entered the study at varying times from early October to late December 2011, with the first rater completing the orientation on October 4th, and the last on December 28th.

5.5.2.2 Scoring sessions. For the scoring sessions raters worked from their own locations as described earlier, downloading three instruments for the session: one for scoring prompt A, another for scoring prompt B, and a short wrap-up survey to collect rater perceptions of their own performance. Examples of the scoring and wrap-up instruments can be found in Appendices G and H. Each scoring session included 100 responses in total to be scored, with an additional 20 responses used for verbal reports of rater cognition, which provided data for Research Question 2. The responses were evenly divided between prompts A and B, with each instrument containing 50 responses to be scored immediately followed by 10 responses for verbal report. Eight responses per prompt (16 total) were repeated across all sessions to allow for additional analyses of rater consistency over time. Each rater received the same set of responses for a given scoring session, with the presentation order of prompts (A or B) and responses within
each prompt randomized for each rater. Raters were instructed to take breaks as needed, although they were encouraged to keep breaks short while working through a prompt, and take a longer break between prompts if needed. The time required for each prompt (including recalls) was approximately 2.5 hours, with the whole scoring session taking approximately 5 hours.

When scoring a prompt the raters were directed to review the scoring rubric and required to open a copy of the rubric to continue (except for session 1). Opening of the rubric then unlocked the first of six exemplars, one for each point of the scale (Figure 5.2). The exemplars were chosen on the basis of scores produced by the ETS scoring leaders; responses were chosen which had reference scores close to center of the score band as well as high levels of rater agreement (as determined by both low infit values as well as inspection of the spread of raw scores). Clicking a “play” button for the exemplar caused the associated audio file to play, recorded the time the button was clicked, and advanced a hidden click counter by a value of one. Raters were required to listen to each exemplar once, in order of lowest to highest score, until all the exemplars had been reviewed (clicking the play button re-locked the current exemplar while unlocking the next). Once the last exemplar had been reviewed, all the exemplars were unlocked and could be freely reviewed thereafter.
Figure 5.2. The first page of the scoring instrument showing the review of exemplars in progress. The exemplar for “Score = 2” is available for review, and the exemplar for “Score = 1” has been checked and re-locked.
Figure 5.3. The first page of the scoring instrument showing scoring in progress. All exemplars are unlocked and examinee response number 5 is being scored.
Playing of the final exemplar unlocked the first examinee response to be scored. A button to play the response was provided along with a drop-down list from which they could select a score on a scale of 1-6 (Figure 5.3). For scoring, raters were given an additional choice of “N/A” to be used if technical problems such as poor sound quality made it impossible to score the response. Selecting a score then unlocked a comment box and an “OK” button to confirm the score. Raters were allowed to play the response and/or alter their score as many times as they liked, but once the OK button was clicked the next response was unlocked and the current response was re-locked, preventing any further review or alteration. This provision was included to preserve any scoring sequence effects in the data, investigated as part of research question 3.

The design of the scoring interface was intended to record both the scores awarded for the responses and information describing the way the rater was interacting with the form. Hidden data fields recorded the number of times the response was played and the score was modified, along with a timestamp for each instance when a button was clicked or the score was changed. The comment box was provided for the raters to record any technological or other problems that they felt compromised their ability to score the response, and was usually left blank. (Raters were instructed to generally avoid writing comments regarding the quality of the response or the scoring process, since their perceptions would be recorded later via verbal reports.)

Following scoring, raters completed a short wrap-up survey where they were asked to judge their own scoring performance in terms of consistency, accuracy (how well their scores matched the scale indicated by the exemplars), and the degree of difficulty they experienced in making scoring decisions. Raters were also asked to
comment on their degree of use of the scoring rubric and exemplars during the scoring session, and they were invited to leave additional written comments.

5.5.2.3 Rater training. The rater training protocol used in the study was based on the use of scoring benchmarks and a calibration exercise (Figure 5.4). This approach is commonly used in a variety of speaking tests (Fulcher, 2003) and has also been used to study rater behavior within the TOEFL Speaking Test (Xi & Mollaun, 2009). The goals of the training were to instill in the raters a common understanding and application of the scoring criteria, and reduce differences in severity among raters (Fulcher, 2003).

Training was done on a self-access basis, using the same sort of interactive pdf instrument as used throughout the study. Rater progress through the session was controlled by locking the document content; as one step was completed the next step would unlock. The training session began with raters reading through the scoring rubric used for the study, followed by review of a series of benchmarks demonstrating performance at different levels, accompanied by scoring rationales. A total of ten benchmarks were presented, representing responses from five different prompts. The benchmarks and scoring rationales were taken from actual TOEFL materials: One set of four benchmarks responding to a single prompt was obtained directly from ETS, while other benchmarks were obtained from a TOEFL iBT preparation guide published by ETS (ETS, 2007) and a sample of TOEFL iBT sample test questions available online (ETS, 2011). Scores for the benchmarks were converted from the original TOEFL scale (1-4) to the 1-6 scale used in the study.

All of the benchmarks were examples of responses to TOEFL independent speaking tasks, but none addressed the actual prompts being scored in the study, since no
such exemplars were available. This practice is consistent, however, with ETS’s use of benchmark and calibration materials that are generic rather than identical to the item(s) being scored (Xi & Mollaun, 2009). However, written commentaries on how to score each of the two prompts used in the study were obtained from ETS and presented to raters prior to the scoring calibration exercise. These commentaries focused on the features of topic development that should be expected for the prompt. Scoring calibration exercises followed the exemplars, with separate sections for prompts A and B. Each section of the calibration exercise began with a review of the written scoring commentary for the prompt, followed by a series of exemplars demonstrating performance at each point of the scale (two exemplars for each scale point). This step was followed by a set of ten calibration items (for each prompt), wherein once a score was confirmed by the rater, feedback was provided in the form of the reference score for the response (rounded to the nearest whole number). The calibration items were chosen using a process similar to that used for the exemplars, so that in theory the calibration items should have been clear examples of performance at a given score level. Once the calibration sessions for both prompts were finished the rater training was complete. The training session typically required 1.5 to 2 hours.
5.5.6 Analyses.

A total of 400 scores, 100 scores per scoring session, were collected for each of 20 raters to make a total number of 8,000 scores. In addition, the time taken to score each response was recorded, and the number of times the exemplars were checked was tabulated for each scoring session. The analyses used with these data are described below.

5.5.6.1 Severity and consistency of scores. Multi-faceted Rasch analysis was the primary tool used to quantitatively analyze the severity and consistency of scores produced by raters. The FACETS software package (Version 3.62; Linacre, 2007b) was used to calculate the following:
- Rater severity, as measured in terms of logits (log odds). Raters of average severity are expected to have a logit value of zero; more lenient raters have negative logit values while more severe raters show positive values. Logit values are useful because they lie on a true interval scale (Bond & Fox, 2007), a necessary condition for the measures to be used with many other types of statistical tests.

- Rater consistency, as measured by model infit values for each rater (infit mean square and infit z values). Infit values have an expected value of 1.0, with higher values indicating less internal consistency (McNamara, 1996). In operational tests, infit mean square values between approximately 0.75 and 1.3 are considered acceptable, although for many applications, overfit (values below 0.75) has little practical impact other than to inflate the reported reliability for the test (Bond & Fox, 2007, p. 240). FACETS provides a variety of rater consistency measures, but infit values are often used because they are less sensitive to outlying values (Bond & Fox, 2007, p. 57).

- Bias towards either of the two prompts or specific test-takers (as indicated by scores that are significantly different from expected values for specific rater-by-prompt or rater-by-examinee combinations). FACETS provides a report of rater-prompt and rater-examinee pairings where the score(s) awarded are statistically different from what would be predicted from the scoring patterns observed in the data.

Two different approaches were taken to the analysis. First, separate analyses were conducted for each scoring session, providing independent measures of rater performance
for each session. Second, a combined analysis was conducted in which scores from each rater were separately entered by date. For example, scores from Rater 101 from sessions 1, 2, 3, and 4 were entered under the labels 1011, 1012, 1013, 1014, as if there were four different raters. This procedure has the potential to violate the assumption of local independence of items required for Rasch analysis given that the rater facet contains repeated measures, but the practical effect of this issue can be examined using the common responses repeated across sessions (John Linacre, personal communication, June 25, 2011). If the logit measures of the repeated responses are the same regardless of analysis approach (separate vs. combined analyses) then local dependence in the rater facet is not a problem. A common person linking plot (Bond & Fox, 2007, p. 80) was used to evaluate the equivalence of the repeated examinee measures, and the measures were found to indeed be equivalent within the bounds of error (Appendix I). Moreover, the rater severity measures themselves were essentially identical regardless of the analysis approach used; a simple regression analysis of the rater severity measures produced by the two analyses produced a Pearson’s $r$ value of .99, with an $R^2$ value of .98. The measures produced by the combined analysis are reported in the results section.

In addition to the results produced by many-facet Rasch analysis, inter-rater consistency was also examined in terms of inter-rater correlations and agreement. Within each of the four scoring sessions Pearson product-moment correlations were calculated for all possible rater pairings using SPSS (Gradpack version 13). The pairwise Pearson $r$ values were calculated using all 100 scores produced by each rater in each scoring session. Inter-rater reliability was then examined across time by entering the resulting $r$ values into a one-way repeated measures ANOVA, where scoring session was
the factor, with four levels. Effect size for the ANOVA main effect was calculated both in terms of eta squared ($\eta^2$), a commonly used measure of effect size based on the study sample, and omega squared ($\omega^2$) a more conservative statistic that estimates effect size for the population. Where the main effect of session was significant, effect size for specific comparisons between sessions were measured using $r$, calculated using the formula:

$$r = \sqrt{\frac{F(1, df_R)}{F(1, df_R) + df_R}}$$

where $F(1, df_R)$ is the F-ratio for the contrast and $df_R$ is the degrees of freedom (or error degrees of freedom; Field, 2005). Effect size was calculated similarly for all ANOVA analyses used in the study. It should be noted that repeated use of such ANOVA analyses may increase the possibility of type 1 errors (i.e., incorrectly detecting a significant difference across sessions), if the analyses are considered to address to a single phenomenon or family of hypotheses. Effect sizes are provided to aid in interpreting the importance of any statistically significant differences observed. Moreover, these analyses are exploratory in nature and no single analysis should be taken as conclusive evidence of causal relationships.

Rater agreement was also examined within each session, given that inter-rater correlations only indicate the degree to which pairs of raters similarly rank examinees and may mask disagreement in actual scores. Agreement was measured in each session by computing the percentage of exact agreements between pairs of raters, as well as Fleiss' kappa, which is an adaptation of Cohen's kappa for use with more than two raters (Fleiss, Levin, & Paik, 2003). A linearly weighted Fleiss' Kappa was calculated; a
weighted kappa was used because the data are ordinal (rather than nominal) and linear weightings were used because intervals of increasing disagreement were considered to be of equal importance (similar to Eckes, 2011). Calculations were made using AgreeStat (Version 2011.1, Advanced Analytics, 2011).

5.5.6.2 Accuracy of scores. The accuracy of scoring was operationalized as the degree of correlation between a rater's scores and the reference scores produced by the ETS scoring leaders. For each scoring session, a Pearson's $r$ value was calculated for the comparison of each rater's scores against the corresponding reference scores. These $r$ values were then compared graphically across time, as well as via a one-way repeated measures ANOVA with the factor “session” divided into four levels, and using $r$ values as the data. (Pearson $r$ values were adjusted using the Fisher z transformation before analysis to avoid distortions in variance associated with low or high $r$ values; Hatch & Lazaraton, 1991.) As an additional measure of scoring accuracy, agreement with the reference scores was examined by similarly calculating Cohen's kappa (Cohen, 1968) for each rater, within each session. Cohen's kappa is an index of agreement that takes into account the likelihood of chance agreement between judges (Fleiss, Levin, & Paik, 2003).

5.5.6.3 Scoring behavior. Scoring behavior was analyzed in terms of the number of times the exemplars were reviewed during the scoring session and the time required to make a scoring decision, calculated as the difference between the time one score was confirmed until the time the next score was confirmed. The time difference between confirmation of one score and clicking “play” for the next response was also calculated. Usually this was on the order of a few seconds, although longer values could occur if the examinee response was played more than once (only the final click of the play button was
time stamped). Values greater than 3 minutes were taken to indicate that the rater had taken a break while scoring and were tabulated separately.

The two measures of exemplar use and time required for a decision were examined because, to the researcher's knowledge, there is no published data regarding how these aspects of rater behavior associate with scoring patterns, and the technology used in this study (i.e., interactive forms) made it possible to collect such data. Also, conflicting predictions regarding the connection and between such behavior and rater expertise were considered equally plausible, making this an interesting issue for investigation. For example, increased use of the exemplars might indicate a careful, accurate rater who is repeatedly re-calibrating during the scoring session. On the other hand, high use of the exemplars might indicate a rater who is unsure of the rating scale and therefore feels the need to keep checking the scoring aids. Similarly, taking longer to make scoring decisions might indicate a rater taking extra care to thoroughly consider the merits of the response, or a rater who simply has difficulty in making scoring judgments. Exemplar use and the time required to make a scoring decision were each analyzed using a one-way repeated-measures ANOVA with scoring session as the repeated factor and four levels, one for each scoring session.

5.6 Research Question 2: Differences Between More-proficient and Less-Proficient Raters

5.6.1 Overview. The purpose of this analysis was to compare the behavior and cognition of specific individuals who demonstrated desirable or undesirable scoring patterns. Where research question 1 examined the effect of rater training and experience
as interventions, the focus here was to compare the qualities of groups of raters who produced specific scoring patterns. Based on the results of the analyses for research question 1, nine raters were chosen for this focused investigation, divided into three groups: more-proficient raters, less-proficient raters, and developing raters. For the purpose of answering this research question, scoring proficiency was defined in terms of three different measures: (a) rater severity measured via multi-faceted Rasch analysis, (b) rater consistency measured in terms of infit mean square values produced by Rasch analysis, and (c) scoring accuracy measured in terms of correlation with the reference scores produced by ETS raters. The specific criteria used to select raters for each group are given in Table 5.3.

Table 5.3

Selection criteria for raters used to answer research question 2

<table>
<thead>
<tr>
<th>More-proficient raters</th>
<th>Severity</th>
<th>Severity consistently within +/- .5 logit of the mean (zero) in all scoring sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consistency</td>
<td>Infit mean square values .9 or lower for all scoring sessions</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Pearson correlation with the reference scores .8 or higher for all sessions</td>
</tr>
<tr>
<td>Less-proficient raters</td>
<td>Severity</td>
<td>At least one session with severity greater than +/- 1 logit away from the mean; relatively large changes in severity across sessions that do not trend towards the mean</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>No specific criterion.</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Pearson correlation with the reference scores .75 or lower, with no consistent upward trend over time</td>
</tr>
<tr>
<td>Improving raters</td>
<td>Severity</td>
<td>No specific criterion.</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>Infit mean square values consistently decreasing over time, with a total decrease of at least .2</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Pearson correlations with the reference scores consistently increase over time, with a total increase of least .1</td>
</tr>
</tbody>
</table>
There was little difference in the backgrounds of raters selected for each group (Table 5.4) although the more-proficient raters had, on average, somewhat more teaching experience: 8.5 years versus 5.7 and 6.0 years, respectively, for the less-proficient and improving rater groups. Other differences included the fact that none of the improving raters were students, while all of the other raters were current students, and none of the less-proficient raters had taught a TOEFL preparation class, while the majority (5 of 6) of the more-proficient and improving raters had taught some sort of TOEFL preparation class.

Table 5.4

\[\textit{Rater background information}\]

<table>
<thead>
<tr>
<th>Educational background</th>
<th>More-Proficient</th>
<th>Less-Proficient</th>
<th>Improving</th>
</tr>
</thead>
<tbody>
<tr>
<td>TESL Certificate</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M.A. TESOL or similar degree</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Other M.A.</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Currently a student</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Taken a course(s) in language testing</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teaching Experience</th>
<th>More-Proficient</th>
<th>Less-Proficient</th>
<th>Improving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average years teaching ESL/EFL</td>
<td>8.5</td>
<td>5.7</td>
<td>6.0</td>
</tr>
<tr>
<td>Currently teaching ESL/EFL</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing experience</th>
<th>More-Proficient</th>
<th>Less-Proficient</th>
<th>Improving</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL iBT speaking test rater</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Taught a TOEFL preparation class</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>TOEFL Speaking Test class/tutoring</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Rater in a different speaking test</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
5.6.2 Instruments and procedures.

5.6.2.1 Rater scoring behavior. Indicators of scoring behavior for the nine raters were collected during the scoring sessions as described previously for research question 1. As before, the measures used were the number of times the exemplars were checked during the scoring session and the average time required to make a scoring decision.

5.6.2.2 Rater cognition. Data regarding rater cognition was collected using verbal report, specifically, stimulated recall (Gass & Mackey, 2000). Although concurrent verbal report techniques such as think aloud have been widely used in the study of writing assessment (and at least one study of rater cognition in a speaking test; Meiron, 1998), retrospective verbal reporting techniques such as stimulated recall are generally considered to be less likely to interfere with the scoring process within speaking test contexts (Lumley & Brown, 2005), and may also require less training because they are easier to produce (Gass & Mackey, 2000). Stimulated recall is also used to improve the fidelity and depth of the recall beyond what would be available from a simple recounting of the scoring process; it is thought that providing a reminder of the event improves access to memory (Gass & Mackey, 2000, p. 17).

As described earlier, raters were trained in how to produce stimulated recalls during the orientation session for the study. For concurrent reports, it has been noted that training may increase the completeness of the report (Ericsson & Simon, 1993) or help the participant to focus the report on cognition, avoiding simple “blow by blow” activity descriptions (Green, 1998). On the other hand, while Gass and Mackey (2000) agree that carefully-worded instructions are important for the collection of valid data using stimulated recall, they argue against providing overly extensive training since this may
prime or otherwise influence participant responses (p. 52). The recall training used in this study attempted to balance the need to focus raters' attention on the type of comments they should be producing, while avoiding undue influence on the content of their reports.

As mentioned earlier, the orientation session included a practice scoring activity where the rater first listened to a series of exemplars and then scored a set of ten responses. This activity served to familiarize the raters with the general process of scoring and was followed by training in the stimulated recall procedure. The stimulated recall training started with the researcher giving an explanation of the general purpose for collecting the stimulated recall as well as a description of the type of comments the raters should be making. In particular, it was stressed that the rater should “think out loud” rather than attempting to give an “explanation” or narrative of their scoring. Raters were also asked to try to recall their original thoughts while scoring, to the extent possible, and some questions were provided to guide raters in case they had difficulty in thinking of something to say. The recall process was not demonstrated, to avoid biasing raters towards attending to particular aspects of the examinee's response, but written instructions were placed at the beginning of the each set of recalls throughout the study to guide the recall process (Appendix J).

Following the discussion of the format of the recall, the specifics of how to record the recall were demonstrated using the rater's computer (and external voice recorder, if used). The general procedure was to first score the response; selecting a score would then open a message that asked the rater to record their recall and provided a name for the resulting audio file. The rater would then start recording their own voice (using installed recording software or an external recorder), begin replay of the examinee response, and
start their recall. Raters were asked to pause the recording at least two or three times to record their thoughts, but were also allowed to speak over the recording without pausing, stop and re-play part of the recording to make comments, and comment after the replay was finished. Generally, most raters paused the recording several times to make comments and then made additional comments at the end. Once the rater had confirmed that they had recorded a recall and saved the resulting audio file, the next response was unlocked. During the orientation session raters completed a set of ten recalls. Within a day following the orientation, the researcher listened to the recalls and gave written feedback to the rater, if needed, to make sure the recalls were intelligible and provided useful data. In two cases raters first provided only narrative explanations of why they gave a certain score, rather than the think-aloud comments targeted by the exercise. These individuals were given feedback regarding the type of comments targeted by the recalls and asked to complete another set of recalls before being allowed to continue with the study.

The collection of recalls during the scoring session followed the same process. As mentioned earlier, ten recalls were done at the end of the scoring of the first prompt (halfway through the overall scoring session) with another ten recalls done for the other prompt at end of the scoring session.

5.6.3 Analyses.

5.6.3.1 Scoring patterns. In addition to the analyses of scoring patterns reported earlier for research question 1, within-rater consistency across time was calculated in terms of rater agreement as measured by a linearly weighted Fleiss' kappa. In addition, use of the rating scale was examined in a secondary many-facet Rasch analysis.
employing a “partial credit model” (Masters, 1982) which does not assume a common interpretation of the rating scale. This differs from the earlier analysis which used the “rating scale model” (Andrich, 1978), which assumes all raters have a common understanding of the rating scale. Using the partial credit model, the FACETS software produced separate raw scale representations for each rater, as well as logit measures for the boundary points between adjacent score values. These boundary points were used to construct a graphical comparison of raters' interpretations of the rating scale.

5.6.3.2 Scoring behavior. Due to the reduced sample size of the focal groups (three raters per group) no inferential statistics were used. Rather, patterns in the use of the exemplars and the time required to make scoring decisions were explored graphically.

5.6.3.3 Rater cognition. All stimulated recalls were transcribed using a word-for-word transcription. Following transcription, ten recalls for each rater (five for prompt A and five for prompt B) for scoring sessions 1, 2, and 4 were randomly selected for coding, which represented half of the recalls actually produced. This sampling of the data was done to limit the time required for coding to a manageable level, given that 1,598 transcripts were collected with a total of approximately 200,000 words. The transcripts of the selected recalls were then segmented into units wherein each segment represented a single cognitive process or action (Green, 1998). Often such segments corresponded to a clause, but ranged from a single phrase to several clauses long, and in any case captured specific actions that are not obviously decomposable into sub-actions, based on the data. A simple coding scheme was then developed based on three randomly selected recalls from each rater, collected in scoring session 1. The coding book was further refined as
coding continued, with final codings for all recalls reviewed and revised as needed at the end of the coding process.

The coding scheme focused on two aspects of rater cognition: (a) language features attended to and (b) mention of scores and the scoring process. Language features mentioned by raters were coded to evaluate the claim that rater training functions to clarify the criteria used for scoring (e.g., Fulcher, 2003; Weigle, 2002). Accordingly, specific language features were coded into one of the three categories used in the TOEFL scoring rubric: delivery (pronunciation and fluency), language use (grammar and vocabulary), and topical development (elaboration, precision, coherence). Comments regarding scoring were divided into two categories: (a) mention of a score for all or part of a response (such as “This person is about a three so far”) and (b) comments regarding the scoring process (scoring meta-talk such as “I checked the rubric here”).

Other categories that required coding included a “general” category used to code comments addressing the response as a whole, an “unclear” category where the language focus of the comment could not be determined (such as when the rater simply repeated an examinee's utterance), a “technical” code used for comments about technical issues (such as poor sound quality), and an “other” category used for comments that did not fit elsewhere, such as speculation regarding the examinee's background or intentions. The codes used and examples of each from the data are provided in Appendix K.

A total of 40% of the stimulated recall transcripts were coded by a second individual, who was an experienced teacher of English as a Foreign language and was trained to use the coding system. Stratified random sampling was used to select the recalls to be double-coded; for each rater, two recalls for each prompt were double coded
for each scoring session. The overall inter-coder agreement was 82.3% (974 of a total of 1184 coded units).

5.7 Research Question 3: Evaluation of the Relative View of Magnitude Judgments

5.7.1 Overview. This part of the study investigated a possible psychological mechanism for magnitude judgments of speaking ability; namely, that such judgments are fundamentally based on comparisons of one thing with another rather than application of an internalized scale. It has also been argued that this model is directly relevant to scoring judgments made in examination contexts (Laming, 2003), and the goal of this section of the study was to evaluate the usefulness of the relative judgment model for understanding rater decision making within a speaking test context.

The relative judgment view suggests that magnitude judgments may be informed by examples of similar items in the immediate environment. In a language testing situation, this would predict that when making a scoring decision raters compare the test taker response to other recently encountered performances. Given that more recently encountered examples should exert a stronger influence (Laming, 1997; Mathews & Stewart, 2009) a sequence effect in scoring might be expected in which each score is correlated with the previous score (also termed anchor and adjustment bias; Plous, 1993). Additionally, it might be expected that as a novice rater encounters more and more examinees, the increasingly robust collection of examinee performances contained in the rater's memory should exert a stabilizing effect on decision making, reducing the magnitude of any sequence effect. The first part of this section of the study investigated
whether such a sequence effect was apparent in the scores collected to answer research question 1.

Another prediction of the relative judgment view is that more precise judgments can be made when a point of reference is provided for comparison, and therefore discrimination between pairs of items should be more precise than magnitude judgments made in isolation (Laming, 1997). This prediction was tested in the present study by a separate experiment where a set of examinee responses were judged in both isolated and pairwise judgment conditions. In the isolated condition, individual scores were awarded one by one, similar to regular scoring, while in the pairwise judgment condition, pairs of responses were presented and the rater identified whether the responses merited the same or different scores. The sensitivity of discrimination in both conditions was then compared using the statistic d-prime ($d'$), which is used in the domain of signal detection theory as a measure of the sensitivity of discrimination between two stimuli (Macmillan & Creelman, 2005).

5.7.2 Instruments and procedures.

5.7.2.1 Sequence effect. As mentioned above, the examination of sequence effect in scoring was conducted using scores collected as part of research question 1. Scores from all four scoring sessions were used, except for those generated while producing verbal reports.

5.7.2.2 Comparison of pairwise and isolated judgments. The precision of discrimination of scoring judgments made individually was compared to the precision of judgments made when two responses were presented in immediate succession. The experiment employed the same raters used throughout the study, with data collection
coming after completion of the final scoring session (session 4). Accordingly, raters had considerable experience with the scoring context by the beginning of this portion of the experiment, having produced over 500 scores by this point. The experiment was carried out in two sessions separated by a period of one week: one session for the isolated judgment condition and another for the pairwise condition. The first session was generally scheduled to start 5-7 days after scoring session 4, although due to a delay in finalizing the data collection instruments the interval between scoring session 4 and the start of the experiment was up to three weeks in a few cases. Like the scoring sessions, three instruments were used in each of the two experimental sessions: separate scoring instruments for each prompt (A and B) and a post-scoring wrap-up survey with questions regarding how the raters felt about the session. The presentation order of the session (isolated or pairwise) and prompt (A or B) were individually randomized for each rater.

A key issue in the design of the experiment was the need for some form of reference measure of examinee ability to use in selecting the responses for the experiment. For example, to investigate discrimination between two examinee responses at the level of a “4” versus a “5”, there would need to be some sort of independent measure that can be used to select responses that truly represent a “4” and a “5”. Similar studies within psychophysics address this issue by using objective physical measures to select stimuli (e.g., weight measured using a scale). For oral language proficiency no such objective measures are available, but this work took the view that measures of speaking ability, as obtained from multiple ratings trained and experienced raters, provided an analog to the physical measures used in psychophysics studies. Similar approaches have been taken by others attempting to apply psychophysical approaches to
complex judgments of magnitude. For example, Ozer (1993) used a written psychological inventory as the basis for comparison with rater judgments of personality traits, Stricker (1997) used student ratings of international teaching assistants as a comparison metric for scores on the SPEAK test, and Matthews and Stewart (2009) used catalog prices as a basis for examining participant judgments of the price of chairs and footwear. In the current study, the reference scores produced by TOEFL Speaking Test scoring leaders were used as measures of the “true” scores for the responses used as stimuli in the experiment.

The same 60 examinee responses were used in both the isolated and pairwise scoring conditions. For each prompt (A and B) ten responses were chosen to represent performances at each of the score levels of 3, 4, and 5. Responses were selected on the basis of the following criteria:

1. The reference score for the response was as close as possible to the intended nominal score. Ideally the reference score was within +/- 0.1 score point of the nominal score, if possible.

2. Responses with Rasch analysis infit mean square values greater than 1.3 were excluded, since such responses poorly fit the Rasch model used to produce the reference scores.

3. Responses previously used as exemplars or between-session repeats were excluded to minimize any effect of familiarity with the response.

By the time the experiment was conducted, all available responses had been used at least once and so responses were reused, except for those mentioned in criterion #3 above.

Descriptive statistics for the responses chosen are given in Table 5.5.
Table 5.5

*Descriptive statistics for responses used in relative judgment experiments*

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Reference score&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score = “5”</td>
<td>10</td>
<td>4.95</td>
<td>4.84</td>
<td>5.10</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Score = “4”</td>
<td>10</td>
<td>3.99</td>
<td>3.92</td>
<td>4.06</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Score = “3”</td>
<td>10</td>
<td>3.00</td>
<td>2.89</td>
<td>3.10</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Prompt B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score = “5”</td>
<td>10</td>
<td>4.99</td>
<td>4.86</td>
<td>5.13</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>Score = “4”</td>
<td>10</td>
<td>3.99</td>
<td>3.90</td>
<td>4.09</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Score = “3”</td>
<td>10</td>
<td>3.00</td>
<td>2.86</td>
<td>3.15</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Reference scores were adjusted mean (fair-M) values obtained from multifaceted Rasch analysis of raw scores from TOEFL iBT scoring leaders (N=11)

### 5.7.2.3 Isolated judgment

The procedure used for the isolated judgment scoring condition was largely similar to the previous scoring sessions. Raters began by reviewing the rubric and exemplars for the prompt and then began scoring. Like previous scoring, raters could freely replay the response while deciding on a score, but once a decision was confirmed the response was locked and could not be changed or reviewed. One substantial modification to the scoring process was that, between each response, raters were required to carry out three simple arithmetic calculations in their heads. The arithmetic problems were of the general format \((a+b)+(c+d)=?\) where each letter represents a single digit integer (including negative numbers) and any plus sign may be replaced with a minus sign. Everyone completed the same set of problems, with the presentation order separately randomized for each rater; a pilot study suggested that the three problems would take 20-40 seconds to complete. This additional step was added to the scoring process to distract the rater, making it more difficult to remember details of the previous response and reducing the likelihood that the previous response would be
used as a point of reference when scoring the next response. Other modifications to the scoring procedure were as follows:

1. Only three score choices were provided (3, 4, or 5) to correspond to the truncated range of responses used.

2. Exemplars were limited to these three score levels, with two exemplars provided for each score band.

3. A review of the scoring commentary for the prompt, as provided in the rater training session, occurred prior to the beginning of scoring.

4. Raters were prohibited from reviewing the exemplars/rubric/commentary while scoring to avoid comparisons between the scoring aids and the response being scored.

The session as a whole required approximately 2 hours to score the 60 responses.

5.7.2.4 Pairwise judgment. Like the isolated judgment condition, the pairwise judgment sessions started with review of the rubric, the scoring commentary for the prompt, and the exemplars (two exemplars each for scores 3, 4, and 5). Raters were also prohibited from using the scoring aids once they started the data collection exercise, which consisted to listening to a pair of responses presented in immediate succession and then deciding whether the two responses merited the same score or different scores. Raters were allowed to listen to the response pair as many times as they liked while making a decision, but once a decision was confirmed the response pair was locked; no arithmetic problems or other distracters were used between decisions.

For each prompt (A or B) a total of 60 pairs of responses were presented, with pairs constructed using the same responses used for the isolated judgment condition.
Pairs were presented in ten blocks, with each block containing one example of each of the six possible response pairings (5-5, 4-4, 3-3, 3-4, 4-5, 3-5). No response appeared more than once in a block, although the design required that each response appear in four different blocks across the session. The same blocks were used with all raters, with the presentation orders of pairings within block and blocks within session separately randomized for each rater. Responses were paired by first sorting them by reference score and then matching the highest “5” with the highest “4” and so on down the list; this measure was taken to increase the consistency of the score difference represented by a particular pairing. Once paired, responses were combined into an audio file with 2 seconds of silence between the responses. The choice of which response was presented first was also randomized by creating two versions of each audio file, one for each presentation order, and then randomly selecting one of the two versions. The total time required to complete the whole session of 120 pairings was approximately five hours.

5.7.3 Analysis.

**5.7.3.1 Sequence effect.** The analysis of sequence effect followed the approach of Mathews and Stewart (2009), where multiple regression was used to examine the portion of variance in scores awarded on trial \( n \) (\( \text{Judgment}_n \), or \( J_n \)) with the following independent variables:

1. the reference score value for the response on trial \( n \) (\( R_n \)),
2. the judgment in the previous trial (\( J_{n-1} \)), and
3. the reference score from the previous trial (\( R_{n-1} \)).

The multiple regression was conducted using SPSS statistical software (GradPack ver. 13.0) with all three independent variables (\( R_n, J_{n-1}, R_{n-1} \)) entered into the regression.
model simultaneously. Analysis of regression coefficients for pairwise contrasts between $J_n$ and the other variables were also produced along with significance tests for these contrasts.

Matthews and Stewart noted that the variables $J_{n-1}$ and $R_{n-1}$ are very likely to be correlated; that is, rater judgment of the response presented at trial $n-1$ should be correlated with the “true” value of the response being judged in the same trial. Such multicollinearity increases the standard error of the correlation coefficients for pairwise comparisons of variables and reduces the likelihood of observing statistical significance (Field, 2005). Following Matthews and Stewart’s lead, variance inflation factors (VIF; Neter, Wasserman, & Kutner, 1985) were calculated for each variable within each regression analysis to determine if multicollinearity was excessively influencing error estimates. Relatively high VIF values and high correlations between $J_n$ and $J_{n-1}$ were in fact observed; since the relationship between $J_n$ and $J_{n-1}$ were of most interest, a second set of analyses was conducted in which was $R_{n-1}$ was excluded.

The correlation coefficients for $J_n$ vs. $J_{n-1}$ were taken from each multiple regression analysis, transformed to Fisher’s $z$ to avoid distortions in variance associated with low or high $r$ values (Hatch & Lazaraton, 1991), and entered into a single-sample $t$-test to determine if the average of correlation coefficients was greater than zero (Matthews & Stewart, 2009). Separate $t$-tests were done for each of the four scoring sessions, with significance level adjusted for multiple comparisons using a Bonferroni adjustment (Brown, 2001). Finally, the correlation coefficients were entered into a one-way repeated measures ANOVA to examine whether the correlation between judgments
at trials $n$ and $n-1$ diminished with time, as would be predicted by the relative judgment model.

5.7.3.2 **Comparison of pairwise and isolated judgment.** The relative precision of judgments made in the isolated and pairwise conditions was examined using two measures. First, the proportion of correct decisions was compared across both conditions. For the pairwise condition, the proportion correct was simply the proportion of decisions where a pairing was correctly identified as “same” or “different” (i.e., the two responses should receive the same or different scores). For the isolated judgment conditions, scores for individual responses were first paired to replicate the pairings used in the pairwise condition and then the scores compared to arrive at a “same” or “different” decision. The proportion of correct decisions was then calculated as for the isolated condition. A higher proportion of correct decisions was taken as evidence of greater ability to discriminate between responses at different score levels.

As a metric, however, the proportion correct is vulnerable to certain kinds of bias (Macmillan & Creelman, 2005). For example, when there are an equal number of same and different pairs, a person who always responds “same” can achieve a correct decision rate of 50% without even listening to the response pair. To correct for such bias, the statistic $d'$ (d-prime) was calculated as a measure of raters' ability to distinguish different levels of performance. D-prime is calculated as the $z$-score of the hit rate (correctly saying “same” when two responses are the same) minus the $z$-score of the false alarm rate (saying “same” when the responses are actually different). Rather than calculate an overall statistic for the whole session (as done for proportion correct) $d'$ values are calculated for specific contrasts; for this study this procedure resulted in three $d'$ values.
for each session corresponding to discrimination between scores of 3 vs. 4, 4 vs. 5, and 3 vs. 5 (d'(3,4), d'(4,5) and d'(3,5)).

Another advantage of d’ is that it could be directly calculated for both the pairwise and isolated judgment designs without any post hoc manipulation, as was required for the proportion correct calculation. For example, in the pairwise condition d'(3,4) was calculated based on the number of times the rater responded “same” when a 3-3 pair was presented (the number of hits) and the number of times “same” was chosen when a 3-4 pair was presented (the number of false alarms). Calculations for the isolated condition were similar, with hits counted when the raters awarded a “3” when the response was actually a 3, and false alarms counted when the rater responded “3” but the response was actually a “4” (Macmillan & Creelman, 2005). Once d’ values had been calculated for both conditions, values were compared across conditions using a paired samples t-test.

In addition, data from the wrap-up survey were also used to examine rater perceptions regarding the relative ease or difficulty of each approach, and to understand how raters approached the task. The questions asked in the survey are given in Appendix L.
CHAPTER 6

Results

This chapter reports the results of the study, organized by research question. As described in the previous chapter, these questions were as follows:

1. What effects do training and experience have on raters’ scoring patterns and scoring behavior?

2. What features of scoring behavior and cognition distinguish more-proficient and less-proficient raters?

3. Can rater decision making and expertise be understood in terms of a relative view of magnitude judgment?

Section 6.1 presents results for research question 1 and contains findings regarding scoring patterns and scoring behavior, including measures of scoring severity and consistency obtained from multi-faceted Rasch analysis. The accuracy of scores in terms of correlation and agreement with reference scores is reported, and rater behavior, in terms of the use of exemplars while scoring and the time taken to reach a scoring decision, is described. Section 6.2 addresses research question 2, where rater behavior and cognition are examined for a subset of raters who demonstrated more- or less-desirable scoring patterns. As with research question 1, behavior while scoring is examined in terms of exemplar use and decision time. This examination is followed by a description of the relative frequency with which different raters attended to various language features while scoring as well as the extent to which they were explicitly focused on the scoring process itself. Finally, section 6.3 reports the results for research question 3, describing the results of efforts to evaluate two assumptions of a relative view.
of magnitude judgment: (a) whether or not a sequence effect is observable in the scoring data, and (b) whether discrimination of test takers of different ability levels is improved when test takers can be judged side-by-side compared to scoring individuals in isolation.

6.1 Effects of training and experience on scoring patterns and behavior

6.1.1. Many-facet Rasch analysis. Figure 6.1 shows the Rasch measurement rulers for all scores generated in the study. The first column shows a logit (log odds) scale, which is a common interval scale on which measures of the facets of the scoring situation (examinee, prompt, and rater) are plotted; in all cases a value of zero logits represents the average value for the facet (much as zero is the mean value for a standardized score, or $z$-score). The examinee column plots test takers’ language ability, where higher values represent increasing levels of language ability. The topic column plots the difficulty of each prompt while the rater column shows the severity of each rater; in these columns higher logit values indicate more difficult topics or more severe raters. The scale column is a graphical representation of the raw scores showing the relationship of raw scores to the logit scale. The logit scale integrates all of these facets by giving the probability that a particular test taker/topic/rater combination will result in a particular score. For example, in Figure 6.1 an examinee with an ability of 1 logit, tested using a topic and rater of average difficulty (zero logits), would be expected to receive a score of four or better about 50% of the time.

Figure 6.1 shows that examinee abilities were spread over a range of approximately 8 logits and were approximately normally distributed. Both prompts were approximately equal in difficulty; prompt A was slightly more difficult with a measure of
0.13 logits while prompt B was slightly easier with a measure of -0.13 logits. This difference was statistically significant (fixed effects chi-square: $\chi(1) = 71.5, p < 0.01$) but the magnitude was small (0.26 logits); for a dichotomous item, this would correspond to an 8% increase in the probability of a correct answer if given the easier prompt (Linacre, 2010). In addition, such difference in topic difficulty is accounted for by the Rasch model when estimating rater attributes, so any difference is of little concern for the analysis of raters.
<table>
<thead>
<tr>
<th>Logit</th>
<th>Examinee</th>
<th>Prompt</th>
<th>Rater</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>6 +</td>
<td>+</td>
<td>+</td>
<td>+ (6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>5 +</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>4 +</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>3 + *</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>2 + ******</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>********</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>********</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>********</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>1 + ******</td>
<td>+</td>
<td>+ *</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>********</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>********</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>********</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>0 + *********</td>
<td>* A B</td>
<td>* ******</td>
<td>--- *</td>
</tr>
<tr>
<td></td>
<td>*********</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*********</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*********</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>-1 + ****</td>
<td>+</td>
<td>+ **</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>****</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>-2 + **</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>-3 + *</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>-4 + *</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>-5 +</td>
<td>+</td>
<td>+</td>
<td>+ (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logit</th>
<th>* = 2</th>
<th>Prompt</th>
<th>* = 2</th>
<th>Scale</th>
</tr>
</thead>
</table>

*Figure 6.1. Measurement rulers for many-facet Rasch analysis of scores.*
Table 6.1

Rater severity measures from multifaceted Rasch analysis

<table>
<thead>
<tr>
<th>Rater</th>
<th>Session 1 Ratings</th>
<th>Session 1 Measure S.E.</th>
<th>Session 2 Ratings</th>
<th>Session 2 Measure S.E.</th>
<th>Session 3 Ratings</th>
<th>Session 3 Measure S.E.</th>
<th>Session 4 Ratings</th>
<th>Session 4 Measure S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>100</td>
<td>0.07 0.13</td>
<td>100</td>
<td>-0.10 0.13</td>
<td>100</td>
<td>0.40 0.13</td>
<td>100</td>
<td>0.22 0.13</td>
</tr>
<tr>
<td>102</td>
<td>100</td>
<td>-0.61 0.13</td>
<td>100</td>
<td>-0.05 0.13</td>
<td>100</td>
<td>0.01 0.13</td>
<td>100</td>
<td>-0.21 0.13</td>
</tr>
<tr>
<td>103</td>
<td>100</td>
<td>0.08 0.13</td>
<td>100</td>
<td>0.75 0.13</td>
<td>100</td>
<td>0.50 0.13</td>
<td>100</td>
<td>0.59 0.13</td>
</tr>
<tr>
<td>105</td>
<td>100</td>
<td>-0.42 0.13</td>
<td>100</td>
<td>-0.99 0.13</td>
<td>100</td>
<td>-0.29 0.13</td>
<td>100</td>
<td>-0.09 0.13</td>
</tr>
<tr>
<td>107</td>
<td>100</td>
<td>0.66 0.13</td>
<td>100</td>
<td>0.59 0.13</td>
<td>100</td>
<td>0.29 0.13</td>
<td>100</td>
<td>-0.12 0.13</td>
</tr>
<tr>
<td>108</td>
<td>100</td>
<td>-1.81 0.14</td>
<td>100</td>
<td>-0.52 0.13</td>
<td>100</td>
<td>-1.19 0.13</td>
<td>100</td>
<td>-1.10 0.13</td>
</tr>
<tr>
<td>109</td>
<td>100</td>
<td>1.16 0.13</td>
<td>100</td>
<td>-0.61 0.13</td>
<td>100</td>
<td>-1.06 0.13</td>
<td>100</td>
<td>-0.76 0.13</td>
</tr>
<tr>
<td>111</td>
<td>100</td>
<td>-0.42 0.13</td>
<td>100</td>
<td>-0.05 0.13</td>
<td>100</td>
<td>0.50 0.13</td>
<td>100</td>
<td>-0.38 0.13</td>
</tr>
<tr>
<td>112</td>
<td>100</td>
<td>-0.65 0.13</td>
<td>100</td>
<td>0.36 0.13</td>
<td>100</td>
<td>0.67 0.13</td>
<td>100</td>
<td>-0.10 0.13</td>
</tr>
<tr>
<td>113</td>
<td>99</td>
<td>0.22 0.13</td>
<td>100</td>
<td>1.60 0.14</td>
<td>100</td>
<td>0.67 0.13</td>
<td>99</td>
<td>0.30 0.13</td>
</tr>
<tr>
<td>115</td>
<td>100</td>
<td>0.31 0.13</td>
<td>100</td>
<td>-0.51 0.13</td>
<td>100</td>
<td>-0.57 0.13</td>
<td>100</td>
<td>-0.41 0.13</td>
</tr>
<tr>
<td>116</td>
<td>100</td>
<td>0.48 0.13</td>
<td>100</td>
<td>0.95 0.13</td>
<td>100</td>
<td>1.17 0.14</td>
<td>100</td>
<td>0.92 0.13</td>
</tr>
<tr>
<td>117</td>
<td>99</td>
<td>0.25 0.13</td>
<td>98</td>
<td>0.14 0.13</td>
<td>100</td>
<td>0.31 0.13</td>
<td>100</td>
<td>0.10 0.13</td>
</tr>
<tr>
<td>119</td>
<td>100</td>
<td>-0.11 0.13</td>
<td>100</td>
<td>-0.51 0.13</td>
<td>100</td>
<td>0.03 0.13</td>
<td>100</td>
<td>0.08 0.13</td>
</tr>
<tr>
<td>120</td>
<td>96</td>
<td>-0.95 0.14</td>
<td>100</td>
<td>-0.74 0.13</td>
<td>100</td>
<td>-0.89 0.13</td>
<td>100</td>
<td>-0.64 0.13</td>
</tr>
<tr>
<td>121</td>
<td>100</td>
<td>0.54 0.13</td>
<td>100</td>
<td>0.20 0.13</td>
<td>100</td>
<td>0.38 0.13</td>
<td>100</td>
<td>0.31 0.13</td>
</tr>
<tr>
<td>122</td>
<td>100</td>
<td>-0.09 0.13</td>
<td>100</td>
<td>0.75 0.13</td>
<td>100</td>
<td>0.72 0.13</td>
<td>98</td>
<td>0.09 0.13</td>
</tr>
<tr>
<td>123</td>
<td>100</td>
<td>-0.28 0.13</td>
<td>100</td>
<td>-0.67 0.13</td>
<td>100</td>
<td>-0.38 0.13</td>
<td>100</td>
<td>-0.14 0.13</td>
</tr>
<tr>
<td>124</td>
<td>100</td>
<td>-0.33 0.13</td>
<td>100</td>
<td>0.32 0.13</td>
<td>100</td>
<td>0.08 0.13</td>
<td>100</td>
<td>0.07 0.13</td>
</tr>
<tr>
<td>125</td>
<td>100</td>
<td>0.96 0.13</td>
<td>100</td>
<td>0.84 0.13</td>
<td>100</td>
<td>-0.24 0.13</td>
<td>100</td>
<td>-0.64 0.13</td>
</tr>
</tbody>
</table>

Avg. -0.047 0.087 0.055 -0.096
SD 0.686 0.686 0.635 0.475

Model, Sample: RMSE .13 Adj (True) S.D .60 Separation 4.54 Reliability (not inter-rater) .95
Model, Fixed (all same) chi-square: 1663.8 d.f.: 79 significance (probability): .00
Inter-Rater agreement opportunities: 114133 Exact agreements: 40833 = 35.8% Expected: 39240.9 = 34.4%
6.1.2 Between-rater consistency in scoring. Measures of rater severity were spread across approximately 3 logits and appeared normally distributed, with most (82%) of the severity measures falling within one logit of the mean value of zero (Figure 6.1, Table 6.1). (The large number of raters shown, 80 in total, is a result of each rater being entered into the analysis four times, once for each scoring session.) This range of severity values is not unusual for language performance tests (Eckes, 2011; McNamara, 1996), but it nonetheless indicates that some raters tended to grade more severely while others graded more leniently. Differences in rater severity were statistically significant (fixed effects chi-square: \( \chi(79) = 1663.8, p < .01 \)). Given that all raters scored the same responses, this result suggests that there was a significant degree of disagreement between raters regarding the scores to be awarded to particular examinees.

Variation in rater severity both within and across scoring sessions can be seen in Figure 6.2. Throughout the four scoring sessions, rater severity consistently spread across a band of plus/minus one logit from the mean, suggesting that training and experience had little influence on inter-rater consistency at the overall group level. Of particular interest is the finding that, in the first scoring session, 90% of the raters had severity measures that were within one logit of the mean. This indicates that novice raters were able to achieve levels of inter-rater consistency typical for operational language tests before they had completed the rater training, or had even seen the scoring rubric. A similar lack of change in collective rater severity over time was observed in the means of scores given to a set of 16 responses that were repeated across all scoring sessions (Appendix M). A one-way repeated measures ANOVA performed on the mean value from each rater indicated no difference between sessions, \( F(1.539,0.143) = 1.419, p = \)
.229, $\eta^2 = .075$, $\omega^2 = .016$; degrees of freedom corrected because of unequal variances using Greenhouse-Geisser estimates of sphericity, $\varepsilon = .625$). This result again suggests that on average, raters maintained a consistent degree of severity across the four scoring sessions.

Figure 6.2. Rater severity (in logits) across scoring sessions. Each line represents a single rater.

The severity data reported above only indicate that raters, on aggregate, did not tend to become more or less severe over time and maintained a consistent range of variability in severity. On the other hand, when rater variability was examined in terms of inter-rater correlations of raw scores, there appeared to be an increase in inter-rater consistency following training. Table 6.2 shows the average of pairwise Pearson product-moment correlations for all possible rater pairings within each session. All raters scored the same set of responses in a given session, and so the pairwise Pearson $r$ values were calculated using 100 scores each (or in some cases slightly fewer if a rater declined to
score a response due to technical problems). Mean Pearson $r$ values rose from 0.596 in session 1 to .673 following the training session, and then were somewhat lower for the remaining two scoring sessions. A one-way repeated measures ANOVA (where each rater pairing was considered to be a context repeated across sessions) was conducted using data transformed using a Fisher $z$ transformation. There was a statistically significant difference in values across sessions, $(F(3, 567) = 77.855, p < .001, \eta^2 = .292, \omega^2 = .143)$ The effect size ($\eta^2$) of .292 suggests an moderate effect of scoring session (Ferguson, 2009). Pairwise comparisons of means, following a Bonferroni adjustment for multiple contrasts, found all means to be significantly different from each other at the $p < .05$ level, with the exception of session 3 versus session 4 where inter-rater Pearson correlation had dropped slightly and appeared to stabilize at a level of .65-.66. The magnitude of the difference in means was modest, with the largest difference being .08, seen between session 1 and session 2 (before and shortly after training). Nonetheless, this difference amounted to an increase of 13% in the inter-rater correlation, and constituted a relatively large effect ($r = .71$; Table 6.3).
Table 6.2

*Mean pairwise inter-rater correlations within scoring sessions*

<table>
<thead>
<tr>
<th>Session</th>
<th>Mean</th>
<th>N</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>.596</td>
<td>190</td>
<td>.074</td>
<td>.413</td>
<td>.802</td>
<td>.585</td>
<td>.606</td>
</tr>
<tr>
<td>Session 2</td>
<td>.673</td>
<td>190</td>
<td>.067</td>
<td>.430</td>
<td>.813</td>
<td>.664</td>
<td>.683</td>
</tr>
<tr>
<td>Session 3</td>
<td>.652</td>
<td>190</td>
<td>.073</td>
<td>.409</td>
<td>.809</td>
<td>.641</td>
<td>.662</td>
</tr>
<tr>
<td>Session 4</td>
<td>.659</td>
<td>190</td>
<td>.066</td>
<td>.432</td>
<td>.828</td>
<td>.650</td>
<td>.669</td>
</tr>
</tbody>
</table>

*Note:* The N value is the number of pairwise Pearson product-moment correlations used to calculate summary statistics.

Table 6.3

*Effect sizes for changes across sessions in inter-rater correlations.*

<table>
<thead>
<tr>
<th>Contrast</th>
<th>df</th>
<th>F</th>
<th>Significance (p)</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1 vs. Session 2</td>
<td>1</td>
<td>187.291</td>
<td>&lt;.001</td>
<td>.71</td>
</tr>
<tr>
<td>Session 2 vs. Session 3</td>
<td>1</td>
<td>19.228</td>
<td>&lt;.001</td>
<td>.30</td>
</tr>
<tr>
<td>Session 3 vs. Session 4</td>
<td>1</td>
<td>1.480</td>
<td>.225</td>
<td>.09</td>
</tr>
</tbody>
</table>

Similarly, inter-rater agreement in absolute terms increased over time (Table 6.4).

The percentage of exact agreements (where a pair of raters gave exactly the same score to the same response) rose from 33.9% before training to 34.6% following training, and continued to rise in the subsequent scoring session finally reaching a value of 38.5%. The percentage of exact agreements does not take into account the relative magnitude of different types of disagreement between raters however; a disagreement of one point between raters (perhaps acceptable) counts the same as a disagreement of five points (a serious problem). To account for this issue as well as for agreement expected by chance, linearly-weighted Fleiss kappa values were calculated for each scoring session. Fleiss
kappa values also generally increased over time (with the exception of session 3), showing a similar trend to the percentage of exact agreements; the general level of agreement across sessions would be considered moderate in magnitude (Landis & Koch, 1977). So, with training and increasing experience, raters were somewhat more likely to award the same score for the same performance.

Table 6.4
Agreement indices within scoring sessions

<table>
<thead>
<tr>
<th>Session</th>
<th>Exact agreements (%)</th>
<th>Linearly-weighted Fleiss' Kappa</th>
<th>Kappa SE</th>
<th>Kappa 95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>33.9%</td>
<td>0.353</td>
<td>0.030</td>
<td>0.293 to 0.413</td>
</tr>
<tr>
<td>Session 2</td>
<td>34.6%</td>
<td>0.417</td>
<td>0.029</td>
<td>0.36 to 0.473</td>
</tr>
<tr>
<td>Session 3</td>
<td>35.2%</td>
<td>0.400</td>
<td>0.032</td>
<td>0.336 to 0.464</td>
</tr>
<tr>
<td>Session 4</td>
<td>38.3%</td>
<td>0.430</td>
<td>0.032</td>
<td>0.366 to 0.494</td>
</tr>
</tbody>
</table>

6.1.3 Within-rater consistency in scoring. In addition to the between-rater variation in severity observed from the many-facet Rasch analysis, considerable within-rater variability in severity was seen across scoring sessions for some individual raters (Table 6.1). For example, the severity of rater 109 varied from 1.16 logits in the first session (i.e., relatively strict) to -1.06 logits in the third session (i.e., relatively lenient). In addition to rater 109, raters 108 and 113 showed changes in severity over time on the order of 1 logit or more. Rater 108 started out being the most lenient rater (-1.81 logits), then adjusted towards average severity in session 2 (-0.52), eventually drifting away from the mean and becoming more lenient in sessions 3 and 4 (-1.19 and -1.10 logits, respectively). Rater 113 started out close to average severity (0.22 logits), but became
much more severe following training (1.60 logits) and finally drifted back towards the mean in the remaining sessions.

Moreover, other individuals were consistently severe or lenient compared to other raters. For example, rater 120 had a severity value of -0.95 logits in the first session (making her the second most lenient rater), and in subsequent sessions her severity ranged from -0.64 (session 4) to -0.89 (session 3), showing no change in severity when measurement error is taken into account. Finally, severity measures for a number of other individuals remained close to the mean throughout the study (e.g., raters 101, 102, 123).

To summarize, while little difference in overall severity was seen over time, there were considerable differences seen in individual patterns.

Within-rater consistency was also investigated by calculating the degree that each rater's scoring patterns fit the predictions of the Rasch model; that is, observed scores should be similar to the scores expected for a given examinee/prompt/rater combination. Model fit was examined in terms of infit mean square, which is calculated using standardized residuals (based on observed minus expected values) that have been squared and then averaged for each rater, with the average weighted to be less sensitive to outlier values (Eckes, 2011). Infit mean square values can range from zero to infinity; a value of 1 indicates agreement with the predictions of the Rasch model. Values less than one indicate that the rater is more consistent than expected (overfitting), while values greater than one indicate that the rater's scores are less consistent than predicted (misfitting). Values within the range of 0.75-1.3 have been suggested to indicate acceptable scoring consistency (Bond & Fox, 2007), although Linacre (2007a) has suggested that values within the range of 0.5 to 1.5 are adequate. Table 6.5 provides infit mean square values...
for all raters as well as a standardized infit statistic (z-score) which has an expected value of zero and standard deviation of one.

Table 6.5

Infit mean square values from multifaceted Rasch analysis

<table>
<thead>
<tr>
<th>Rater</th>
<th>Session 1</th>
<th>Session 2</th>
<th>Session 3</th>
<th>Session 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Infit</td>
<td>Infit</td>
<td>Infit</td>
<td>Infit</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>$MnSq$</td>
<td>$z$</td>
<td>$N$</td>
</tr>
<tr>
<td>101</td>
<td>100 0.9 -0.6</td>
<td>100 0.93 -0.4</td>
<td>100 0.6 -3.2</td>
<td>100 0.73 -2.1</td>
</tr>
<tr>
<td>102</td>
<td>100 0.62 -3</td>
<td>100 0.62 -3.1</td>
<td>100 0.71 -2.3</td>
<td>100 0.69 -2.4</td>
</tr>
<tr>
<td>103</td>
<td>100 0.99 0</td>
<td>100 1.1 0.7</td>
<td>100 0.91 -0.5</td>
<td>100 0.94 -0.3</td>
</tr>
<tr>
<td>105</td>
<td>100 0.95 -0.3</td>
<td>100 0.95 -0.3</td>
<td>100 0.79 -1.5</td>
<td>100 0.88 -0.8</td>
</tr>
<tr>
<td>107</td>
<td>100 0.96 -0.2</td>
<td>100 1.03 0.2</td>
<td>100 0.93 -0.4</td>
<td>100 0.92 -0.5</td>
</tr>
<tr>
<td>108</td>
<td>100 0.84 -1.1</td>
<td>100 1.16 1.1</td>
<td>100 1.26 1.7</td>
<td>100 1.27 1.8</td>
</tr>
<tr>
<td>109</td>
<td>100 1.18 1.2</td>
<td>100 0.85 -1</td>
<td>100 0.97 -0.1</td>
<td>100 0.85 -1.1</td>
</tr>
<tr>
<td>111</td>
<td>100 1.28 1.9</td>
<td>100 0.87 -0.9</td>
<td>100 1.45 2.8</td>
<td>100 0.87 -0.9</td>
</tr>
<tr>
<td>112</td>
<td>100 1.15 1</td>
<td>100 0.95 -0.3</td>
<td>100 1.67 4</td>
<td>100 1.18 1.2</td>
</tr>
<tr>
<td>113</td>
<td>99 0.98 0</td>
<td>100 1.04 0.3</td>
<td>100 1.14 1</td>
<td>99 0.62 -3.1</td>
</tr>
<tr>
<td>115</td>
<td>100 0.74 -2</td>
<td>100 0.91 -0.6</td>
<td>100 0.75 -1.9</td>
<td>100 0.47 -4.7</td>
</tr>
<tr>
<td>116</td>
<td>100 1.2 1.3</td>
<td>100 0.93 -0.4</td>
<td>100 0.8 -1.5</td>
<td>100 1.11 0.7</td>
</tr>
<tr>
<td>117</td>
<td>99 1.78 4.6</td>
<td>98 1.24 1.6</td>
<td>100 1.24 1.6</td>
<td>100 1.04 0.3</td>
</tr>
<tr>
<td>119</td>
<td>100 1.17 1.1</td>
<td>100 1.05 0.4</td>
<td>100 0.83 -1.2</td>
<td>100 1.22 1.5</td>
</tr>
<tr>
<td>120</td>
<td>96 1.13 0.9</td>
<td>100 1.13 0.9</td>
<td>100 1.16 1.1</td>
<td>100 0.88 -0.8</td>
</tr>
<tr>
<td>121</td>
<td>100 0.95 -0.2</td>
<td>100 1 0</td>
<td>100 1.34 2.2</td>
<td>100 1.24 1.6</td>
</tr>
<tr>
<td>122</td>
<td>100 1.14 1</td>
<td>100 1.35 2.2</td>
<td>100 0.98 0</td>
<td>98 0.98 0</td>
</tr>
<tr>
<td>123</td>
<td>100 0.79 -1.5</td>
<td>100 0.74 -2</td>
<td>100 0.78 -1.6</td>
<td>100 0.77 -1.7</td>
</tr>
<tr>
<td>124</td>
<td>100 0.69 -2.5</td>
<td>100 1.04 0.3</td>
<td>100 0.77 -1.7</td>
<td>100 1.03 0.2</td>
</tr>
<tr>
<td>125</td>
<td>100 1.4 2.5</td>
<td>100 1.19 1.3</td>
<td>100 1.08 0.5</td>
<td>100 0.95 -0.3</td>
</tr>
</tbody>
</table>

Note: Misfitting raters are marked in gray, while overfitting raters are enclosed in boxes. Misfitting raters were those with infit mean square values greater 1.3, while overfitting raters had values less than 0.75.
Infit mean square values varied between raters and across scoring sessions, but were generally within acceptable bounds. A total of 78 observations (97.5%) fell inside the bounds proposed by Linacre (2007a), while 66 observations (82.5%) fell within the more restrictive bounds cited by Bond and Fox (2007). Moreover, of the 14 observations that fell outside of these more restrictive bounds, 11 were overfitting indicating that the raters were more consistent than expected. Similar to the severity data, infit values observed in the first scoring session were already generally within acceptable bounds, before raters underwent training or had access to the scoring rubric.

Figure 6.3. Rater scoring consistency in terms of infit mean square. Each line represents a single rater.

In addition, training and experience appeared to have little effect on infit measures, much like the pattern seen in rater severity. Visual inspection of Figure 6.3 suggests that in session 2 there may have been some minor clustering of infit mean square values around the expected value of one with an opposite trend towards greater
variability in session 3, but on closer examination these trends appear to be the result of variation in a few individuals and probably do not represent a general response to training and/or scoring experience.

Rater consistency was also examined in terms of rater bias towards specific examinees, defined as instances where the score awarded to an examinee is higher or lower than would be predicted based on the ability of the examinee and the severity of the rater. When examining scores for bias the FACETS software package also provides a $t$-statistic testing the hypothesis that the observed bias is due only to measurement error, calculated as the magnitude of bias (observed minus expected score, in logits) divided by the standard error of the bias measure (Linacre, 2007a). For the purpose of this analysis, scores with bias measures greater than 2.0 standard errors were counted.

The number of scores showing bias, for each rater and across sessions, is given in Table 6.6. The frequency of biased scores was quite low in all scoring sessions, with an average of 1.9 instances out of 100 scores produced by each rater in each scoring session. Instances of bias ranged from a low of zero to a high of five; there were no noticeable trends in the frequency of bias over time. Like the results for infit mean square, the bias data suggested that raters were already scoring fairly consistently before the rater training or being given access to the scoring rubric.
### Table 6.6

*Frequency of scores showing rater-by-examinee bias*

<table>
<thead>
<tr>
<th>Scoring Session</th>
<th>Rater</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.8</td>
<td>0.96</td>
</tr>
<tr>
<td>102</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1.0</td>
<td>0.82</td>
</tr>
<tr>
<td>103</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>3.8</td>
<td>1.50</td>
</tr>
<tr>
<td>105</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.5</td>
<td>1.29</td>
</tr>
<tr>
<td>107</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td>0.50</td>
</tr>
<tr>
<td>108</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>3.3</td>
<td>1.50</td>
</tr>
<tr>
<td>109</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2.0</td>
<td>1.41</td>
</tr>
<tr>
<td>111</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.82</td>
</tr>
<tr>
<td>112</td>
<td>3a</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>3.0</td>
<td>1.41</td>
</tr>
<tr>
<td>113</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2b</td>
<td>0</td>
<td>2.3</td>
<td>0.96</td>
</tr>
<tr>
<td>115</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.50</td>
</tr>
<tr>
<td>116</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>1.8</td>
<td>1.50</td>
</tr>
<tr>
<td>117</td>
<td>5a</td>
<td>2b</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>3.8</td>
<td>1.26</td>
</tr>
<tr>
<td>119</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1.8</td>
<td>1.50</td>
</tr>
<tr>
<td>120</td>
<td>2c</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2.3</td>
<td>1.26</td>
</tr>
<tr>
<td>121</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1.8</td>
<td>0.50</td>
</tr>
<tr>
<td>122</td>
<td>3</td>
<td>3</td>
<td>1b</td>
<td>0</td>
<td>1</td>
<td>2.5</td>
<td>1.00</td>
</tr>
<tr>
<td>123</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.58</td>
</tr>
<tr>
<td>124</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1.8</td>
<td>2.06</td>
</tr>
<tr>
<td>125</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2.0</td>
<td>1.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>2.2</th>
<th>1.9</th>
<th>1.7</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>1.54</td>
<td>1.25</td>
<td>1.42</td>
<td>1.57</td>
</tr>
</tbody>
</table>

*Note.* Frequency is calculated as the number of biased scores per scoring session, with each session having a total of 100 scores (except where marked).

a 99 scores used
b 98 scores used
c 96 scores used

**6.1.4 Accuracy of scores.** While the analyses reported so far provide information regarding raters' scoring patterns, they do not necessarily address the issue of whether the scores accurately reflect the scale used in the study. To examine this issue, scores produced by each rater in each session were compared to reference scores for the same responses using Pearson product-moment correlations (Figure 6.4, Table 6.7). Correlations between raters' scores and reference scores ranged from .62 to .90, with
most values falling in the range of .70 to .90. Inspection of Figure 6.4 suggests there was a slight increase in accuracy following training; a one-way repeated measures ANOVA performed following a Fisher $z$ transformation detected a statistically significant difference across sessions, $F(3,57) = 12.287, p < .001, \eta^2 = .393, \omega^2 = .121$. Pairwise contrasts indicated that values increased significantly from session 1 to session 2 ($F(1,19) = 21.181, p < .001, r = .726$) but no difference was seen between session 2 and session 3, or session 3 and session 4. Once again, rater scoring performance was reasonably good across all sessions but nonetheless improved following training, where a large effect (.726) was apparent.

**Figure 6.4.** Rater accuracy in terms of Pearson correlations with reference scores. Each line represents a single rater.
Table 6.7

*Rater accuracy in terms of Pearson correlation with reference scores*

<table>
<thead>
<tr>
<th>Scoring Session</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Mean</th>
<th>SD</th>
<th>Low-High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 101</td>
<td>.81</td>
<td>.85</td>
<td>.90</td>
<td>.85</td>
<td>.853</td>
<td>.037</td>
<td>.81-.90</td>
</tr>
<tr>
<td>Rater 102</td>
<td>.86</td>
<td>.85</td>
<td>.84</td>
<td>.86</td>
<td>.789</td>
<td>.010</td>
<td>.84-.86</td>
</tr>
<tr>
<td>Rater 103</td>
<td>.72</td>
<td>.70</td>
<td>.73</td>
<td>.75</td>
<td>.753</td>
<td>.021</td>
<td>.70-.75</td>
</tr>
<tr>
<td>Rater 105</td>
<td>.79</td>
<td>.78</td>
<td>.78</td>
<td>.77</td>
<td>.792</td>
<td>.008</td>
<td>.77-.79</td>
</tr>
<tr>
<td>Rater 107</td>
<td>.80</td>
<td>.78</td>
<td>.80</td>
<td>.83</td>
<td>.759</td>
<td>.021</td>
<td>.78-.83</td>
</tr>
<tr>
<td>Rater 108</td>
<td>.71</td>
<td>.75</td>
<td>.71</td>
<td>.69</td>
<td>.714</td>
<td>.025</td>
<td>.69-.75</td>
</tr>
<tr>
<td>Rater 109</td>
<td>.67</td>
<td>.76</td>
<td>.69</td>
<td>.73</td>
<td>.747</td>
<td>.040</td>
<td>.67-.76</td>
</tr>
<tr>
<td>Rater 111</td>
<td>.74</td>
<td>.82</td>
<td>.80</td>
<td>.78</td>
<td>.779</td>
<td>.034</td>
<td>.74-.82</td>
</tr>
<tr>
<td>Rater 112</td>
<td>.66</td>
<td>.85</td>
<td>.77</td>
<td>.83</td>
<td>.730</td>
<td>.085</td>
<td>.66-.85</td>
</tr>
<tr>
<td>Rater 113</td>
<td>.64</td>
<td>.74</td>
<td>.62</td>
<td>.74</td>
<td>.743</td>
<td>.064</td>
<td>.62-.74</td>
</tr>
<tr>
<td>Rater 115</td>
<td>.75</td>
<td>.81</td>
<td>.80</td>
<td>.85</td>
<td>.790</td>
<td>.041</td>
<td>.75-.85</td>
</tr>
<tr>
<td>Rater 116</td>
<td>.76</td>
<td>.80</td>
<td>.80</td>
<td>.76</td>
<td>.754</td>
<td>.023</td>
<td>.76-.80</td>
</tr>
<tr>
<td>Rater 117</td>
<td>.67</td>
<td>.73</td>
<td>.77</td>
<td>.75</td>
<td>.779</td>
<td>.043</td>
<td>.67-.77</td>
</tr>
<tr>
<td>Rater 119</td>
<td>.78</td>
<td>.84</td>
<td>.87</td>
<td>.84</td>
<td>.785</td>
<td>.038</td>
<td>.78-.87</td>
</tr>
<tr>
<td>Rater 120</td>
<td>.70</td>
<td>.75</td>
<td>.74</td>
<td>.79</td>
<td>.781</td>
<td>.037</td>
<td>.70-.79</td>
</tr>
<tr>
<td>Rater 121</td>
<td>.71</td>
<td>.86</td>
<td>.85</td>
<td>.86</td>
<td>.781</td>
<td>.073</td>
<td>.71-.86</td>
</tr>
<tr>
<td>Rater 122</td>
<td>.67</td>
<td>.77</td>
<td>.78</td>
<td>.75</td>
<td>.789</td>
<td>.050</td>
<td>.67-.78</td>
</tr>
<tr>
<td>Rater 123</td>
<td>.81</td>
<td>.84</td>
<td>.84</td>
<td>.86</td>
<td>.783</td>
<td>.021</td>
<td>.81-.86</td>
</tr>
<tr>
<td>Rater 124</td>
<td>.73</td>
<td>.75</td>
<td>.73</td>
<td>.70</td>
<td>.753</td>
<td>.021</td>
<td>.70-.75</td>
</tr>
<tr>
<td>Rater 125</td>
<td>.69</td>
<td>.80</td>
<td>.81</td>
<td>.81</td>
<td>.778</td>
<td>.059</td>
<td>.69-.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>.734</th>
<th>.792</th>
<th>.782</th>
<th>.790</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>.060</td>
<td>.047</td>
<td>.066</td>
<td>.055</td>
</tr>
<tr>
<td>Low-High</td>
<td>.64-.86</td>
<td>.70-.86</td>
<td>.62-.90</td>
<td>.69-.86</td>
</tr>
</tbody>
</table>

Agreement with the reference scores increased following training, with average values for Cohen's kappa (linearly weighted) increasing from .47 to .52 (Table 6.8). This difference in Cohen’s kappa was statistically significant (one-way repeated measures ANOVA $F(3,57) = 13.506, p < .001, \eta^2 = .415, \omega^2 = .140$), with agreement increasing significantly with each session (Table 6.9). In addition, individual raters showing the lowest levels of agreement with the reference scores in session 1 improved substantially after training. The four raters who had kappa values less than .40 before training
improved by .18 on average, and rater 108 actually more than doubled the level of agreement, from .22 to .51. Conversely, one individual (rater 113) actually worsened after training (dropping from .42 to .26), but individuals with the poorest agreement at the beginning generally tended to show the most improvement following training.

Table 6.8

*Rater accuracy in terms of agreement (Cohen's Kappa) with reference scores*

<table>
<thead>
<tr>
<th>Rater</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Mean</th>
<th>SD</th>
<th>Low-High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 101</td>
<td>.57</td>
<td>.58</td>
<td>.63</td>
<td>.56</td>
<td>.59</td>
<td>.03</td>
<td>.56-.63</td>
</tr>
<tr>
<td>Rater 102</td>
<td>.63</td>
<td>.63</td>
<td>.58</td>
<td>.61</td>
<td>.62</td>
<td>.02</td>
<td>.58-.63</td>
</tr>
<tr>
<td>Rater 103</td>
<td>.56</td>
<td>.41</td>
<td>.44</td>
<td>.52</td>
<td>.48</td>
<td>.07</td>
<td>.41-.56</td>
</tr>
<tr>
<td>Rater 105</td>
<td>.59</td>
<td>.49</td>
<td>.57</td>
<td>.56</td>
<td>.55</td>
<td>.04</td>
<td>.49-.59</td>
</tr>
<tr>
<td>Rater 107</td>
<td>.51</td>
<td>.45</td>
<td>.54</td>
<td>.65</td>
<td>.54</td>
<td>.08</td>
<td>.45-.65</td>
</tr>
<tr>
<td>Rater 108</td>
<td>.22</td>
<td>.51</td>
<td>.46</td>
<td>.41</td>
<td>.40</td>
<td>.13</td>
<td>.22-.51</td>
</tr>
<tr>
<td>Rater 109</td>
<td>.34</td>
<td>.53</td>
<td>.41</td>
<td>.50</td>
<td>.44</td>
<td>.09</td>
<td>.34-.53</td>
</tr>
<tr>
<td>Rater 111</td>
<td>.47</td>
<td>.62</td>
<td>.53</td>
<td>.55</td>
<td>.54</td>
<td>.06</td>
<td>.47-.62</td>
</tr>
<tr>
<td>Rater 112</td>
<td>.42</td>
<td>.57</td>
<td>.50</td>
<td>.56</td>
<td>.51</td>
<td>.07</td>
<td>.42-.57</td>
</tr>
<tr>
<td>Rater 113</td>
<td>.42</td>
<td>.26</td>
<td>.36</td>
<td>.46</td>
<td>.37</td>
<td>.08</td>
<td>.26-.46</td>
</tr>
<tr>
<td>Rater 115</td>
<td>.52</td>
<td>.57</td>
<td>.52</td>
<td>.63</td>
<td>.56</td>
<td>.06</td>
<td>.52-.63</td>
</tr>
<tr>
<td>Rater 116</td>
<td>.48</td>
<td>.42</td>
<td>.43</td>
<td>.45</td>
<td>.45</td>
<td>.03</td>
<td>.42-.48</td>
</tr>
<tr>
<td>Rater 117</td>
<td>.43</td>
<td>.52</td>
<td>.52</td>
<td>.53</td>
<td>.50</td>
<td>.05</td>
<td>.43-.53</td>
</tr>
<tr>
<td>Rater 119</td>
<td>.53</td>
<td>.63</td>
<td>.69</td>
<td>.64</td>
<td>.62</td>
<td>.07</td>
<td>.53-.69</td>
</tr>
<tr>
<td>Rater 120</td>
<td>.39</td>
<td>.53</td>
<td>.51</td>
<td>.51</td>
<td>.49</td>
<td>.06</td>
<td>.39-.53</td>
</tr>
<tr>
<td>Rater 121</td>
<td>.46</td>
<td>.62</td>
<td>.56</td>
<td>.60</td>
<td>.56</td>
<td>.07</td>
<td>.46-.62</td>
</tr>
<tr>
<td>Rater 122</td>
<td>.46</td>
<td>.44</td>
<td>.47</td>
<td>.53</td>
<td>.48</td>
<td>.04</td>
<td>.44-.53</td>
</tr>
<tr>
<td>Rater 123</td>
<td>.63</td>
<td>.64</td>
<td>.65</td>
<td>.68</td>
<td>.65</td>
<td>.02</td>
<td>.63-.68</td>
</tr>
<tr>
<td>Rater 124</td>
<td>.47</td>
<td>.44</td>
<td>.48</td>
<td>.51</td>
<td>.48</td>
<td>.03</td>
<td>.44-.51</td>
</tr>
<tr>
<td>Rater 125</td>
<td>.38</td>
<td>.49</td>
<td>.59</td>
<td>.58</td>
<td>.51</td>
<td>.10</td>
<td>.38-.59</td>
</tr>
</tbody>
</table>

Note: Individuals with the lowest pre-training values are enclosed in boxes.
Table 6.9

*Effect sizes for changes across sessions in agreement (Cohen's kappa) with reference scores.*

<table>
<thead>
<tr>
<th>Contrast</th>
<th>df</th>
<th>F</th>
<th>Significance (p)</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1 vs. Session 2</td>
<td>1</td>
<td>21.919</td>
<td>&lt;.001</td>
<td>.73</td>
</tr>
<tr>
<td>Session 2 vs. Session 3</td>
<td>1</td>
<td>5.103</td>
<td>.036</td>
<td>.46</td>
</tr>
<tr>
<td>Session 3 vs. Session 4</td>
<td>1</td>
<td>7.752</td>
<td>.012</td>
<td>.54</td>
</tr>
</tbody>
</table>

6.1.5 **Rater behavior while scoring.** Rater behavior while scoring was investigated in two ways. First, the number of times the exemplar performances were checked while scoring was counted. The exemplars served as examples of performance at a given score level and exemplar use was seen as an indicator of the degree to which raters recalibrated during the scoring session. Second, the time required to reach a scoring decision was measured from the completion of scoring one response to the completion of scoring the next response. Longer times might indicate that the rater was having difficulty reaching a decision, and/or that more time was taken in re-listening to the examinee response or exemplars.

Values for frequency of exemplar use are provided in Table 6.10 and illustrated in Figure 6.5. Thirteen of the raters tended to view the exemplars the minimum twelve times per scoring session. (Raters were required to review six exemplars for each prompt to unlock the responses to be scored.) Within this group, four individuals reviewed the exemplars more than twelve times during the first session, but reverted to the minimum in subsequent sessions (R103, R109, R117, R120). Of the seven raters who regularly checked the exemplars more than twelve times per session, there was no clear pattern across time. Five of these seven raters showed a decrease in exemplar use after training.
(R101, R102, R122, R123, R124), but usage in the following sessions showed no consistent pattern (Figure 6.5). Two raters, 102 and 122, showed a dramatic and nearly identical decrease in exemplar use after training, but in both cases exemplar use reached the highest levels in later sessions. Figure 6.5 suggests that use of exemplars dropped somewhat following training, which might be expected given that the exemplars were the only scoring aid provided in the first session. However, a one-way repeated measures ANOVA showed no difference in the number of times the exemplars were checked in each scoring session, \((F(3, 57) = 2.670, p = .056, \eta^2 = .123, \omega^2 = 0; \) data transformed using an inverse transformation due to positive skew in the raw data). Overall, the majority of raters made little or no use of the exemplars once the scoring session was unlocked. For the minority who did use the exemplars while scoring, usage patterns were highly variable.
Figure 6.5. The number of times the exemplars were reviewed during a scoring session. Each line represents a single rater. Multiple lines overlap at approximately 12 exemplar checks per session.

Table 6.10

The number of times the exemplars were checked while scoring

<table>
<thead>
<tr>
<th>Scoring Session</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Mean</th>
<th>SD</th>
<th>Low-High</th>
</tr>
</thead>
<tbody>
<tr>
<td>R101</td>
<td>37</td>
<td>30</td>
<td>37</td>
<td>27</td>
<td>32.8</td>
<td>5.06</td>
<td>27-37</td>
</tr>
<tr>
<td>R102</td>
<td>43</td>
<td>21</td>
<td>47</td>
<td>38</td>
<td>37.3</td>
<td>11.44</td>
<td>21-47</td>
</tr>
<tr>
<td>R103</td>
<td>15</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12.8</td>
<td>1.50</td>
<td>12-15</td>
</tr>
<tr>
<td>R105</td>
<td>22</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>14.8</td>
<td>4.86</td>
<td>12-22</td>
</tr>
<tr>
<td>R107</td>
<td>21</td>
<td>29</td>
<td>28</td>
<td>15</td>
<td>23.3</td>
<td>6.55</td>
<td>15-29</td>
</tr>
<tr>
<td>R108</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12.0</td>
<td>0.00</td>
<td>12-12</td>
</tr>
<tr>
<td>R109</td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12.5</td>
<td>1.00</td>
<td>12-14</td>
</tr>
<tr>
<td>R111</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>12.3</td>
<td>0.50</td>
<td>12-13</td>
</tr>
<tr>
<td>R112</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12.0</td>
<td>0.00</td>
<td>12-12</td>
</tr>
<tr>
<td>R113</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>12.3</td>
<td>0.50</td>
<td>12-13</td>
</tr>
<tr>
<td>R115</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>15</td>
<td>12.8</td>
<td>1.50</td>
<td>12-15</td>
</tr>
<tr>
<td>R116</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>12.8</td>
<td>0.96</td>
<td>12-14</td>
</tr>
<tr>
<td>R117</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12.3</td>
<td>0.50</td>
<td>12-13</td>
</tr>
<tr>
<td>R119</td>
<td>17</td>
<td>21</td>
<td>25</td>
<td>20</td>
<td>20.8</td>
<td>3.30</td>
<td>17-25</td>
</tr>
<tr>
<td>R120</td>
<td>19</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>14.0</td>
<td>3.37</td>
<td>12-19</td>
</tr>
<tr>
<td>R121</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12.0</td>
<td>0.00</td>
<td>12-12</td>
</tr>
<tr>
<td>R122</td>
<td>43</td>
<td>22</td>
<td>27</td>
<td>53</td>
<td>36.3</td>
<td>14.31</td>
<td>22-53</td>
</tr>
<tr>
<td>R123</td>
<td>21</td>
<td>16</td>
<td>14</td>
<td>16</td>
<td>16.8</td>
<td>2.99</td>
<td>14-21</td>
</tr>
<tr>
<td>R124</td>
<td>28</td>
<td>27</td>
<td>19</td>
<td>27</td>
<td>25.3</td>
<td>4.19</td>
<td>19-28</td>
</tr>
<tr>
<td>R125</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13.0</td>
<td>0.00</td>
<td>13-13</td>
</tr>
</tbody>
</table>

| Mean | 19.5 | 16.2 | 18.0 | 17.9 |
| SD   | 10.33| 6.35 | 9.88 | 10.82|
| Low-High | 12-43| 12-30| 12-47| 12-53|
Figure 6.6 summarizes the results regarding the time required to reach a scoring decision; exact values are provided in Appendix N. As with exemplar use, there is little apparent difference in decision times across scoring sessions. This conclusion is also supported by the results of a one-way repeated measures ANOVA conducted on the mean decision times for each rater, which found no difference between sessions \( (F(3,57) = 0.447, p = .720, \eta^2 = .023, \omega^2 = 0) \). The majority of raters took from 60 to 80 seconds to listen to the examinee response and reach a score. Figure 6.6 suggests that a smaller group of five or six raters (depending on the session) typically took more time to reach a decision, usually in the range of approximately 90 to 100 seconds. Analysis of the distribution of scoring times for these raters suggests that the difference might be a result of these individuals being more likely to listen to the examinee response a second time. That is, the decision time values for each rater showed a bimodal distribution, with the first (and invariably largest) mode occurring at 60-80 seconds and indicating decisions for which the 45 second response was played once, and a smaller mode (often at about 120 seconds) probably indicating decisions where either the response was played twice or an exemplar was reviewed. (Specific examples of such distributions are provided in section 6.2.1 below.)
6.2 Scoring Behavior and Cognition of More-Proficient and Less-Proficient Raters

6.2.1 Scoring patterns. As described in chapter 5, individuals were chosen for further analysis who showed scoring patterns which were more or less desirable in terms of severity, consistency, and accuracy. These individuals were divided into three groups of raters: more-proficient, less-proficient, and developing. The specific criteria for membership in each group are summarized below:

1. More-proficient raters were those who in every session showed severity values within +/-0.5 logits of the mean, infit mean square values of 0.9 or lower, and accuracy values (correlation with reference scores) greater than .8.
2. Less-proficient raters were those who showed wide variation in severity values across sessions, with at least one occasion where severity was greater than 1 logit away from the mean, and where accuracy was consistently .75 or less.

3. Developing raters were those with infit mean square values that consistently decreased over time, with a total decrease of at least 0.2, and accuracy values that consistently increased over time, with a total increase of at least .1.

Three individuals were placed into each group; the specific scoring patterns of these individuals can be compared in Appendix O.

In addition to the features used to define the rater categories, other measures of scoring performance were also compared between the groups. First, within-rater scoring consistency across scoring sessions was investigated in terms of agreement in scores given to a set of 16 responses repeated in all four scoring sessions. For each rater, the percentage of exact agreements were compared for all possible pairings of scoring session (six in all), and an overall within-rater agreement percentage was calculated; these values are shown in Figure 6.7. As expected, the more-proficient raters showed the highest levels of consistency across sessions, with agreement values ranging from 51% to 66%. The less-proficient and improving rater groups had lower values, generally in the range of 30% to 40% except for rater 117 who reached a value of 42%.
Figure 6.7. The percentage of within-rater exact agreement across scoring sessions. The values are based on 16 responses repeated in all sessions.

Second, the scale use of the individual raters was also compared to examine the degree to which raters' application of the rating scale was similar or different. As shown earlier in Figure 6.1, the FACETS Rasch analysis software aligns the rating scale against the logit scale used to measure examinee ability and other facets of the testing context. In this alignment, the boundary between adjacent scores (such as a three and four, for example) is associated with a logit value which corresponds to the level of examinee ability for which an examinee has an equal chance on being awarded either the higher or lower score (if all other facets of the testing situation are of average difficulty). In the case of Figure 6.1, the alignment of the rating scale to the logit scale was made using all score from all raters pooled together; this alignment is recreated as the “pooled scale” shown at the top of Figure 6.8. In more technical terms, this pooled scale was the result of specifying a rating scale model (Andrich, 1978) in the analysis, which assumes all raters
share a common understanding of the scale (Linacre; 2007a; McNamara, 1996). In addition, the data were also analyzed using a *partial credit model* (Masters, 1982) which does not assume a common interpretation of the rating scale; in this case raters are assumed to have idiosyncratic understandings of the rating scale and FACETS produces separate rating scale representations for each rater. Results of this second analysis are shown in Figure 6.8, where scale representations for individual raters are plotted.

![Figure 6.8](image)

*Figure 6.8.* Scale use of more-proficient, less-proficient, and improving raters. The tick marks on the horizontal bars indicate boundary points between raw score values; that is, the point at which a response has an equal probability of getting the higher or lower score.
The pooled scale at the top of the figure arguably represents the best estimate of the measurement scale as operationalized by the raters as a group. More-proficient raters best approximated the pooled scale and generally showed relatively even intervals between raw score values, a desirable trait if raw scores are intended to approximate an interval scale. The range of the scale was somewhat compressed among the more-proficient raters, however, and reluctance to award extremely high or low scores (central tendency) can result in Rasch model overfit (Eckes, 2011). The restriction in range seen here may have therefore contributed to the relatively low Rasch model infit values seen for this group. On the whole, scale use for the less-proficient raters appeared least like the scale produced using all data, with raters 108 and 109 both showing a relatively large degree of unevenness in scale increments. Improving raters generally seemed to be intermediate in their application of the rating scale. Values for the scale boundaries along with standard errors are provided in Appendix P.

6.2.1 Behavior while scoring. As with research question 1, behavior while scoring was examined in terms of use of the exemplars and the time required to reach a scoring decision. As was reported earlier, there was little consistent difference across scoring sessions in the frequency of exemplar use or the average time taken to reach a scoring decision; rather, more variation was seen between individuals, which might suggest that different individuals had separate styles or approaches to the scoring task. Therefore, in the current analysis values were averaged across all four scoring sessions for each rater.

Figure 6.9 shows exemplar use for more-proficient, less-proficient, and improving raters. A clear difference is visible between the more-proficient raters, who viewed the
exemplars from 14 to 47 times per session, and the other groups where exemplar use was close to the minimum required to unlock the scoring session (12 views). It was also apparent from timestamp data recorded by the scoring instrument that the more-proficient raters were checking the exemplars during the scoring session, rather than repeatedly playing the examples to prepare for scoring. Although no causal link can be made between exemplar use and scoring patterns, the more-proficient raters essentially seemed to be re-calibrating periodically during the session, which could have contributed to the high levels of score accuracy seen during the first scoring session, when no other scoring aids were available.
Figure 6.9. Use of exemplars while scoring for proficient, non-proficient, and developing raters.
The more-proficient raters also took longer to make scoring decisions, with overall average times ranging from 83-100 seconds (Figure 6.10). The lone exception was rater 113 from the less-proficient group, who took an average of 95 seconds to reach a decision. Closer analysis of the data, however, suggests that the behavior of rater 113 may be qualitatively different from the raters in the more-proficient group. Figure 6.11 shows the distribution of decision times for each rater; all of the more-proficient raters show a bimodal distribution with a large peak at about 60 seconds and a smaller peak at about 120 seconds. It seems likely that this second mode represents decisions where either the 45-second examinee response was played a second time, or one of the exemplar recordings was reviewed. In contrast, the distribution of decision times for rater 113 shows no obvious second mode but rather a single mode with a greater spread in values; this rater may have replayed examinee recordings, but another possibility is that the longer decision times resulted from checking the rubric, difficulty in reaching a decision, or mental distraction.

Like rater 113, the other two less-proficient raters showed a unimodal distribution of decision times. The distribution of decision times for the developing raters was intermediate, with raters 117 and 125 showing small secondary peaks, and rater 120 showing an essentially unimodal distribution. Compared with the other groups, the more proficient raters also tended to show a relatively long tail in the distribution of times, with decision times in the range of 150-400 seconds being more common than the other groups, although still representing a small proportion of all decisions. It is unclear what kinds of behavior produced these extended decisions times, but time intervals of this
length would allow two or more exemplars to be checked, and/or the examinee response to be played additional times.

**Figure 6.10.** The average time required to make a scoring decision for proficient, non-proficient, and developing raters.

*Figure 6.10.* The average time required to make a scoring decision for proficient, non-proficient, and developing raters.
Figure 6.11. Distribution of decision times for more-proficient, less-proficient, and improving raters. The vertical axis is the number of decisions while the horizontal axis is the decision time, in 15-second intervals.
In addition to the time required to reach a decision, the number of breaks taken while scoring are shown in Figure 6.12. Breaks were excluded from the decision time results described above, and were defined as periods of five minutes or more occurring either (a) between the end of scoring one response and starting the next or (b) while scoring a response (i.e., decision times greater than five minutes were counted as breaks). No clear difference was seen between groups in the number of breaks taken per scoring session, although two of the three highest values belonged to raters in the more-proficient group (raters 102 and 123). In any case, the number of breaks was on average about seven or fewer per session, which is perhaps a bit surprising given that the scoring for each session typically required three to four hours to complete (not including time spent doing stimulated recalls or reviewing the exemplars and rubric prior to scoring).

To summarize, the more-proficient raters viewed the exemplars more often and took more time to score compared to the other groups. While it cannot be proven that these behaviors lead to more desirable scoring patterns, the overall impression is that perhaps the more-proficient raters may have taken greater care in making their scoring decisions.
Figure 6.12. The number of breaks (5 minutes or longer) taken while scoring by proficient, non-proficient, and developing raters.
6.2.3 Cognition while scoring. As described in chapter 5, rater cognition was investigated using rater comments collected via stimulated recall. In each scoring session, raters first scored 50 responses for each prompt; these were then followed by ten verbal reports in which a response was first scored and then played again as the rater thought through the scoring process out loud. Transcriptions from ten (half) of the verbal reports collected in scoring session 1 (before training), session 2 (after training), and session 4 (at the conclusion of the study) were coded for each of the three groups of raters. Comments regarding specific language features were coded into categories of “delivery,” “language use,” and “topical development” as used in the TOEFL Speaking Test rubric, while comments providing an overall evaluation of language ability were coded into a “general” category, and ambiguous comments were coded as “unclear.” Finally, comments addressing scoring or scoring issues were coded into a “scoring” category, while other miscellaneous rater comments were placed in a “miscellaneous” category.

Figure 6.13 shows the frequency (in percentages) with which different types of language features were mentioned in raters' comments; raw values are provided in Appendix Q. There was considerable individual variation in the frequency of comments made concerning various language features, and on the whole there appeared to be few obvious or consistent differences between the three groups of raters. On the other hand, a few trends over time were apparent. In particular, following training (session 2), the proportion of comments focused on topical development increased and became more consistent across raters. For some individuals a dramatic increase in the frequency of such comments was seen; for example rater 109 devoted only 8% of comments to issues of topical development in session 1 but increased to 38% in session 2. Moreover, by
session 2 most raters made more comments regarding topical development than they did for either delivery or language use. While comments regarding delivery remained at approximately the same level across sessions (typically 15%-40%), comments addressing language use decreased in frequency for all but one rater, suggesting that the increase in attention given to topical development may have come at the expense of attention to language use issues. This trend reversed somewhat by the final session where a more equal distribution of comments was seen.

The frequency with which raters made general, unclear, or miscellaneous comments is shown in Figure 6.14. Comments in each of these categories typically did not exceed 10% of all comments, and by session 4 the combined comments in all three categories generally did not surpass 10% of the total for most raters. One exception was rater 108 in the less-proficient group, who made general comments 19%-27% of the time. On the whole, there appeared to be a steady decrease over the three scoring sessions in comments made in these categories, suggesting that raters' comments became more focused on specific language features, or that comments were more clearly related to language features mentioned in the scoring rubric.
Figure 6.13 The frequency of raters' comments on various aspects of language performance, collected during stimulated recall. More proficient raters are shown in white, less-proficient raters are shown in black, and developing raters are shown in gray.
Figure 6.14. The frequency of raters' comments on various aspects of language performance, collected during stimulated recall. More proficient raters are shown in white, less-proficient raters are shown in black, and developing raters are shown in gray.
Figure 6.15. The frequency of raters' comments regarding scoring or the scoring process, collected during stimulated recall. More proficient raters are shown in white, less-proficient raters are shown in black, and developing raters are shown in gray.

Figure 6.16. The average ratio of rater comments per utterance. More proficient raters are shown in white, less-proficient raters are shown in black, and developing raters are shown in gray.
Like comments regarding topical development, comments about scoring were also generally more frequent following rater training, particularly for less-proficient raters (Figure 6.15). In session 1, the frequency of scoring comments ranged from 3%-12% for this group, increasing to 13%-28% following training and suggesting a more explicit focus on the scoring process, perhaps associated either with the training and/or the provision of the scoring rubric. Comments made by rater 109 illustrate both of the possibilities. In comments made during session 2, rater 109 repeatedly referred to the rater training, in which she received feedback that her scores were too low. (In fact, in session 1 she was the most severe of all 20 raters in the study.) Comments included several statements like the following:

```
mm so far I'm thinking this is a four. uh I tend to grade a little bit harsher than what happened in the rater training, so I’m trying to see if it would hit five or not, or, yeah., but so far, I’m thinking it's, perfect four, which, I end up doubting this judgment now because, on the rater training when I thought something was a perfect four, uh it was a five so.
```

In addition, it was also clear in both session 2 and session 4 that Rater 109 was consulting the scoring rubric while scoring. In one example, the rubric is quoted verbatim (with quoted material enclosed in quotation marks):

```
so I'm seeing a number two with “the response demonstrating a limited range and control of grammar and vocabulary”, and it “prevents the full expression of ideas”, and there's difficulty with fluidity and, although sentences are very simple, and some connections, (to) things are a little bit unclear, so that part, is two.
```

In this sequence, Rater 109 provides two direct quotes and a paraphrase of the scoring rubric description of language use expected for a score of two. The other two less-proficient raters did not obviously reference the training session or the scoring rubric, but for whatever reason verbalized more about scoring in session 2.
Finally, the ratio of rater comments to utterances is shown in Figure 6.16, as an indicator of the degree of elaboration in raters' recalls. An utterance was defined as the comments made during a period when playback of the examinee response was paused. Alternately, when the rater spoke over the recording, and utterance was defined as a set of comments made in succession, separated from previous or subsequent comments by a gap of at least 3 seconds. Comments were defined as a single idea unit, and were the basic unit for coding. As can be seen in Figure 6.16, the ratio of comments to utterances increased following training for eight of nine raters, indicating that raters were making more comments and/or elaborating more on their comments in session 2. No consistent differences were seen between rater groups, but, interestingly, ratios for individual raters were virtually identical in sessions 2 and 4, suggesting that the degree of elaboration when making comments may be a matter of individual style. Informal inspection of raters’ comments tends to support the idea that different raters showed different approaches to the verbal report task. For example, rater 101 tended to make extended comments on relatively few issues, rater 102 often made a series of short comments each addressing a different issue, and rater 109 seemed to be checking off items from rubric, as described above.

6.3 Usefulness of a Relative View of Magnitude Judgment

6.3.1 Sequence effect. One prediction of the relative view of judgment is that examples in the environment may be used for comparison when making magnitude judgments. In a testing context, this situation may lead to sequence effect (or anchor and adjustment bias; Plous, 1993), where the magnitude of one score (the anchor) biases
perception of a later score. Such a bias effect should be observable as a correlation between sequential scores (Laming, 2003). Sequence effect was examined using the scores produced for research question 1, where scores produced in trial \( n \) were compared to scores produced immediately previously in trial \( n-1 \). A Pearson product-moment correlation was computed separately across scoring sessions for each rater, with the resulting correlation values used as data in later analyses. The results are presented in Table 6.11.

The average correlation between scores at trial \( n \) versus trial \( n-1 \) ranged from .104 in session 1 to .054 in session 4. One-sample \( t \)-tests (with significance level corrected using a Bonferroni adjustment for multiple tests) found that the correlation values averaged across all raters were significantly different from zero in all sessions \( (p < .05) \); or in other words, a statistically significant sequence effect was seen in all scoring sessions. It should be noted, however, that while significant the magnitude of the correlations was small and would account for 1% or less of the variation in scores. There also appeared to be a general decrease over time in the magnitude of the correlation, both in terms of average value as well as the number of raters with correlations greater than .10 (with the number of such individuals being 9, 7, 7, 5 in sessions 1-4, respectively). This would also be predicted by the relative judgment view in that as judges develop increasing experience with the testing context, they may rely less on examples in the environment when making decisions. On the other hand, a one-way repeated measures ANOVA, conducted on values transformed to Fisher’s \( z \), found no difference between scoring sessions \( (F(3, 57) = 2.253, p = .092, \omega^2 = .04) \). This finding may be partly due to the relatively low power of the analysis associated with relatively few participants.
(observed power = .541), but it must be concluded that the current data do not show any difference in sequence effect over time.

Table 6.11

Average Pearson product-moment correlations for successive pairs of scores

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Mean</th>
<th>SD</th>
<th>Low-High</th>
<th>t(19)</th>
<th>p(two-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R101</td>
<td>.076</td>
<td>.077</td>
<td>.049</td>
<td>.033</td>
<td>.059</td>
<td>.022</td>
<td>.033-.077</td>
<td>5.337</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>R102</td>
<td>.096</td>
<td>.110</td>
<td>-.018</td>
<td>.024</td>
<td>.053</td>
<td>.060</td>
<td>-.018-.110</td>
<td>5.397</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>R103</td>
<td>.047</td>
<td>.102</td>
<td>.144</td>
<td>.040</td>
<td>.083</td>
<td>.049</td>
<td>.040-.144</td>
<td>3.936</td>
<td>.001</td>
</tr>
<tr>
<td>R105</td>
<td>.136</td>
<td>.066</td>
<td>.090</td>
<td>.087</td>
<td>.095</td>
<td>.030</td>
<td>.066-.136</td>
<td>3.962</td>
<td>.001</td>
</tr>
<tr>
<td>R107</td>
<td>.186</td>
<td>.079</td>
<td>.017</td>
<td>.019</td>
<td>.075</td>
<td>.079</td>
<td>.017-.186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R108</td>
<td>-.047</td>
<td>-.028</td>
<td>.122</td>
<td>.113</td>
<td>.040</td>
<td>.090</td>
<td>-.047-.122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R109</td>
<td>.256</td>
<td>.130</td>
<td>.085</td>
<td>.037</td>
<td>.127</td>
<td>.094</td>
<td>.037-.256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R111</td>
<td>.096</td>
<td>.014</td>
<td>.043</td>
<td>.053</td>
<td>.052</td>
<td>.034</td>
<td>.014-.096</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R112</td>
<td>.021</td>
<td>.021</td>
<td>-.018</td>
<td>.081</td>
<td>.026</td>
<td>.041</td>
<td>-.018-.081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R113</td>
<td>.132</td>
<td>.119</td>
<td>.171</td>
<td>.119</td>
<td>.135</td>
<td>.025</td>
<td>.119-.171</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R115</td>
<td>.127</td>
<td>.072</td>
<td>.166</td>
<td>.064</td>
<td>.107</td>
<td>.048</td>
<td>.064-.166</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R116</td>
<td>.104</td>
<td>.145</td>
<td>.167</td>
<td>.086</td>
<td>.126</td>
<td>.037</td>
<td>.086-.167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R117</td>
<td>.024</td>
<td>-.012</td>
<td>-.045</td>
<td>-.122</td>
<td>-.039</td>
<td>.062</td>
<td>-.122-.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R119</td>
<td>.062</td>
<td>.052</td>
<td>.082</td>
<td>.085</td>
<td>.070</td>
<td>.016</td>
<td>.052-.085</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R120</td>
<td>.083</td>
<td>.190</td>
<td>-.060</td>
<td>.071</td>
<td>.071</td>
<td>.102</td>
<td>-.060-.190</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R121</td>
<td>.292</td>
<td>.044</td>
<td>.054</td>
<td>.002</td>
<td>.098</td>
<td>.131</td>
<td>.002-.292</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R122</td>
<td>.109</td>
<td>.063</td>
<td>.049</td>
<td>.115</td>
<td>.084</td>
<td>.033</td>
<td>.049-.115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R123</td>
<td>-.028</td>
<td>.021</td>
<td>-.029</td>
<td>.093</td>
<td>.014</td>
<td>.057</td>
<td>-.029-.093</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R124</td>
<td>.090</td>
<td>-.003</td>
<td>.192</td>
<td>-.050</td>
<td>.057</td>
<td>.107</td>
<td>-.050-.192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R125</td>
<td>.225</td>
<td>.111</td>
<td>.102</td>
<td>.126</td>
<td>.141</td>
<td>.057</td>
<td>.102-.225</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean: .104  .069  .068  .054
SD: .087  .057  .077  .061
Low-High: -.047-.292  -.028-.190  -.060-.192  -.122-.126

Note: Each r value in the matrix is based on approximately 98 comparisons.
Significance tests were calculated using transformed values (Fisher’s z)
6.3.2 Comparison of pairwise and isolated judgments. Another prediction of the relative judgment view is that the precision of judgment should be greater when items can be directly compared as opposed to situations where magnitude must be judged in isolation. This prediction was tested by examining the precision of rater judgments in two conditions: an isolated judgment condition where single responses were awarded a score, much like the scoring done throughout the study, and a pairwise judgment condition in which pairs of examinee responses were presented in immediate succession and the rater was asked to decide if the two responses deserved the same or different scores. The precision of judgment in the two conditions was measured as the proportion of correct discriminations as well as the statistic $d'$ (d-prime) which takes into account any tendency to favor a particular response, such as “same” or “different.”

The pairwise condition did indeed show a higher proportion of correct discriminations as predicted (Figure 6.17). In the pairwise condition the raters made correct discriminations in 62% of all possible comparisons while in the isolated condition correct discriminations were made 51% of the time, a statistically significant difference ($t(18) = 6.90, p < .001, r = .48$). On the other hand, the opposite situation was seen in the $d'$ values (Figure 6.18). Separate $d'$ values were calculated for discrimination of responses with reference scores of three vs. four, $d'(3,4)$, and four vs. five, $d'(4,5)$. In both cases the $d'$ value for the isolated judgment condition was approximately three times greater than in the pairwise condition, a significant difference at the $p < .01$ level (following Bonferroni adjustment for multiple tests).
Figure 6.17. The proportion of correct discriminations in pairwise and isolated judgment conditions.

Figure 6.18. D-prime values for discrimination of responses with reference scores of three vs., four and four vs. five (mean + SE). In both cases, the difference in means is statistically significant (paired t-test; for $d'(3,4)$, $t(18) = -9.483, p < .001$; $d'(4,5)$, $t(18) = -7.863, p < .001$).

The marked difference in the results produced by the two different measures is likely the result of bias in the raters' decision making, as can be seen in the hit rate and false alarm values used to calculate the $d'$ statistic. The hit rate is the proportion of the time two responses with the same reference score were correctly identified as the same.
(pairwise condition) or a score was correctly identified (isolated condition); the false alarm rate is the proportion of the time when a pair of responses with different reference scores was incorrectly identified as the same (pairwise condition), or one score was incorrectly identified (isolated condition).

For discrimination of responses with references scores of three or four, the average hit rate is lower for the pairwise condition (.56 vs..70 ), but perhaps more importantly, the false alarm rate is nearly double that of the isolated condition (.24 vs. .42; Figure 6.19). The difference in false alarm rate suggests that raters had more of a tendency to incorrectly identify the responses as the same in the pairwise condition; that is, they had difficulty distinguishing when the responses were in fact different. The presence of rater bias is also suggested by the results for discrimination of scores of four vs. five, but the situation is somewhat different. For the pairwise condition, the hit and false alarm rates were nearly unchanged, but for the isolated condition both metrics increased with the hit rate increasing from .70 to .88, and the false alarm rate increasing from .24 to .47 (Figure 6.20). This increase in both hit rate and false alarm rate suggest that raters were awarding the score of four relatively often, correctly answering “four” when the response was a four, but also answering “four” when the response was actually a five. Given that a score of four represented the center of the scale available to the raters (only responses with references scores of 3, 4, or 5 were used) it is possible that the $d'$ value for the 4,5 comparison reflects some degree of central tendency in scoring. In any case, given the biases present in the raters' responses, the $d'$ statistic would seem to be the more appropriate measure of discrimination, and it must therefore be concluded that discrimination was greater overall in the isolated judgment condition.
In addition, raters’ feedback collected in questionnaires following data collection for each condition suggested that raters felt that decision making was easier in the isolated judgment condition compared to the pairwise condition (Table 6.12). When asked the question “Generally, how easy was it to decide what score to give?” the average score was 4.5 for the isolated condition compared to 3.3 for the pairwise condition (1 = “very difficult, 7 = “very easy”). Similar trends were seen in raters’ self
appraisal of the accuracy and consistency of their decisions, with decisions made in the isolated condition reported as being more accurate and consistent.

A common comment, both in the questionnaire and in informal conversations with the researcher was that for pairwise judgments raters found it difficult to remember the first response while scoring the second, with several raters also reporting that the pairwise condition was more fatiguing than either the isolated condition or the usual scoring procedure used in the study. Moreover, a few people revealed that in the pairwise condition they simply gave a score to both responses and then compared the scores in order to reduce the mental burden of remembering the details of both responses. This strategy was also apparent in rater responses to the questionnaire. Following the pairwise judgments condition, raters were asked if they scored both responses (Table 6.12, question 5); the average score was 4.9 on a scale of 1 to 7 (1 = never, 7 = always), with two people responding “7” or “always.” Scoring both responses arguably represents a subversion of the original intent of the experiment, given that the pairwise condition was designed to elicit a single same/different decision, not separate scores for each response that were then distilled into a same/different answer. Overall, the raters’ comments suggest the presence of methodological problems in the experiment, which brings into doubt the usefulness of the findings for testing the predictions of the relative judgment view. On the other hand, the cognitive difficulty many raters reported in making pairwise comparisons suggests there may in any case be practical hurdles to the use of pairwise comparisons in speaking tests.
Table 6.12  
*Responses to questionnaires administered immediately after the pairwise and isolated judgment data collection sessions*

<table>
<thead>
<tr>
<th>Isolated Judgment condition</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Low-High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Generally, how easy was it to decide what score to give? 1-7; very difficult - very easy</td>
<td>17</td>
<td>4.5</td>
<td>1.28</td>
<td>1-6</td>
</tr>
<tr>
<td>2) Compared to the “regular” scoring earlier in this study, making a decision using this method was... 1-7; 1 = much easier; 4 = about the same; 7 = much more difficult</td>
<td>17</td>
<td>3.4</td>
<td>1.58</td>
<td>1-6</td>
</tr>
<tr>
<td>3) How accurate are your scores? (Do they match the scale represented by the exemplars)? 1-7; not accurate at all - highly accurate</td>
<td>17</td>
<td>4.6</td>
<td>1.23</td>
<td>1-6</td>
</tr>
<tr>
<td>4) How consistent is your scoring? 1-7; not consistent at all, highly consistent</td>
<td>17</td>
<td>4.6</td>
<td>1.33</td>
<td>1-6</td>
</tr>
<tr>
<td>5) When giving a score, remembering what the previous test taker was like was... 1-7; very easy - very difficult</td>
<td>17</td>
<td>4.2</td>
<td>1.20</td>
<td>2-7</td>
</tr>
<tr>
<td>6) I thought about the previous test taker when deciding on a score. 1-7; not at all - all the time</td>
<td>17</td>
<td>2.4</td>
<td>1.32</td>
<td>1-5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pairwise judgment condition</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Low-High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Generally, how easy was it to decide what score to give?</td>
<td>17</td>
<td>3.3</td>
<td>1.45</td>
<td>1-6</td>
</tr>
<tr>
<td>2) Compared to the “regular” scoring earlier in this study, making a decision using this method was...</td>
<td>17</td>
<td>4.9</td>
<td>1.56</td>
<td>1-7</td>
</tr>
<tr>
<td>3) How accurate are your scores? (Do they match the scale represented by the exemplars)?</td>
<td>17</td>
<td>3.7</td>
<td>1.26</td>
<td>1-5</td>
</tr>
<tr>
<td>4) How consistent is your scoring?</td>
<td>17</td>
<td>3.5</td>
<td>1.23</td>
<td>1-5</td>
</tr>
<tr>
<td>5) I decided on a score for each person, then compared the scores to decide “same” or “different” 1-7; never - always</td>
<td>17</td>
<td>4.7</td>
<td>1.36</td>
<td>2-7</td>
</tr>
<tr>
<td>6) I decided if the test takers were “same” or “different” without considering the score each should get. 1-7; never - always</td>
<td>17</td>
<td>3.1</td>
<td>1.69</td>
<td>1-6</td>
</tr>
</tbody>
</table>
CHAPTER 7

Summary and Discussion

In this chapter, the results of the study are summarized and discussed in the light of previous studies of raters in performance tests, as well as findings from the general literature on expertise. Each research question is discussed in turn, as follows:

1. What effects do training and experience have on scoring patterns and scoring behavior?

2. What features of scoring behavior and cognition distinguish more-proficient and less-proficient raters?

3. Can rater decision making and expertise be understood in terms of a relative view of magnitude judgment?

7.1 Effects of Training and Experience

7.1.1 Summary of results. Scores and indicators of scoring behavior (use of scoring exemplars and the time taken to reach a scoring decision) were collected for 20 raters who were experienced teachers of ESL/EFL, but none of whom had engaged in this kind of language assessment performance rating before. Raters participated in four scoring sessions, one prior to training (and without access to the scoring rubric) and three sessions following training, at approximately one-week intervals. The results are summarized below.

1. In the first session, prior to training or exposure to the scoring rubric, raters’ scoring performance as examined via many-faceted Rasch model (MFRM) analysis was, on average, already of a standard typical for operational language
performance tests (e.g., Bachman, Lynch, & Mason, 1995; Bonk & Ockey, 2003; Eckes, 2005; Lumley & McNamara, 1995; Myford & Wolfe, 2000). The severity of 90% (18 of 20) of the raters was within 1 logit of the mean, suggesting that as a group there was some, but not a great difference in scores between “hard” and “easy” raters. Also, only two raters (of 20) showed significant internal inconsistency problems (their scores did not fit well with the predictions of the Rasch model). The number of rater-by-examinee bias interactions was low, with a total of 44 statistically significant interactions out of a possible number of 2000 (2.2% of the total). Correlations between raters’ scores and established reference scores were also reasonably high, with an average of .734 (range of .64-.86). Conversely, the average of pairwise inter-rater correlations was lower at .596. Inter-rater agreement was also quite modest; exact agreements amounted to 33.9% of possible comparisons, and agreement measured in terms of weighted Fleiss Kappa was .353 (where a value of one represents perfect agreement and zero indicates the level of agreement expected by chance).

2. Training had little apparent effect on rater severity, but there appeared to be some improvement in rater internal consistency. Values for Rasch model fit tended to cluster closer to the expected value following training, with the number of misfitting raters decreasing from two to one and overfitting raters decreasing from three to one. Inter-rater agreement and correlations improved following training, with the Fleiss Kappa values for agreement increasing from .353 to .417 (33.9% exact agreements to 34.6% exact agreements). Average pairwise inter-rater correlations also increased from .596 to .673, suggesting a fairly large effect of
Scoring accuracy improved somewhat, with average correlations to reference scores rising from .734 to .792. Instances of rater-by-examinee bias dropped from 44 to 38, but the frequency of bias was low in any case. Overall, training improved scoring performance in a variety of ways, although rater severity continued to vary across a range of approximately 2 logits.

3. Additional experience following training seemed to have little effect on rater scoring performance overall. Rater severity in sessions 3 and 4 continued to be within 1 logit of the mean. More raters (three individuals) were found to be misfitting in session 3 than in the previous sessions, although no raters were misfitting in the final session. Other features of rater scoring performance, including inter-rater agreement, inter-rater reliability, and the number of bias interactions, showed no additional improvement beyond that observed following session 2 (i.e., immediately following initial training). Scoring accuracy showed mixed results, with no change in correlations with the reference scores, but increasing agreement over time.

4. Scoring behavior differed somewhat among different raters, with inter-rater differences being fairly consistent over time. Most raters checked the scoring exemplars (or benchmark responses) little more than the minimum number of times required to unlock the scoring instrument, but six raters consistently checked the exemplars more often, reviewing them while scoring. Most raters took an average of 60-80 seconds to score each response, consistent with approximately 45 seconds to listen to the response followed by 15-30 seconds to make and record a decision and move to the next examinee. A subset of 6 raters
took longer, on average 80-110 seconds, which may suggest that this group was more likely to listen to the examinee response a second time.

7.1.2 Discussion. The rater training used in the study modified rater scoring patterns in a variety of different ways. Inter-rater reliability and agreement both improved following training: an increase of .077 was seen for average inter-rater correlation and .067 for Fleiss Kappa values, amounting to increases of 13% and 19%, respectively. Scoring accuracy also improved, with the average correlation with reference scores increasing by .058 (8%) and average agreement (Cohen’s kappa) increasing by .05 (11%). These findings suggest that rater variability in application of the rating scale decreased following training, and they are substantially in agreement with results obtained within studies of rater training in writing assessment (Fahim & Bijani, 2011; McIntyre, 1993; Shohamy, Gordon, & Kraemer, 1992). For example, Shohamy, Gordon, and Kraemer (1992) reported increases in intra-class correlation coefficients of .01 to .13 following training (depending on rater group and subscale), similar to the increase in average inter-rater correlation seen in this study. Raters also appeared to be able to better approximate the intended rating scale, as indicated by improved accuracy following scoring. Finally, rater internal scoring consistency as indicated by Rasch model fit also appeared to improve somewhat following training, although rater fit values were generally within acceptable bounds from the start.

On the other hand, training seemed to have little effect on variability in rater severity, with the range of severity values, as measured using MFRM, consistently extending across a range of about 2 logits. This result is also in keeping with the findings
of studies employing MFRM, where it has been demonstrated repeatedly that differences in severity are durable over time (e.g., Lim, 2011; Lumley & McNamara, 1995; Weigle, 1998). Moreover, the finding that training improved internal consistency (model fit) but had little effect on severity is similar to the results reported by Weigle (1998, 1999) as well as the conventional wisdom that training may make raters more consistent in their own scoring, but does not necessarily make scores more consistent between raters (Eckes, 2011; Fulcher, 2003; McNamara, 1996).

While differences in rater severity seems relatively stable, measures of inter-rater reliability and agreement improved following training. Given that variability in rater severity is also an indicator of the extent to which raters agree on scores, the MRFM results would seem to conflict with the patterns seen in the other measures of inter-rater reliability and agreement. The reason for such a discrepancy is not clear, although Eckes (2011) has pointed out that measures of rater reliability and agreement, such as correlations and kappa values, may actually conceal various kinds of rater variability, where Rasch analysis takes this variability into account when estimating rater severity. For example, two lenient raters may agree on relative rankings and absolute scores for a set of examinees, all the while being consistently less strict than other raters (Eckes, 2011, p. 72). Similarly, MFRM provides estimates of severity and consistency for individual raters, while the reliability and agreement indices here were reported in aggregate. While trends were seen in the aggregate data, individual patterns were more variable.

It should also be noted that the impact of differences in rater severity on scores do not always cause problems. Previous studies have observed that while multifaceted Rasch
analysis may detect differences between raters, this variability did not necessarily mean that raters contributed greatly to the overall variance in scores as measured using generalizeability analyses (Bachman, Lynch, & Mason, 1995; Lynch & McNamara, 1998). Moreover, in operational tests rater differences may be accounted for by using multifaceted Rasch analysis to produce the final scores (McNamara, 1996), or differences may be minimized by administrative measures such as monitoring rater performance and excluding those who do not meet minimum standards of agreement with established reference scores (e.g., Xi & Mollaun, 2009).

Training also appeared to have relatively little effect on the prevalence of rater-by-examinee bias. This result is not surprising, given that attempts to influence scoring bias with training have generally met with poor results (Elder, Barkhuizen, & von Randow, 2005; Elder, Barkhuizen, Knoch, & von Randow, 2007; Knoch, 2011; Knoch, Read, & von Randow, 2007; O’Sullivan & Rignal, 2007). In addition, training appeared to have little effect on raters’ use of exemplars or the time taken to reach a scoring decision. This is also not surprising given that the training activities did not provide explicit guidance regarding the use of exemplars or the speed at which raters should work. However, a commonly noted feature of expertise generally is that experts work faster, at least in the performance of basic tasks (Feltovich, Pietula, & Ericsson, 2006). Rater training has also been found to increase the working speed of raters scoring open-ended written responses in a secondary-level biology test (Nádas & Suto, 2010). However, in Nádas and Suto’s study about half of the items required only very short answers (a phrase or less) and so scoring of such answers might be more readily
automatized than scoring of the more complex examinee responses used in the current study.

Following training, little additional change in scoring patterns or behavior was seen over time, suggesting additional scoring experience had little impact. This finding may be the result of the relatively brief period in which data were collected (about two weeks), and additional changes in scoring patterns or rater behavior might have been observed if the study had included more scoring sessions or extended over a longer period of time. However, it should be noted that raters scored 120 responses in each session and by the end of the study had scored over 500 responses (including scoring done in the training session), assumedly enough to establish a high degree of familiarity with the testing context.

Another possibility is that the training and experience raters received in the study was enough to establish competence but not to reach an elite level of expertise, and raters essentially reached a plateau following training. Feltovich, Pietula, and Ericsson (2006) note that studies of expertise suggest that proficient performance may be achieved in tens of hours for many domains, but outstanding performance may require much longer periods, perhaps 10 years or more. They also observe that elite performance usually requires focused experience and practice, and that simple experience is not enough. Following training, raters in the current study were given no such additional focused practice and so a lack of further progress might not be surprising.

In particular, once they had completed training the raters did not receive any additional feedback on the quality of their scores. If the raters had received some form of ongoing feedback, such as viewing the reference score for each response immediately
after each scoring decision (as was done in the training session) then perhaps additional improvement might have been achieved. However, in the current study this approach would have been in conflict with the goal of investigating whether experience, by itself, can improve rater performance. The effectiveness of feedback in modifying scoring patterns of established raters is also unclear, given the limited influence of feedback on rater bias patterns as mentioned earlier and the fact that the raters used in the study were also experienced teachers who presumably had established conceptions regarding the features of strong (or weak) oral language performance. Nonetheless, while simple experience following training seemed to result in little additional improvement, more, or different forms of experience might have produced more positive results and questions remain regarding what human raters might be capable of in terms of scoring performance with enough of the right kind of training.

It should also be noted that, prior to training, raters had already achieved a level of scoring performance typical of raters in operational speaking tests, at least in terms of the features measured by MFRM. Moreover, average correlation with reference scores was .734, an impressive result for raters who had not yet seen the scoring criteria used to produce the reference scores. Rater scoring performance in session 1 becomes even more impressive when it is remembered that the original TOEFL iBT Speaking Test rating scale was expanded from the original four point scale (1-4) to a six point scale (1-6) with the express purpose of making it more difficult to achieve reliable and accurate scoring. Although good scoring performance among untrained raters is not unheard of (e.g., Lim, 2011; Shohamy, Gordon, & Kraemer, 1992), the question remains as to how novice raters were able to achieve this level of performance. One possibility is that the previous
teaching and other experiences of the raters provided them with enough knowledge about the examinee population to be able to score examinee responses in a fairly consistent way. This explanation fits well with the approach of operationalizing rater expertise in terms of teaching experience, as has been done in some studies of language tests (e.g., Cumming, 1990; Delaruelle, 1997). A second possibility is that the TOEFL iBT Speaking Test scoring criteria intuitively “make sense” or are otherwise aligned with ES/FL teachers’ native perceptions of language proficiency. Such a situation would not be surprising given that comments from teachers of English for academic purposes were used to construct and confirm the TOEFL scoring rubric (Brown, Iwashita, & McNamara, 2005; Enright, Bridgeman, Eignor, Lee, & Powers, 2008).

A positive effect of rater background experience on scoring performance might also be consistent with the observation that expertise is domain specific (Feltovich et al., 2006; Glaser & Chi, 1988; Shanteau, 1992). That is, raters’ experience in evaluating learners in the course of their teaching may extend to the scoring of examinee responses in a language performance test, although the degree to which a rater’s particular background experiences are relevant will undoubtedly vary. Relative judgment theories, such as Decision by Sampling (Stewart, Brown, & Chater, 2006) suggest a complementary view, which is that as teachers gain more experience with a particular population of students, they develop an increasingly robust internal distribution of language proficiency samples. To the extent that this population of learners is comparable to the learners taking the test, such experience should contribute to more reliable scoring.

Use of the scoring exemplars might have provided another mechanism by which raters were able to achieve good scoring performance prior to training. Exemplar
responses with associated scores were provided in the first scoring session (as well as a preceding orientation session) and may have provided enough of a reference for raters to score fairly accurately. If this was the case, then it is possible that provision of benchmark responses in itself may be largely adequate to help raters (or raters with appropriate background experience) produce acceptably consistent scores.

This possibility then raises the issue of what the scoring rubric contributes to the process. The conventional wisdom is that the scoring rubric helps to ensure the reliability and validity of scores by providing guidance regarding the language features to be considered, as well as how to evaluate these features (Alderson, 1991; Fulcher, 2003). In contrast, Lumley (2005) noted that rating scales represent simplified and idealized descriptions of actual performance, leaving considerable latitude to the rater. He suggested that rather than basing decisions on the rubric, raters made decisions based on intuitive impressions, with the scoring rubric used to justify decisions rather than make them. If Lumley is correct, then one might speculate that the scoring rubric was less important in aligning rater perceptions to the rating scale; rather, the exemplars may have played a larger role in promoting accuracy by influencing raters’ intuitive perceptions.

On the other hand, scoring performance did improve following training, and the rubric may have played a role in this improvement. In particular, the rubric appeared to focus raters’ attention on issues of topical development (see section 7.2 below) and provided guidance regarding what should be considered good “content” for an academic speaking task. This guidance could have reduced differences in raters’ perceptions regarding what content or organizational style was appropriate for examinee responses, thus increasing scoring accuracy and reliability.
Finally, it is worth considering what these results say about the nature of rater expertise. If one takes the view that training increases expertise, then those aspects of rater performance that improve with training might be considered characteristics of “expert” raters. The current results then suggest that internal scoring consistency, inter-rater reliability and agreement, and scoring accuracy are features that should distinguish experts from novices. In other words, as raters develop expertise, they become generally more consistent in scoring and better aligned with the rating scale. Features of rater performance that did not change with training, including rater severity and the prevalence of bias might represent an intractable noise that is simply part of rater variability, or perhaps, are issues that are only resolved at truly elite levels of performance.

7.2 Differences Between More- and Less-Proficient Raters

7.2.1 Summary of results. From the group of 20 raters studied in research question 1, more-proficient, less-proficient, and developing raters (3 each) were identified on the basis of their severity, internal consistency (Rasch model fit), and accuracy of scores. Their use of the rating scale, scoring behavior, and cognition was then compared. The results are summarized as follows.

1. In addition to greater scoring consistency measured in terms of Rasch model fit, more-proficient raters also showed higher levels of within-rater agreement across scoring sessions, compared to less-proficient and developing raters. The percentage of exact agreements for scores awarded to a set of 16 responses repeated in all scoring sessions ranged from 50%-65% for more-proficient raters and 30%-40% for less-proficient and developing raters. In terms of use of the
rating scale examined via MFRM, increasing raw score values from more-proficient raters corresponded to relatively even increases of examinee ability, better approximating an interval measurement scale. In absolute terms, boundaries between adjoining raw scores were more likely to correspond to the scale estimated using data from all raters (i.e., the best estimate of the scale as operationalized in the assessment). In other words, scores of more proficient raters showed better alignment to the rating scale operationalized in the assessment as a whole.

2. More-proficient raters tended to use the exemplars more and reviewed exemplars while scoring; less-proficient and developing raters essentially only viewed the exemplars the minimum number of times required to begin each scoring session.

3. More-proficient raters generally used more time to make decisions, on average 83-100 seconds, while less-proficient and developing raters took an average of 56-70 seconds. (But, one individual in the less-proficient group spent an average of 95 seconds.) Closer examination of the data suggested that experts were more likely to listen a second time to the examinee response or to an exemplar; less-proficient raters appeared to seldom listen a second time, if ever, and developing raters showed an intermediate pattern. Experts also seemed somewhat more likely to take breaks while scoring, although considerable individual variation was seen in this behavior.

4. Considerable individual variation was seen in the frequency with which different language features were mentioned by raters while scoring. Following training, mention of language features related to the “topical development” category of the
scoring rubric increased for all groups to become the single most-often mentioned feature, with the magnitude of change being greatest for developing raters and least for more-proficient raters. By the final scoring session, however, mention of topical development was lower and more similar to the first session. General, miscellaneous, or unclear comments were generally low in frequency in all sessions, and appeared to decrease somewhat over time. More-proficient raters consistently made very few such comments, with greater individual variation seen in the other groups.

5. The frequency of rater comments on the scoring process generally increased following training, suggesting raters were giving more explicit attention to scoring. In addition, less-proficient and developing raters tended to make relatively more comments about the scoring process; in at least one case this was associated with feedback received during rater training the rater’s scores consistently differed from the reference scores.

6. Each rater appeared to have her or his own style of commenting, as reflected in both the sequence of topics covered in comments as well as the amount of elaboration on particular points.

7.2.2 Discussion. Rater behavior appeared to differ between raters with better scoring performance compared to those whose performance was less good in terms of scoring severity, internal consistency, and scoring accuracy. One major difference was that more-proficient raters used the exemplars more, reviewing the exemplars while scoring. This behavior might be viewed as an ongoing recalibration, and as such it seems
intuitive that it would lead to improved scoring accuracy. From a relative judgment view, the exemplars provided reference points that could be compared to examinee responses, increasing the accuracy and reliability of scores (Laming, 2003; Laming, 2004). In fact, Laming (2003), discussing the British university examination system, argued that provision of extraneous standards is one of only two possible ways to improve rater accuracy (the other being to independently judge the component parts of the response). While the more-proficient raters examined in this study did not check the exemplars for every decision (the highest rates of exemplar use corresponded to one exemplar check for every 4-5 examinee responses), the fact that exemplar use was associated with more-proficient scoring performance is nonetheless consistent with the view that magnitude judgments are ultimately relative in nature.

Interestingly, use of the scoring rubric did not necessarily result in better scoring performance. One of the less-proficient raters (rater 109) appeared to make considerable use of the scoring rubric following training as indicated by verbal report, but this did not result in improved scoring accuracy. Although no conclusions can be drawn from a single case, this observation does raise the issue of the degree to which use of the scoring rubric actually influences scoring patterns, as was discussed in the previous section.

More proficient raters also took more time in making scoring decisions. This observation would seem to be in contrast with the finding that experts tend to perform basic tasks in a faster and more automatic way, but, as discussed earlier, it is possible that the evaluation of language samples represents a more complex task that is not easily automatized. In addition, more-proficient raters appeared to be more likely to listen to the examinee response a second time, perhaps suggesting that this group took greater care in
their decision making. One possibility is that this extra care was realized as additional “data collection” by the rater before making a decision; this interpretation would be consistent with the finding that expert raters tend to withhold judgment longer before making a decision (Huot, 1993; Wolfe, 2006). Alternately, the extra time might have been spent confirming a decision made earlier, an interpretation that fits well with models of the writing scoring process (i.e., Lumley, 2005; Milanovic, Saville, & Shen, 1996; Wolfe, 1997).

In contrast to the differences in behavior seen between more- and less-proficient raters, relatively little difference was seen in the attention given to different scoring criteria, as determined using stimulated recall. Rather, there was considerable individual variation in the frequency that different language features were mentioned, and this, along with the various ways that different raters organized and elaborated on their comments, suggests that different raters had different approaches to scoring that were not necessarily tied to the quality of their scores.

One pattern apparent in the data was that mention of language features related to topical development increased after training. This was true for raters in all groups, with more dramatic differences seen for less-proficient and developing raters. This increase may be related to the scoring instructions provided during the rater training session. The instructions covered the scoring of the two prompts raters actually used in the study and focused exclusively on matters of topical development; given that organization and content may differ depending on the nature of the prompt, prompt-specific guidelines help to ensure consistent interpretation of scoring criteria. No such instructions were provided for the other two categories of the scoring rubric (delivery and language use).
since these features are less likely to systematically differ between prompts, at least within the range of prompts used as independent tasks in the TOEFL iBT Speaking Test. It seems possible that the prompt-specific scoring instructions focused raters’ attention on issues of topical development, resulting in more comments made on this feature following training. This focus seemed to diminish by the final scoring session, however, where roughly equal mention was made of the three language categories included in the scoring rubric.

In addition, more-proficient raters appeared to consistently forego comments unrelated to specific rubric categories, and there also appeared to be a tendency towards fewer such comments as raters gained more experience. The frequency of these kinds of comments was generally low, however, which may be partly due to the coding scheme used in the study. Comments that could be logically placed in a rubric category were coded as such, even if the specific feature mentioned was not referenced in the rubric. For example, comments regarding “accent” were coded as delivery (which covers pronunciation and fluency), even though accent is not specifically mentioned in the scoring criteria. If a more strict interpretation had been used, the number of “unclear” comments would have undoubtedly been greater. Nonetheless, these findings might suggest that more-proficient raters, or raters with more experience, were focused more tightly on language features relevant to the scoring task.

Finally, less-proficient and developing raters tended to make more frequent mention of the scoring process itself, especially after training (although again, there was considerable individual variation). Given that reflective practice has been reported as one characteristic of experts (Feltovich et al., 2006; Tsui, 2003) this finding seems
counterintuitive. On the other hand, it is possible that the increased mention of scoring reflects difficulty in reaching a decision or a lack of confidence in scoring. One less-proficient rater (rater 109) repeatedly commented that her scores had been “off” during training, and in response seemed to be giving extra care to the scoring process in the following session.

7.3 Relevance of the Relative Judgment View to Judgments of Language Ability

7.3.1 Summary of results. Two predictions of the relative judgment view were examined: (a) that scores awarded to previous responses serve as a point of reference when judging subsequent responses (i.e., sequence effect) and (b) discrimination between responses at different levels of performance will be improved when responses can be directly compared. Scores collected as part of research question 1 were analyzed for sequence effect. In addition, following the conclusion of data collection for research question 1, all raters completed two additional scoring sessions to examine the effect of scoring method on raters’ ability to discriminate between examinees. In one session responses were judged one by one with a break in between each, while in the other session the same responses were presented in pairs (one immediately after the other) and raters judged whether the two responses merited the same or different scores. The results are summarized as follows.

1. A small but statistically significant sequence effect was found; that is, a correlation was observed between scores awarded in immediate succession. Average $r$-value, computed from individual $r$-values for each rater, was significantly different from zero for all scoring sessions. The average $r$-value also
appeared to decrease over time, starting at .104 in session 1, dropping to .069 after training, and then decreasing to .054 by the end of the study. Although these differences did not reach statistical significance, they are suggestive that the magnitude of the sequence effect decreased as raters completed training and gained scoring experience.

2. Raters’ ability to successfully discriminate between examinees, measured in terms of $d'$, was on the order of three times higher when examinees were judged separately than when judged in pairs. This result seemed to be influenced by both a difficulty in distinguishing responses in the pairwise condition (i.e., raters answering “same” when the two responses actually differed in level) as well as a tendency to favor the center of the scale in the isolated judgment condition. Rater comments suggested that many felt that the pairwise judgment condition was cognitively more difficult, and some reported awarding separate scores to responses in the pairwise condition, contrary to the intent of the task.

### 7.3.2 Discussion

A small sequence effect was found in scores, suggesting that one score (or examinee) was used as a reference when scoring the next examinee. This finding is consistent with studies of English L1 essay marking (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes, Keeling, & Tuck, 1980) as well as studies of magnitude judgments of physical phenomena (e.g., Jesteadt, Luce, & Green, 1977; Ward, 1979). The findings also support the theoretical position that judgments of magnitude are relative in nature, where it is predicted that recently experienced items may serve as points of reference for subsequent judgments (Laming, 2003, 2004; Stewart,
Brown, & Chater, 2005, Stewart, Chater, & Brown, 2006). Moreover, the magnitude of the sequence effect decreased over time, something also predicted by the relative judgment view (Matthews & Stewart, 2009; Stewart, Chater, & Brown, 2006). This view theorizes that as raters encounter more examinee responses, they develop a more robust internal distribution of examples that can be used for comparison when making judgments; therefore, they have less need to rely on comparisons with other examples (responses) in the environment. This in turn should reduce the magnitude of a sequence effect. In any case, the sequence effect observed in the current study was quite small in magnitude, with average $r$-values of .1 or less, which would correspond to 1% or less shared variance between scores. This finding does not preclude the possibility of substantial sequence effect in specific cases, however, such as when a particularly poor response is followed by an average one (where the average response might be judged more generously, Hales & Tokar, 1975). However, on aggregate, the level of sequence effect observed here would likely introduce little additional variability into scores.

In contrast to the predictions of the relative judgment view, scoring examinee responses in pairs did not result in improved discrimination between responses of differing levels of performance. In fact, the results were exactly the opposite, with much better discrimination seen in the isolated judgment condition. In the pairwise judgment condition, raters were both more likely to mistakenly respond “different” when the two responses had the same reference score, as well as more likely to respond “same” when the two responses were actually different. Many raters also reported that they felt it very difficult to jointly consider two responses at the same time, holding the features of the first response in mind while listening to the second. As a result, a number of individuals
simply scored the first response, then scored the second response, and finally compared the scores to arrive at a “same” or “different” answer, defeating the intent of the study and casting doubt on the results.

This finding highlights the fact that complex multidimensional stimuli were used in the current study (extended oral responses) which are quite different from the simple unidimensional kinds of physical stimuli typically used in psychophysical studies. It cannot be assumed that the psychological processes involved in making magnitude judgments are the same for both simple and complex stimuli. On the other hand, the holistic rating scale used in the study is based on the assumption that responses can be usefully represented by a single score, even if the capabilities underlying the performance are multidimensional. However, in the pairwise comparison experiment it appears that the raters could not (or simply did not) make a holistic or gestalt type of comparison between test taker responses, and rather attempted to compare specific aspects of each response. That raters should have difficulty with such a detail-by-detail comparison is perhaps not surprising; the observation that humans are limited in their ability to simultaneously evaluate specific language features is one argument for the use of automated scoring, which employs machines that have no such limitation (Bridgman, 2012). Rather, human raters seem to be better at producing holistic judgments that better capture the totality of the construct being measured (Bridgeman, Powers, Stone, & Mollaun, 2012), and it seems possible that if raters had been more explicitly instructed to make gestalt judgments of examinee performances, then perhaps the comparison task might have been less challenging and the results different.
Nonetheless the experiment did provide some useful practical insights into the use of a pairwise approach to scoring speaking tests; comparing two speech samples, even relatively short ones, was cognitively very challenging and probably not practical for operational scoring. Conversely, not every rater felt that the pairwise condition was more difficult. In the post-scoring questionnaire three raters indicated that they felt the pairwise method was actually easier than the typical scoring procedure used throughout the study. Interestingly, the two raters who checked the exemplars the most (raters 101 and 102) both answered that they felt the pairwise method was more difficult than “regular” scoring, despite the fact that during such regular scoring they frequently seemed to compare exemplar responses to examinee responses. While anecdotal, this observation might suggest that when it comes to evaluating speech samples, not all comparisons are the same, and that the task that raters were asked to do in the experiment happened to be particularly challenging. As mentioned earlier, a gestalt judgment might have reduced the difficulty of the task. Moreover, a more focused comparison, where the rater already has something in mind to check when listening to the second speech sample (as might be the case when reviewing an exemplar), could also be cognitively easier than making an unguided comparison of two responses. In the latter situation, the features that distinguish the responses will not be clear at first, so raters may try to remember as much as possible, greatly increasing the cognitive difficulty of the task.

Finally, the expertise literature suggests that experts are better able to organize and chunk information, using their knowledge of patterns within the domain to organize new information in a more principled way, which reduces the load on memory (Glaser & Chi, 1988; Feltovich et al., 2006). So, it is also possible that as raters become more
experienced with the typical patterns seen in examinee responses, they might be able to more efficiently manage the cognitive demands of comparing different examinees. In the present study, perhaps the raters did not get enough experience to achieve such improvement in information processing.

To close, the experimental comparison of pairwise and isolated judgments conducted in this study likely suffered from methodological problems, and may not have targeted the kind of comparisons that raters actually make while scoring. So, the data probably make little contribution to a theoretical understanding of rater decision making. On the other hand, the experiment arguably makes a useful practical observation; specifically, that unguided comparisons of oral responses are difficult to do. Such comparisons are likely not practical for operational scoring, although might be useful for specific situations where fewer comparisons can be made and more time is available for repeated listening, such as the construction of rating scales (e.g., Turner & Upshur, 2002; Upshur & Turner, 1999) or for equating different administrations of a test (Black & Bramley, 2008; Pollitt, 2004; Pollitt & Crisp, 2004).
CHAPTER 8

Conclusion

This study investigated the nature of rater expertise within a speaking test context and examined: (a) how training and experience influenced rater scoring patterns and behavior, (b) the ways in which raters showing better or worse scoring performance differed in behavior and cognition, and (c) whether a recent theory of magnitude judgment from the domain of psychophysics is relevant to scoring judgments in a language performance test. In the previous chapter, specific findings of the study were summarized and discussed. This final chapter discusses the broader implications of study findings in terms of both our understanding of rater expertise, as well as practical issues related to the use of raters in speaking performance tests. The chapter begins with a description of the limitations of the study, followed by a discussion of study implications, and finally offers suggestions for future work in this area.

8.1 Limitations of the Study

This study has a number of limitations that should be considered in the interpretation of study findings. First, with the exception of a face-to-face orientation prior to data collection, raters worked largely unsupervised from their own locations, accessing data collection materials online. This measure was necessary given that the local area did not contain enough participants with both appropriate background qualifications and adequate time to devote to the study. Additionally, the long scoring sessions and large time commitment required of study participants (approximately 32 hours over a period of 4-6 weeks) made it impractical to ask raters to come into the
laboratory for data collection. While some degree of supervision was provided via email and phone, the end result was that there was very little real control over the exact circumstances in which raters carried out the various study tasks. For example, scoring sessions were completed at different times of the day (and night, in some cases), and in some instances the first half of the scoring session was done earlier in the day, with the second half done later in the day. Such situations were very difficult to avoid given the busy schedules of the teachers and graduate students who participated in the study, as well as the fact that each half of the scoring session (each covering a different prompt) took 2-3 hours to complete and had to be scheduled in whatever free time was available. Moreover, while raters were asked to do the scoring in a location where they could work undisturbed, choice of location was ultimately at the rater’s discretion. Recordings of stimulated recalls included occasional disturbances and background noise, suggesting that at times raters may have been distracted from the task at hand. In sum, the circumstances of data collection fell short of a controlled laboratory study.

On the other hand, it is worth noting that operational scoring of the TOEFL iBT Speaking Test uses a similar approach, with a dispersed network of raters accessing materials online and scoring from their individual locations. So, although far from perfect from a data collection standpoint, the unsupervised scoring used in the current study at least approximates a real-world approach to scoring a large-scale speaking performance test. In addition, time stamp data collected from the scoring instruments suggested that raters were conscientious in completing the work. The average number of breaks (of five minutes or more) taken ranged from 2.4 to 4.0 per session, modest numbers given that each session involved scoring 100 responses and producing stimulated recalls for another
20 responses. While unsupervised, evidence suggests that raters were diligent in their work.

Fatigue may also have influenced raters’ performance given that a full scoring session required about 5 hours to complete. Informal comments by several raters indicated that they felt the sessions were tiring. However, once again it should be noted that long scoring sessions are representative of the scoring done in the operational TOEFL iBT Speaking Test, where raters work shifts of 4, 6, or 8 hours in length (Pam Mollaun, personal communication, November 5, 2010). In addition, the order in which responses were presented to each rater was individually randomized, so fatigue effects, if any, should be distributed throughout the scoring data. Fatigue effects might have been more substantial for the stimulated recalls, however, since they were conducted at the end of each half of the scoring session.

In addition, the quality of the stimulated recall data was perhaps most vulnerable to the lack of close supervision during data collection. While raters were trained in the stimulated recall technique during the face-to-face orientation, lack of supervision while carrying out the recalls may have influenced the amount and quality of the data produced. Conventional wisdom regarding verbal reports is that it is useful for the researcher to be present to ensure the right type of comments are produced and to prompt the participant if he or she falls silent (Ericsson & Simon, 1993; Gass & Mackey, 2000; Green, 1998). The unsupervised procedure used in the current study represents a departure from the usual practice in collecting verbal report data, although lack of supervision is not unprecedented (Ericsson & Simon, 1993). Nonetheless, the stimulated recall data collected in the study may be less well focused than would have been the case otherwise.
Analysis of the results so far, however, suggests that raters generally followed the instructions, producing adequate amounts of usable data.

Another limitation of the verbal report data is simply that most of it remains unanalyzed. While rough transcriptions have been produced for all recalls, time constraints have made it impossible to analyze the entire data set of 1,598 transcripts and approximately 200,000 words. In this dissertation 270 recalls were analyzed (16.9% of the total) in the comparison of more-proficient, less-proficient, and developing raters. Clearly, the current effort represents only a first step in the analysis of these data.

Finally, it should be noted that the time course of data collection was relatively short for a longitudinal study, which may have obscured changes in rater behavior resulting from experience and reduced the possibility of developing true expertise. Other studies of rater expertise have collected scoring data on as many as 21 occasions over a period as long as 20 months (Lim, 2011). In contrast, within the current study the effects of experience were examined over a period of about 2 weeks following training. Raters’ scoring patterns and behavior changed little during this period, beyond changes observed in response to initial training, and it is possible that the results might have been different had data collection extended over a longer period of time. Nonetheless, raters scored over 500 responses by the end of scoring session 4 (including responses scored during training), representing a non-trivial level of scoring experience.

8.2 Theoretical Implications of the Study

The study findings have a number of implications for theory. First, the study provided insight into the characteristics rater expertise. Specifically, it was observed that
internal scoring consistency, scoring accuracy, and inter-rater reliability and agreement increased with training. Additionally, more-proficient raters showed greater scoring consistency over time and produced scores that better approximated the rating scale operationalized in the assessment. In terms of scoring patterns, rater expertise might therefore be conceptualized as increased consistency in scoring as well as more accurate implementation of the rating scale. These results largely confirm findings regarding the nature of expertise from the domain of writing assessment (e.g., Weigle, 1998, 1999; Wolfe, 2006), extending these findings into a speaking test context.

It should be made clear, however, that this view of rater expertise privileges consistency in judgments and applies to testing contexts where there is an emphasis on reliability of scores. In other assessment contexts diversity in rater perception may be a virtue, providing a broader view of the construct(s) of interest (McNamara, 1996; Moss, 1994). One example would be a dissertation committee, where a diversity of viewpoints among committee members allows for a more thorough evaluation of the candidate’s work, and variability in judgments can be seen as a strength rather than a shortcoming. Where the focus is on providing feedback or otherwise producing contextualized results, the features that make an effective rater might be quite different than those presented here.

In addition, behavioral differences were found between raters with more-proficient scoring patterns and those showing poorer scoring performance. More-proficient raters used the exemplars more, took longer to make decisions, and may have been more likely to re-listen to examinee responses. The overall impression is that more-proficient raters may have been more careful in their decision making. This finding
stands in contrast to the general expertise literature which suggests that experts tend to work faster. A possible implication of this divergence is that the task of scoring oral language may be qualitatively different from the kinds of routine tasks that experts learn to perform quickly. Rather, it may be that the scoring of communicative oral language performance involves a degree of complexity or unpredictability, such that judgment cannot easily be automatized, but rather benefits from an extra degree of care.

While training improved scoring performance in a variety of ways, untrained raters were nonetheless able to achieve acceptable levels of scoring consistency and accuracy, raising the issue of the relative importance of background experience and test-specific training. It seems reasonable that previous experience with English language learners should be helpful in judging oral performance within a test of English as a foreign language, and the relative judgment view predicts that experience with similar learners should improve raters’ ability to discriminate between examinees. Theories of relative judgment also predict that having exemplars available while scoring (as was done in the first scoring session) should improve scoring accuracy (Laming, 2003), and the testing context itself may have contributed to rater performance. Raters scored responses that were quite short and addressed a single type of speaking task, perhaps making it easier for raters to quickly get a sense of the range of examinee performance and achieve consistency in distinguishing different degrees of ability. Nonetheless, the fact naive raters produced useful scores seems rather counterintuitive given that it is generally accepted that training is necessary for reliable scoring in language performance tests (Fulcher, 2003). The current findings suggest that while typical rater training activities
are helpful, they may not be an absolute requirement for reliable and accurate scoring, at least given the right circumstances.

While rater expertise may involve particular scoring patterns and include a behavioral element, the importance of raters’ internalized scoring criteria for defining rater expertise is less clear. More- and less-proficient raters differed little in the frequency with which they mentioned different scoring criteria, and a diligent effort by a less-proficient rater to rigorously apply the scoring criteria published in the rubric did not lead to greatly improved scoring performance. This finding also raises the question of the degree to which explicit descriptions of performance, such as those found in scoring rubrics, actually influence scores awarded. As mentioned in the previous chapter (section 7.1.2), raters are faced with the task of applying relatively vague scoring criteria to complex phenomena, and written scoring criteria might be used more to justify a decision made intuitively than to actually make the decision itself (Lumley, 2005). Rather than being a definitive statement of the language features measured by a particular test, it seems possible that the scoring rubric simply provides one piece of evidence, and the actual construct operationalized in a particular test resides just as much in specific exemplars and training procedures used to develop a shared understanding among raters.

Also, the study provides limited initial support for a new way of conceptualizing the fundamental psychological nature of the scoring process; namely, that magnitude judgments of language ability are relative in nature. Specifically, it was shown that scoring judgments are influenced by similar examples of performance available in the immediate environment, as indicated by sequence effect in scores. Additional support for the relative judgment view comes from the finding that use of exemplars was associated
with scoring accuracy/consistency in more-proficient raters, although this finding is highly tentative given that the association is confounded by other variables such as the time taken in scoring. In any case, this study was the first attempt to evaluate the predictions of the relative judgment view within a language assessment context, and represents one of only a handful of attempts to apply the predictions of relative judgment theories to domains outside of psychophysics. While the results of the current study were somewhat inconclusive, they arguably suggest that further examination of the predictions of the theory of Decision by Sampling (Stewart, Chater, & Brown, 2006) as well as related claims made regarding the judgment process (e.g., Laming, 2003, 2004) are worth pursuing. Although the empirical findings produced so far are quite modest, the current study introduces to the language testing community a new, and testable, framework for thinking about the way the decisions are made, and for guiding studies of rater cognition.

8.3 Practical implications of the Study

In addition to the theoretical implications, the study also has several practical implications for the training and management of raters in speaking tests. First, the results highlight the importance of examples for aligning rater perceptions to the rating scale. With only the exemplars for guidance, raters in the study were able to achieve reasonably good score reliability and accuracy in the first scoring session. So, it might be useful to make benchmark samples available to raters and encourage their use, especially for difficult cases. Moreover, the finding that raters were able to achieve reasonable scoring accuracy and consistency with only the exemplars for guidance would suggest that review of benchmark samples or practice scoring with feedback prior to starting work
may improve scoring performance. Some examinations already require raters to score a sample set with a certain degree of accuracy before being allowed to start a scoring session (e.g., TOEFL iBT Speaking Test; Xi & Mollaun, 2009), and it would be a simple matter to provide the rater with a reference score for each practice response. If a reference score was provided immediately after each scoring decision, each response would then essentially serve as an additional benchmark.

Training still has value, however, as indicated by the improved scoring performance seen after raters completed a series of typical training activities (exposure to the rubric, review of benchmarks with associated scoring rationales, and practice scoring with feedback). On the other hand, subsequent experience did result in further improvement in scoring performance, suggesting that unguided experience may not be effective in developing scoring expertise, at least once a rater has reached a certain level of performance. This view is consistent with the broader literature in expertise, which suggests that elite levels of performance can be obtained only through focused practice. Whether or not such focused practice (such as practice scoring with feedback) would further improve scoring performance in a language test remains an open question.

Also, the difficulties that study participants had in simultaneously evaluating pairs of oral responses suggest that such pairwise comparisons are challenging and may be impractical for scoring operational speaking tests. Such pairwise comparisons may be more tractable in writing assessment, where texts can be physically placed side-by-side, or perhaps in the development or equating of speaking tests, where more time is available for reviewing oral responses. However, the results suggest that making judgments of
paired oral responses is quite challenging, at least using the method employed in the current study.

Finally, this study demonstrated the use of interactive online forms for the purposes of data collection in a language assessment context. In addition to simply collecting scores or type-in data, the Adobe pdf forms used in this study automatically collected a considerable amount of information regarding the behavior of raters while scoring. Normally, such behavioral data would be available only by using specialized software, probably located on a computer in a laboratory setting. Moreover, the instruments were produced using widely available software and with a very modest amount of programming expertise. The current study demonstrates that tools are now available to the average researcher that allow collection of new types of data, with the potential for new insights into raters’ decision making processes.

8.4. Possibilities for Future Research

The findings and implications of the study suggest a number of possible avenues for future research. First, there is the issue of the large amount of unanalyzed verbal report data from the current study. Further analysis of these data might be used to confirm the current findings, as well as extend the findings into the processes by which raters make decisions. Moreover, verbal report data might be more closely integrated with the scoring and behavioral data that were collected at the same time, to explore connections between rater cognition, scoring behavior, and scores. For example, in cases where raters show disagreement in the score awarded to a response, analysis of rater recalls might illuminate the sources of disagreement, perhaps giving insight into what makes particular
responses problematic to score. In another example, comparison of examinee responses and rater comments might clarify the features of performance that contribute to variability in rater perceptions. Informal impressions from the current data suggest the possibility that when a performance has a noteworthy characteristic such as a very long pause, raters tend to comment on it. When the response is less noteworthy, then comments tend to be more variable across raters, and raters may fall back on established individual patterns in making comments.

Second, the finding that raters showed reasonably good scoring performance in the absence of a scoring rubric raises the question of the influence of various types of aids on scores and the scoring process. While a limited number of studies have examined the influence of scoring rubrics (e.g., Lumley, 2005; Pollitt & Murray, 1996) and feedback (e.g., Elder, Barkhuizen, Knoch, & von Randow, 2007; Knoch, 2011; Wigglesworth, 1993), little attempt has been made to systematically document how different aids such as benchmarks, rubrics, scoring rationales, and practice scoring work together to influence rater decision making. One question of interest might be to examine the relative contribution of scoring aids that present the rating scale implicitly versus those that present the scale explicitly, with exemplars and practice scoring (with feedback) falling in the former category, and scoring rubrics and rationales included in the latter category. The current study suggests both sorts of aids may be useful in improving scoring performance, but implicit aids might be more effective for establishing reliable scoring, while explicit aids guide raters to attend to relevant features of performance.

Third, questions remain whether the importance of rater training or rubrics may differ across different test tasks. In addition to the opinion tasks used in the current study,
the TOEFL iBT Speaking Test includes a series integrated tasks where examinees must incorporate information from a written and/or spoken text into their responses. Performance of these tasks is partly judged on the basis of the quality and depth of the incorporated information (ETS, 2007), and it seems unlikely that raters would be able to make consistent or appropriate judgments in this area without some familiarity with the specific information that should be included in the response. In cases where specific content matters, training may very well take on a greater role in promoting the reliability and accuracy of judgments, and reliable scoring may depend more on explicit scoring instructions regarding how to evaluate the information provided in the response.

Fourth, the current study also suggests that the relative view of magnitude judgment may provide useful insights into the basic psychological nature of scoring judgments, and there are a variety of predictions that might be tested in future work. To start with, a new experiment might be designed to test the prediction that discrimination between examinees is improved when responses can be judged side-by-side. This type of study may be difficult to do in a speaking test context, but might be more easily done in a test of writing, where there is the possibility of physically placing written scripts side-by-side. Other predictions also wait to be tested. For example, in terms of rater training, the relative judgment view would predict that it will be important to expose new raters to as many examples of scored examinee performance as possible, to establish a profuse inner distribution of examples to use when scoring. This approach differs from one commonly used in language tests, where a few examples are scored and then scoring is discussed at length. So, it might be interesting to compare the relative efficacy of different training regimens: one that focuses on providing as much exposure as possible to examples of
examinee performance, and another that takes a more traditional approach where emphasis is placed on explicit discussion of how to score a limited number of examples.

Another prediction that could be tested is that the distribution of examples experienced in the past will influence magnitude judgments. For example, a rater whose experience is solely with low-proficient language learners would be expected to be better able to distinguish test takers at the lower end of the spectrum, since the rater has more examples of performance at this level to use for comparison. (An analogous situation might be an expert on French wines who is able to distinguish minute differences between French vintages, while being less sensitive to subtle differences in wines from another region.) This prediction might be tested by exposing groups of naive native speakers, who have little experience with language learners, to samples of L2 speakers that have been skewed towards one end or the other of a speaking proficiency continuum and then investigating whether the two groups show differences in their ability to discriminate proficiency level at higher and lower ends of the spectrum.

Finally, it could be quite informative to study the characteristics of expert raters who truly excel in their scoring performance. The general literature on expertise has identified a number of common characteristics of experts, but these findings are largely based on work done with individuals capable of elite levels of performance. Given the relatively modest amount of training and experience provided in the present study, the raters are probably more accurately described as “competent” rather than “elite,” including the more-proficient raters identified as part of Research Question 2. However, a performance-based definition of expertise might be used to identify individuals who are extraordinarily accurate or consistent in their scoring, perhaps searching for such people
among the ranks of long-time scorers of an operational test. With such individuals it might then be possible to examine whether characteristics of experts described in the expertise literature are actually relevant to raters in language tests, such as the finding that experts are more efficient in chunking and processing information.

8.5. Conclusion

This goal of this study was to illuminate the nature of rater expertise within the context of a speaking performance test. Some degree of insight has been gained regarding the effect of training on raters, as well as the characteristics that separate more-proficient raters from their less proficient colleagues. It is hoped that the results of this study will contribute to a better understanding of what it means to be an expert rater, and will serve as a resource for those interested in investigating ways in which consistency and/or validity in rater-generated scores can be achieved and improved.
### Appendix A: Description of the tasks used in the TOEFL iBT Speaking Test

<table>
<thead>
<tr>
<th>Task</th>
<th>Format</th>
<th>Task description</th>
<th>Example prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tasks 1 &amp; 2</td>
<td>Simple prompt</td>
<td><strong>Prompt:</strong> Describe and give reasons regarding a familiar person, place, event, etc.</td>
<td>Choose a teacher you admire and explain why you admire him or her. Please include specific examples and details in your explanation.</td>
</tr>
<tr>
<td>Task 3</td>
<td>Integrated:</td>
<td><strong>Reading:</strong> Presents a campus issue (various possible genres) <strong>Listening:</strong> One or two speakers discuss the issue. <strong>Prompt:</strong> Describe the position taken by the speaker(s)</td>
<td>The woman expresses her opinion of the announcement by the university president. State her opinion and explain the reasons she gives for holding that opinion.</td>
</tr>
<tr>
<td>Task 4</td>
<td>Integrated:</td>
<td><strong>Reading:</strong> Short passage from an introductory academic text. <strong>Listening:</strong> Academic lecture excerpt on the same or related topic. <strong>Prompt:</strong> Integrate information provided in the materials.</td>
<td>The professor describes the behavior of horses and antelope in herds. Explain how their behavior is related to their suitability for domestication.</td>
</tr>
<tr>
<td>Task 5</td>
<td>Integrated:</td>
<td><strong>Listening:</strong> Conversation regarding a campus problem; two solutions proposed. <strong>Prompt:</strong> Describe the problem, then state your preferred solution and explain.</td>
<td>The students discuss two possible solutions to the woman’s problem. Describe the problem. Then state which of the two solutions you prefer and explain why.</td>
</tr>
<tr>
<td>Task 6</td>
<td>Integrated:</td>
<td><strong>Listening:</strong> Brief excerpt from a lecture on a general academic topic. <strong>Prompt:</strong> Describe information presented in the lecture.</td>
<td>Using points and examples from the talk, explain how the automobile and the radio contributed to a common culture in the United States.</td>
</tr>
</tbody>
</table>
## Appendix B: Scoring rubric used in the study

*(TOEFL iBT Speaking Test rubric adapted to a 6-point scale)*

<table>
<thead>
<tr>
<th>Score</th>
<th>General Description</th>
<th>Delivery</th>
<th>Language Use</th>
<th>Topical Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>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:</td>
<td>Generally well-placed 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.</td>
<td>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.</td>
<td>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).</td>
</tr>
<tr>
<td>5</td>
<td>The response is between a score of 4 and a score of 6.</td>
<td>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).</td>
<td>The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecision 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.</td>
<td>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.</td>
</tr>
<tr>
<td>4</td>
<td>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:</td>
<td>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).</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>The response is between a score of 2 and a score of 4.</td>
<td>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.</td>
<td>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).</td>
<td>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.</td>
</tr>
<tr>
<td>2</td>
<td>The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence may occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:</td>
<td>Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.</td>
<td>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.</td>
<td>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 task and may rely heavily on repetition of the prompt.</td>
</tr>
<tr>
<td>1</td>
<td>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:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C. Academic and professional background of TOEFL scoring leaders

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Average</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TESOL Certificate</td>
<td>3</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATESOL or similar</td>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other M.A. degree(^a)</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ph.D degree</td>
<td>0</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scored other speaking tests?</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English teaching experience (years)</td>
<td></td>
<td></td>
<td>12.0</td>
<td>5.0</td>
<td>20.0</td>
</tr>
<tr>
<td>TOEFL scoring leader experience (years)</td>
<td></td>
<td></td>
<td>4.3</td>
<td>2.8</td>
<td>6.3</td>
</tr>
<tr>
<td>TOEFL typical monthly workload (hours)</td>
<td></td>
<td></td>
<td>80.0</td>
<td>20.0</td>
<td>160.0</td>
</tr>
</tbody>
</table>

\(^a\)Two individuals had both an MATESOL degree and an M.A. degree in another field.
Appendix D: FACETS results for ETS scoring leaders

<table>
<thead>
<tr>
<th>Measure</th>
<th>Examinee</th>
<th>Rater</th>
<th>Task</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ 7 +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>(7) +</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 6 +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 5 *</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 4 +</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 3 +</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 2 +</td>
<td>**</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ 1 +</td>
<td>***</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>****</td>
<td>12</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>11</td>
<td>4</td>
<td>Task A</td>
</tr>
<tr>
<td>* 0 **</td>
<td>**</td>
<td>3</td>
<td>6</td>
<td>Task B</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>2</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ -1 +</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ -2 +</td>
<td>**</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ -3 +</td>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ -4 +</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* = 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure D.1. ETS Scoring leaders FACETS vertical rulers.
Table D.1

*ETS Scoring leaders rater measurement report*

<table>
<thead>
<tr>
<th>Count</th>
<th>Measure</th>
<th>S.E.</th>
<th>MnSq</th>
<th>zStd</th>
<th>MnSq</th>
<th>zStd</th>
<th>Rater</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>0.87</td>
<td>0.05</td>
<td>0.94</td>
<td>-0.9</td>
<td>0.94</td>
<td>-1.0</td>
<td>REF05</td>
</tr>
<tr>
<td>600</td>
<td>0.74</td>
<td>0.05</td>
<td>0.75</td>
<td>-4.5</td>
<td>0.75</td>
<td>-4.4</td>
<td>REF12</td>
</tr>
<tr>
<td>599</td>
<td>0.37</td>
<td>0.05</td>
<td>1.44</td>
<td>6.7</td>
<td>1.55</td>
<td>7.6</td>
<td>REF11</td>
</tr>
<tr>
<td>600</td>
<td>0.19</td>
<td>0.05</td>
<td>0.96</td>
<td>-0.6</td>
<td>1.02</td>
<td>0.3</td>
<td>REF04</td>
</tr>
<tr>
<td>600</td>
<td>0.05</td>
<td>0.05</td>
<td>0.89</td>
<td>-1.8</td>
<td>0.86</td>
<td>-2.3</td>
<td>REF01</td>
</tr>
<tr>
<td>600</td>
<td>-0.14</td>
<td>0.05</td>
<td>0.98</td>
<td>-0.4</td>
<td>0.98</td>
<td>-0.2</td>
<td>REF03</td>
</tr>
<tr>
<td>600</td>
<td>-0.16</td>
<td>0.05</td>
<td>0.95</td>
<td>-0.9</td>
<td>0.96</td>
<td>-0.6</td>
<td>REF06</td>
</tr>
<tr>
<td>600</td>
<td>-0.41</td>
<td>0.05</td>
<td>1.27</td>
<td>4.3</td>
<td>1.23</td>
<td>3.3</td>
<td>REF07</td>
</tr>
<tr>
<td>600</td>
<td>-0.43</td>
<td>0.05</td>
<td>0.98</td>
<td>-0.4</td>
<td>0.92</td>
<td>-1.1</td>
<td>REF08</td>
</tr>
<tr>
<td>599</td>
<td>-0.44</td>
<td>0.05</td>
<td>0.76</td>
<td>-4.6</td>
<td>0.73</td>
<td>-4.4</td>
<td>REF02</td>
</tr>
<tr>
<td>600</td>
<td>-0.66</td>
<td>0.05</td>
<td>1.07</td>
<td>1.1</td>
<td>1.07</td>
<td>1.0</td>
<td>REF09</td>
</tr>
</tbody>
</table>

Model, Populn: RMSE .05 Adj (True) S.D. .47 Separation 9.29 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 963.2 d.f.: 10 significance (probability): .00
Inter-Rater agreement opportunities: 56397 Exact agreements: 21702 = 38.5%
Expected: 20289.4 = 36.0%
Appendix E: Professional background questionnaire

Educational and Professional Background Questionnaire
The purpose of this questionnaire is to collect information which will be used
to better understand your scores. Please answer all the questions that apply to you.
(Note: Entering text that is longer than the text box is no problem - the text box will scroll.)

### Educational Background

1) What degrees and certifications have you completed? (Please check all that apply.)

- [ ] Associate (2 year) Degree Major:
- [ ] Bachelor (4 year) Degree Major:
- [ ] Certificate in TESOL/TEFL
- [ ] Master’s Degree in TESOL/TEFL/Applied Linguistics
- [ ] Master’s Degree in another field Major:
- [ ] PhD in SLA/Second Language Studies/Applied Linguistics
- [ ] PhD in another field Major:

Other degrees or professional certifications (please list):

2) Are you currently a student?  [ ] Yes  [ ] No

If yes, please give the degree and program (e.g., M.A. in TESOL)

3) Have you taken any courses in language testing?  [ ] Yes  [ ] No

If yes, please briefly describe:

4) Have you taken any other coursework related to the teaching or analysis of spoken language?  [ ] Yes  [ ] No

If yes, please briefly describe:

5) Have you done any research involving analysis of spoken language (including discourse)?  [ ] Yes  [ ] No

If yes, please briefly describe:
Professional Background

6) How many years have you taught ESL/EFL?

7) Are you currently teaching ESL/EFL, or have you taught ESL/EFL learners in the last 12 months?  □ Yes  □ No

8) Please briefly describe your teaching experience.
If you have extensive teaching experience, just briefly describe your typical teaching situation(s):
   Country(ies):
   School/institutional context(s):
   Types of classes/subjects taught:

Student characteristics
   Age(s):
   Native Language(s):
   English proficiency level:
   Academic level (if applicable):

9) Have you had any other extensive exposure to non-native speakers of English, outside of teaching?
   Please briefly describe. (Examples: living abroad, international marriage/relationship, etc.)

10) Have you ever worked as a TOEFL iBT speaking test rater?  □ Yes  □ No

11) Have you ever taught a TOEFL preparation class?  □ Yes  □ No
If yes, please briefly describe. Please state if this included (or not) the TOEFL iBT speaking test:

11) Have you ever scored any other speaking tests (not including classroom tests)?  □ Yes  □ No
If yes, please briefly describe. The test name is enough for large-scale standardized tests. For local program tests (placement, achievement, etc.), please briefly describe a few aspects of the context, such as: (1) the test format (e.g., individual interview, pair work, immediate "live" scoring or taped responses, etc.), (2) the type of scoring (e.g., holistic scale/one score, analytic scale/several subscores), and (3) how often and over how long a period you worked as a scorer (e.g., twice a year for two years).
### Appendix F: Modified LEAP-Q Questionnaire

This questionnaire is adapted from Marian, Blumenfeld, & Kaushanskaya (2007) and is shown containing partial sample data.

<table>
<thead>
<tr>
<th>Language Experience and Proficiency Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>The purpose of this questionnaire is to collect information which will be used to better understand your scores. (Note: Entering text that is longer than the text box is no problem - the text box will scroll.)</td>
</tr>
</tbody>
</table>

**Today's Date:** Oct 18, 2011  
**Age:** 30  
**Date of Birth:** 7/7/1981  
**Gender:** Male

1. **Please list the languages you know in order of acquisition (if you know more than 5, list the first five):**
   - 1. English
   - 2. Spanish
   - 3. Japanese
   - 4.
   - 5.

2. **Please list the languages you know in order of overall dominance (strongest or easiest to use; give the strongest five, if you know more):**
   - 1. English
   - 2. Japanese
   - 3. Spanish
   - 4.
   - 5.

3. **What percentage of the time, on average, do you do the following with each language? (Percentages should add up to 100%)**
   - **Listen**
     - 1. English  
       - % 90  
       - Total %: 100
     - 2. Japanese  
       - % 10  
     - 3. Spanish  
       - % 0
     - 4.  
     - 5.  

   - **Speak**
     - 1. English  
       - % 99  
       - Total %: 100
     - 2. Japanese  
       - % 1  
     - 3. Spanish  
       - % 0
     - 4.  
     - 5.  

   - **Read**
     - 1. English  
       - % 90  
       - Total %: 100
     - 2. Japanese  
       - % 10  
     - 3. Spanish  
       - % 0
     - 4.  
     - 5.  

   - **Write**
     - 1. English  
       - % 90  
       - Total %: 100
     - 2. Japanese  
       - % 10  
     - 3. Spanish  
       - % 0
     - 4.  
     - 5.  

226
All questions below refer to your knowledge of English

(1) At what age did you begin acquiring this language? □

(2) Please list the number of years and months you spent in...
   a. a country where English is spoken: Years □ Months □
   b. a family where English is spoken: Years □ Months □
   c. a school and/or working environment where English is spoken: Years □ Months □

(3) On a scale from zero to ten, please select your level of proficiency in speaking, listening, reading, and writing English
(0 = no ability to use the language; 10 = native-like ability to use the language)
   Speaking □ Listening □ Reading □ Writing □

Language: Japanese

All questions below refer to your knowledge of Japanese

(1) At what age did you begin acquiring this language? □

(2) Please list the number of years and months you spent in...
   a. a country where Japanese is spoken: Years □ Months □
   b. a family where Japanese is spoken: Years □ Months □
   c. a school and/or working environment where Japanese is spoken: Years □ Months □

(3) On a scale from zero to ten, please select your level of proficiency in speaking, listening, reading, and writing Japanese
(0 = no ability to use the language; 10 = native-like ability to use the language)
   Speaking □ Listening □ Reading □ Writing □

Language: Spanish

All questions below refer to your knowledge of Spanish

(1) At what age did you begin acquiring this language? □

(2) Please list the number of years and months you spent in...
   a. a country where Spanish is spoken: Years □ Months □
   b. a family where Spanish is spoken: Years □ Months □
   c. a school and/or working environment where Spanish is spoken: Years □ Months □

(3) On a scale from zero to ten, please select your level of proficiency in speaking, listening, reading, and writing Spanish
(0 = no ability to use the language; 10 = native-like ability to use the language)
   Speaking □ Listening □ Reading □ Writing □
Appendix G: Scoring instrument

This example was the scoring instrument used for prompt A, scoring session 2. Only the first two pages of the scoring instrument are shown; in a scoring session these were followed by two more full pages of responses to score, and final fifth page with ten responses used for doing verbal reports (see Appendix J). In this example, the rubric and exemplars have been reviewed, and response number 7 is being scored.

<table>
<thead>
<tr>
<th>Scoring Session 2, Topic A</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Instructions</td>
</tr>
</tbody>
</table>

**Setting up (see the tech notes for details regarding setup and scoring)**

- Please set up for the session in a place where you can record your voice and won't be disturbed.
- Before starting, make sure desired browser is set to "default."
- Computer desktop arrangement:
  - Make sure the pdf is large enough that you can comfortably read and navigate the form.
  - Arrange the pdf, browser, media player, and recorder as shown in the figure (for easy navigation).
- Try a test recording to make sure recorder and microphone are working. Adjust record/play levels as needed.
- You may want to create a folder to store your finished pdfs and recordings for upload.

**Workflow**

The scoring session consists of 2 parts, each covering a different TOEFL topic.

- **Please complete all of this scoring session (both Part 1 and Part 2) in one day (important).**
- But, please take breaks as needed.
- When you first save the pdf, please add a "done" at the end: (Original file name) done.pdf
- Afterwards, just use save when prompted, and "replace the existing file."
- For stimulated recall recording, use the name given for the item (e.g., 112202 Recall.mp3)
- Parts 1 and 2 of the scoring session will likely each take 2-2.5 hours to complete (for 4-5 hours total).

**Use of Rubric/Exemplars**

You must review the rubric and listen to the exemplars to unlock the test taker responses for scoring.

You are welcome to go back to the rubric and refer to the exemplars whenever you wish during regular scoring.

For the stimulated recall, please go directly to your recall after scoring - no additional checking of the rubric, etc.

**Navigation**

Please use the "listen" buttons in the pdf to open and listen to files - do not dig through files already open in the media player.

(This helps prevent confusion as to which audio file you are actually listening to.)

**After you finish...**

- When you finish, please upload to your Google Docs account all of the pdfs you used as well as your recordings.
- Be sure to share the files with me.
- Delete the test taker responses (mp3) from your "downloads" folder, or wherever the browser saved the files.

**If you have questions or experience problems**

Please contact me via either email (deville@hawaii.edu) or phone (808-294-5285).
Session 2, Topic A

1. Rubric
   Please review the rubric.

2. Exemplars
   The test takers were asked to answer the following question:
   Students work hard but they also need to relax. What do you think is the best way for a student to relax after working hard? Explain why.

   Please review the exemplars before starting scoring.
   Once you have listened to the first exemplar, the next exemplar will appear.
   When you have listened to all the exemplars once, click "Done" to reset the exemplars.
   You can now listen to the exemplars as you like. But, please do not take notes - just listen.

   Score = 1  Score = 2  Score = 3  Score = 4  Score = 5  Score = 6

   Done

3. Scoring
   1. To hear the recording, click on the green "Listen" button; select the score from the drop down box.
   2. Comments are optional, but encouraged if you notice anything notable, or experience technical or other problems.

   Students work hard but they also need to relax. What do you think is the best way for a student to relax after working hard? Explain why.

   1)  
   2)  
   3)  
   4)  
   5)  
   6)  
   7)  
   8)  
   9)  
   10)
Appendix H: Scoring session wrap-up questionnaire

**Scoring Session 2**
**Wrap Up**

Please take a moment to answer the questions below regarding the scoring session you just finished (parts 1 and 2). Then, please be sure to upload your finished files (pdfs and recordings) to Google Docs when finished with the scoring session (parts 1 & 2). After uploading, remember to “share” the documents with me.

For more details, please see the technical notes. Also, please contact me if you have any questions: [opens tech notes]

---

**How did the scoring session go? In your opinion...**

How accurate are your scores? (Do they match the scale represented by the exemplars)? __________

Generally, how easy was it to decide what score to give? __________

How consistent is your scoring? __________

---

**Please comment on your use of the rubrics/exemplars while scoring the 100 responses in the scoring session.** *(Not including the stimulated recall.)* *Please be honest - there is no "right answer"*

While scoring, I checked the rubric ___% of the time. __________

While scoring, I checked one or more of the topic exemplars ___% of the time. __________

---

**Optional: Any other comments regarding the session covered in this pdf (including the stimulated recalls)?**

[space for comments]
Appendix I: Common person linking plots.

These plots compare logit ability measures for repeated examinees, produced when the data were analyzed separately by scoring session versus measures produced when all sessions were combined in the analysis. The squares show ability measures (in logits) while the lines are error bars for the 95% confidence interval. The fact that the data fall on a line with slope of 1 (within error) suggest that either approach produces equivalent measures, and local dependence of raters in the combined analysis has not substantively influenced the results (John Linacre, personal communication, June 25, 2011).

Figure I.1. Examinee measures (logits) calculated from session 1 vs. all data combined

Figure I.2. Examinee measures (logits) calculated from session 2 vs. all data combined
Figure I.3. Examinee measures (logits) calculated from session 3 vs. all data combined

Figure I.4. Examinee measures (logits) calculated from session 4 vs. all data combined
Appendix J: Instrument for collecting verbal report data

This example was the fifth page of the scoring instrument used for prompt A, scoring session 2. In this example, response number 7 has been scored, with the stimulated recall to be done next. The stimulated recall checklist is shown open here and is covering the first three items; normally this would be hidden and opened by clicking on a button, if needed.

4. Stimulated Recall
In this section, you will score several test taker responses as above. Immediately after scoring the response, you will re-play the response and record what you were thinking when you scored the response. Please pause the response and record you thoughts every 10-15 seconds (at least 2-3 times per response). It is also okay to talk while the response is playing, or replay the response (in whole or part).

As you replay the response, please describe what comes to mind as you listen to the test taker, following your original thoughts if you can. (But, this is not an explanation or rationale for your scoring. Rather, I am looking for your “stream of thought” as you listen.)

Some questions to help guide your recall:
- What do you notice about the test taker’s performance?
- What score comes to mind?
- Is anything particularly notable or important?

Students work hard but they also need to relax. What do you think is the best way for a student to relax after working hard? Explain why.

Stimulated Recall Checklist
1. Review the recall Instructions before starting this section, and as needed while working.
2. Score the test taker performance.
3. Start recording with your sound recorder (Windows Sound Recorder, QuickTime, Audacity, etc.).
   - Play the test taker response over the computer speakers - please don’t use headphones.
4. Replay the test taker response and talk about what is going through your mind.
5. While talking, you may pause or replay the test taker response. But, do not stop recording until you are finished talking.
6. Save the recording when you are finished with the recall, using the name provided (copy & paste the part in blue).
7. Return to this pdf, and click "OK" (you will not be able to change your score after this).
8. When asked if you saved the file, click "yes" or "no" as appropriate. If "no" you will be returned to the item; if "yes" then click "done" and the next item will appear.

4)  
5)  
6)  
7) Listen  Score  2  Please record your recall & label the file: 112227 Recall OK
8)  
9)  
10)
## Appendix K: categories used to code verbal recall data.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery</td>
<td>Comments regarding pronunciation, fluency, and intelligibility generally</td>
<td>“good rate of speaking” “I didn’t get that word”</td>
</tr>
<tr>
<td>Language Use</td>
<td>Comments regarding range and control of grammar and vocabulary,</td>
<td>so “the relaxation” “the soft music” a couple things with the grammar for use of articles</td>
</tr>
<tr>
<td>Topical Development</td>
<td>Comments regarding elaboration of the topic, precision and coherence of</td>
<td>“okay so that idea's just a little bit incomplete”</td>
</tr>
<tr>
<td></td>
<td>content</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>Comments made regarding the general quality of the response without</td>
<td>“okay so he starts very strong”</td>
</tr>
<tr>
<td></td>
<td>reference to a particular feature</td>
<td></td>
</tr>
<tr>
<td>Unclear</td>
<td>Comments for which the language focus is unclear (such as repetition of the</td>
<td>“we will we will” [Examinee: “in the morning we will feel sleepy and we will, we will, stay at, up…”]</td>
</tr>
<tr>
<td></td>
<td>examinee's utterance without further comment)</td>
<td></td>
</tr>
<tr>
<td><strong>Scoring</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score mention</td>
<td>Comments where a score for part or all of the response is mentioned</td>
<td>“so I actually, thought at this point maybe this is a two”</td>
</tr>
<tr>
<td>Scoring process</td>
<td>Comments regarding the scoring process itself such a reference to the rubric</td>
<td>“I tend to, grade a little bit harsher than what happened in the rater</td>
</tr>
<tr>
<td></td>
<td>or other examinees.</td>
<td>training”</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical</td>
<td>Comments regarding the sound quality of the recorded response, or other</td>
<td>“the sound's a little bit difficult in the beginning of this one”</td>
</tr>
<tr>
<td></td>
<td>technology issues.</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Comments not fitting into the other categories, such as speculations</td>
<td>“which seems to be an indication of, the fact that they can’t think</td>
</tr>
<tr>
<td></td>
<td>regarding the background or intentions of the examinee.</td>
<td>spontaneously of new things to say”</td>
</tr>
</tbody>
</table>
Appendix L. Questions asked of raters after isolated and pairwise scoring sessions

**Questions asked for both isolated and pairwise judgment conditions**
1. How accurate are your scores? (Do they match the scale represented by the exemplars)?
   (Likert scale, 1-7; 1 = not accurate at all, 7 = highly accurate)
2. Generally, how easy was it to decide what score to give?
   (Likert scale, 1-7; 1 = very difficult, 7 = very easy)
3. How consistent is your scoring?
   (Likert scale, 1-7; 1 = not consistent at all, 7 = highly consistent)
4. Compared to the “regular” scoring earlier in this study, making a decision using this method was...
   (Likert scale, 1-7; 1 = much easier, 4 = about the same, 7 = much more difficult)
5. Any other comments regarding the session?
   (Text box)

**Questions asked for the isolated judgment condition only.**
1. For me, the arithmetic problems were...
   (Likert scale, 1-7; 1 = very easy, 7 = very difficult)
2. When giving a score, remembering what the previous test taker was like was...
   (Likert scale, 1-7; 1 = very easy, 7 = very difficult)
3. I thought about the previous test taker when deciding on a score.
   (Likert scale, 1-7; 1 = not at all, 7 = all the time)

**Questions asked for the pairwise judgment condition only.**
1. I decided on a score for each person, then compared the scores to decide “same” or “different”
   (Likert scale, 1-7; 1 = never, 7 = always)
2. I decided if the test takers were “same” or “different” without considering the score each should get.
   (Likert scale, 1-7; 1 = never, 7 = always)
3. I used a different method to decide “same” or “different”. (If yes, please describe in the comment box below.)
   (Likert scale, 1-7; 1 = never, 7 = always)
Appendix M. Average raw scores responses repeated across scoring sessions.

Table M.1

Average raw scores for a set of test taker responses (n = 16) repeated in all scoring sessions

<table>
<thead>
<tr>
<th>Rater</th>
<th>Session 1 Mean</th>
<th>Session 1 SD</th>
<th>Session 2 Mean</th>
<th>Session 2 SD</th>
<th>Session 3 Mean</th>
<th>Session 3 SD</th>
<th>Session 4 Mean</th>
<th>Session 4 SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>3.7</td>
<td>1.35</td>
<td>3.5</td>
<td>1.41</td>
<td>3.1</td>
<td>1.29</td>
<td>3.4</td>
<td>1.36</td>
</tr>
<tr>
<td>102</td>
<td>4.0</td>
<td>1.32</td>
<td>3.6</td>
<td>1.36</td>
<td>3.6</td>
<td>1.36</td>
<td>3.8</td>
<td>1.42</td>
</tr>
<tr>
<td>103</td>
<td>3.4</td>
<td>1.26</td>
<td>3.3</td>
<td>1.24</td>
<td>3.0</td>
<td>1.15</td>
<td>3.2</td>
<td>1.28</td>
</tr>
<tr>
<td>105</td>
<td>3.8</td>
<td>1.53</td>
<td>4.0</td>
<td>1.59</td>
<td>3.6</td>
<td>1.41</td>
<td>3.7</td>
<td>1.66</td>
</tr>
<tr>
<td>107</td>
<td>3.1</td>
<td>1.24</td>
<td>3.3</td>
<td>1.06</td>
<td>3.2</td>
<td>1.17</td>
<td>3.4</td>
<td>1.45</td>
</tr>
<tr>
<td>108</td>
<td>4.4</td>
<td>1.03</td>
<td>3.8</td>
<td>1.47</td>
<td>4.0</td>
<td>1.51</td>
<td>3.9</td>
<td>1.61</td>
</tr>
<tr>
<td>109</td>
<td>2.7</td>
<td>1.20</td>
<td>3.9</td>
<td>1.45</td>
<td>4.2</td>
<td>1.42</td>
<td>4.1</td>
<td>1.31</td>
</tr>
<tr>
<td>111</td>
<td>3.8</td>
<td>1.34</td>
<td>3.7</td>
<td>1.40</td>
<td>3.1</td>
<td>1.75</td>
<td>3.6</td>
<td>1.41</td>
</tr>
<tr>
<td>112</td>
<td>4.2</td>
<td>1.72</td>
<td>2.9</td>
<td>1.45</td>
<td>3.1</td>
<td>1.63</td>
<td>3.4</td>
<td>1.71</td>
</tr>
<tr>
<td>113</td>
<td>3.6</td>
<td>1.15</td>
<td>2.8</td>
<td>1.29</td>
<td>3.1</td>
<td>1.00</td>
<td>3.3</td>
<td>1.00</td>
</tr>
<tr>
<td>115</td>
<td>3.3</td>
<td>1.30</td>
<td>3.8</td>
<td>1.38</td>
<td>3.8</td>
<td>1.18</td>
<td>3.6</td>
<td>1.20</td>
</tr>
<tr>
<td>116</td>
<td>3.2</td>
<td>1.38</td>
<td>3.4</td>
<td>1.20</td>
<td>2.8</td>
<td>1.39</td>
<td>3.3</td>
<td>1.53</td>
</tr>
<tr>
<td>117</td>
<td>3.5</td>
<td>1.59</td>
<td>3.3</td>
<td>1.53</td>
<td>3.2</td>
<td>1.33</td>
<td>3.4</td>
<td>1.26</td>
</tr>
<tr>
<td>119</td>
<td>3.6</td>
<td>1.41</td>
<td>3.7</td>
<td>1.58</td>
<td>3.6</td>
<td>1.46</td>
<td>3.6</td>
<td>1.36</td>
</tr>
<tr>
<td>120</td>
<td>4.2</td>
<td>1.01</td>
<td>4.2</td>
<td>1.28</td>
<td>3.8</td>
<td>1.05</td>
<td>3.9</td>
<td>1.54</td>
</tr>
<tr>
<td>121</td>
<td>3.4</td>
<td>1.41</td>
<td>3.4</td>
<td>1.82</td>
<td>3.4</td>
<td>1.71</td>
<td>3.4</td>
<td>1.59</td>
</tr>
<tr>
<td>122</td>
<td>3.7</td>
<td>1.01</td>
<td>2.9</td>
<td>1.31</td>
<td>3.2</td>
<td>1.56</td>
<td>3.4</td>
<td>1.26</td>
</tr>
<tr>
<td>123</td>
<td>3.8</td>
<td>1.47</td>
<td>4.0</td>
<td>1.26</td>
<td>3.8</td>
<td>1.48</td>
<td>3.6</td>
<td>1.31</td>
</tr>
<tr>
<td>124</td>
<td>3.6</td>
<td>1.03</td>
<td>3.1</td>
<td>1.12</td>
<td>3.4</td>
<td>0.89</td>
<td>3.3</td>
<td>0.93</td>
</tr>
<tr>
<td>125</td>
<td>3.1</td>
<td>1.36</td>
<td>2.9</td>
<td>1.39</td>
<td>3.6</td>
<td>1.45</td>
<td>3.8</td>
<td>1.29</td>
</tr>
</tbody>
</table>

Mean 3.6 3.5 3.4 3.5
SD 0.42 0.42 0.37 0.26
Appendix N. Time required to make scoring decisions

Table N.1
The average time required to make a scoring decision (in seconds). Breaks are periods of 5 minutes or more occurring between examines, or instances where scoring lasted more than 5 minutes.

<table>
<thead>
<tr>
<th></th>
<th>Session 1</th>
<th></th>
<th>Session 2</th>
<th></th>
<th>Session 3</th>
<th></th>
<th>Session 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>N</td>
<td>Breaks</td>
<td>M</td>
<td>SD</td>
<td>N</td>
<td>Breaks</td>
</tr>
<tr>
<td>R101</td>
<td>91.5</td>
<td>49.49</td>
<td>92</td>
<td>4</td>
<td>101.3</td>
<td>50.59</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>R102</td>
<td>81.5</td>
<td>39.64</td>
<td>93</td>
<td>5</td>
<td>74.5</td>
<td>31.02</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>R103</td>
<td>54.6</td>
<td>8.09</td>
<td>96</td>
<td>1</td>
<td>64.8</td>
<td>37.95</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>R105</td>
<td>69.0</td>
<td>38.34</td>
<td>91</td>
<td>7</td>
<td>65.9</td>
<td>32.31</td>
<td>92</td>
<td>6</td>
</tr>
<tr>
<td>R107</td>
<td>81.7</td>
<td>47.57</td>
<td>95</td>
<td>3</td>
<td>55.9</td>
<td>16.98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>R108</td>
<td>58.4</td>
<td>21.90</td>
<td>98</td>
<td>0</td>
<td>57.1</td>
<td>11.74</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>R109</td>
<td>58.9</td>
<td>11.86</td>
<td>97</td>
<td>0</td>
<td>64.6</td>
<td>26.84</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>R111</td>
<td>76.6</td>
<td>26.83</td>
<td>92</td>
<td>6</td>
<td>69.9</td>
<td>25.52</td>
<td>97</td>
<td>1</td>
</tr>
<tr>
<td>R112</td>
<td>80.2</td>
<td>39.07</td>
<td>94</td>
<td>4</td>
<td>63.0</td>
<td>32.59</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>R113</td>
<td>84.1</td>
<td>33.43</td>
<td>96</td>
<td>1</td>
<td>103.9</td>
<td>45.84</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>R115</td>
<td>67.8</td>
<td>20.35</td>
<td>96</td>
<td>2</td>
<td>76.1</td>
<td>25.75</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>R116</td>
<td>75.8</td>
<td>45.72</td>
<td>94</td>
<td>4</td>
<td>64.3</td>
<td>13.11</td>
<td>88</td>
<td>10</td>
</tr>
<tr>
<td>R117</td>
<td>78.6</td>
<td>40.43</td>
<td>89</td>
<td>9</td>
<td>60.7</td>
<td>29.32</td>
<td>92</td>
<td>6</td>
</tr>
<tr>
<td>R119</td>
<td>76.3</td>
<td>34.11</td>
<td>95</td>
<td>3</td>
<td>77.3</td>
<td>33.58</td>
<td>94</td>
<td>4</td>
</tr>
<tr>
<td>R120</td>
<td>72.8</td>
<td>25.11</td>
<td>97</td>
<td>1</td>
<td>78.6</td>
<td>27.13</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>R121</td>
<td>76.9</td>
<td>25.65</td>
<td>94</td>
<td>4</td>
<td>110.3</td>
<td>51.50</td>
<td>91</td>
<td>7</td>
</tr>
<tr>
<td>R122</td>
<td>95.5</td>
<td>58.40</td>
<td>93</td>
<td>4</td>
<td>103.4</td>
<td>52.67</td>
<td>90</td>
<td>8</td>
</tr>
<tr>
<td>R123</td>
<td>87.5</td>
<td>51.04</td>
<td>92</td>
<td>6</td>
<td>101.0</td>
<td>50.93</td>
<td>83</td>
<td>15</td>
</tr>
<tr>
<td>R124</td>
<td>62.4</td>
<td>20.16</td>
<td>96</td>
<td>2</td>
<td>87.1</td>
<td>50.68</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td>R125</td>
<td>65.6</td>
<td>24.45</td>
<td>96</td>
<td>2</td>
<td>66.6</td>
<td>33.05</td>
<td>97</td>
<td>1</td>
</tr>
<tr>
<td>Mean</td>
<td>74.8</td>
<td>3.4</td>
<td>77.3</td>
<td>4.0</td>
<td>74.6</td>
<td>3.4</td>
<td>74.4</td>
<td>2.4</td>
</tr>
<tr>
<td>SD</td>
<td>11.13</td>
<td>2.39</td>
<td>3.92</td>
<td>3.78</td>
<td>3.77</td>
<td>3.00</td>
<td>3.76</td>
<td>1.82</td>
</tr>
</tbody>
</table>
Appendix O. Scoring patterns of more-proficient, less-proficient, and improving raters.

Figure O.1. Rater severity.

Figure O.2. Infit mean square values.
Figure O.3. Rater accuracy in terms of Pearson correlation with reference scores.
Appendix P. Raw score scale boundaries

Table P.1
Measures (in logits) of raw score scale boundaries for more-proficient, less-proficient, and improving raters.

<table>
<thead>
<tr>
<th>Raw Score Boundary</th>
<th>1,2</th>
<th>2,3</th>
<th>3,4</th>
<th>4,5</th>
<th>5,6</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Data Measure</td>
<td>-3.45</td>
<td>-1.56</td>
<td>0.12</td>
<td>1.7</td>
<td>3.2</td>
</tr>
<tr>
<td>SE</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>More-proficient raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 101 Measure</td>
<td>-2.59</td>
<td>-1.74</td>
<td>-0.17</td>
<td>1.55</td>
<td>2.95</td>
</tr>
<tr>
<td>SE</td>
<td>0.22</td>
<td>0.16</td>
<td>0.12</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td>Rater 102 Measure</td>
<td>-3.18</td>
<td>-1.61</td>
<td>-0.2</td>
<td>1.62</td>
<td>3.36</td>
</tr>
<tr>
<td>SE</td>
<td>0.28</td>
<td>0.17</td>
<td>0.13</td>
<td>0.13</td>
<td>0.24</td>
</tr>
<tr>
<td>Rater 123 Measure</td>
<td>-3.12</td>
<td>-1.45</td>
<td>0.23</td>
<td>1.46</td>
<td>2.88</td>
</tr>
<tr>
<td>SE</td>
<td>0.3</td>
<td>0.17</td>
<td>0.13</td>
<td>0.13</td>
<td>0.19</td>
</tr>
<tr>
<td>Less-proficient raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 108 Measure</td>
<td>-2.68</td>
<td>-1.07</td>
<td>-0.51</td>
<td>1.05</td>
<td>3.21</td>
</tr>
<tr>
<td>SE</td>
<td>0.31</td>
<td>0.2</td>
<td>0.15</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>Rater 109 Measure</td>
<td>-2.63</td>
<td>-1.77</td>
<td>-0.26</td>
<td>1.91</td>
<td>2.75</td>
</tr>
<tr>
<td>SE</td>
<td>0.26</td>
<td>0.18</td>
<td>0.13</td>
<td>0.13</td>
<td>0.2</td>
</tr>
<tr>
<td>Rater 113 Measure</td>
<td>-4.24</td>
<td>-1.85</td>
<td>0.09</td>
<td>1.76</td>
<td>4.24</td>
</tr>
<tr>
<td>SE</td>
<td>0.26</td>
<td>0.14</td>
<td>0.12</td>
<td>0.18</td>
<td>0.56</td>
</tr>
<tr>
<td>Improving raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 117 Measure</td>
<td>-3.24</td>
<td>-1.02</td>
<td>0.21</td>
<td>1.74</td>
<td>2.32</td>
</tr>
<tr>
<td>SE</td>
<td>0.24</td>
<td>0.14</td>
<td>0.13</td>
<td>0.15</td>
<td>0.2</td>
</tr>
<tr>
<td>Rater 120 Measure</td>
<td>-4</td>
<td>-1.35</td>
<td>0.24</td>
<td>1.88</td>
<td>3.23</td>
</tr>
<tr>
<td>SE</td>
<td>0.49</td>
<td>0.19</td>
<td>0.13</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>Rater 125 Measure</td>
<td>-3.07</td>
<td>-1.13</td>
<td>0.18</td>
<td>1.11</td>
<td>2.91</td>
</tr>
<tr>
<td>SE</td>
<td>0.22</td>
<td>0.14</td>
<td>0.13</td>
<td>0.14</td>
<td>0.23</td>
</tr>
</tbody>
</table>
## Table Q.1

Raw counts of comments made during stimulated recall

<table>
<thead>
<tr>
<th></th>
<th>More-proficient</th>
<th>Less-proficient</th>
<th>Developing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R101 R102 R123</td>
<td>R108 R109 R113</td>
<td>R117 R120 R125</td>
</tr>
<tr>
<td><strong>Session 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery</td>
<td>41 31 11</td>
<td>9 28 15</td>
<td>11 10 14</td>
</tr>
<tr>
<td>Use</td>
<td>28 35 27</td>
<td>10 26 19</td>
<td>15 13 10</td>
</tr>
<tr>
<td>Topic</td>
<td>25 16 39</td>
<td>19 5 33</td>
<td>28 11 21</td>
</tr>
<tr>
<td>General</td>
<td>1 10 5</td>
<td>18 0 3</td>
<td>2 19 5</td>
</tr>
<tr>
<td>Unclear</td>
<td>0 7 4</td>
<td>10 3 2</td>
<td>5 10 4</td>
</tr>
<tr>
<td>Misc</td>
<td>6 2 9</td>
<td>7 2 11</td>
<td>0 10 5</td>
</tr>
<tr>
<td>Scoring</td>
<td>6 3 14</td>
<td>2 5 11</td>
<td>11 17 22</td>
</tr>
<tr>
<td>Technical</td>
<td>0 3 0</td>
<td>3 6 1</td>
<td>0 1 0</td>
</tr>
<tr>
<td><strong>Total segments</strong></td>
<td>107 107 109</td>
<td>78 75 95</td>
<td>72 91 81</td>
</tr>
<tr>
<td><strong>Total Utterances</strong></td>
<td>32 60 55</td>
<td>35 64 47</td>
<td>27 35 29</td>
</tr>
<tr>
<td><strong>Session 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery</td>
<td>42 34 39</td>
<td>7 30 16</td>
<td>12 21 10</td>
</tr>
<tr>
<td>Use</td>
<td>19 29 35</td>
<td>3 27 8</td>
<td>18 16 10</td>
</tr>
<tr>
<td>Topic</td>
<td>47 40 56</td>
<td>39 36 39</td>
<td>32 36 44</td>
</tr>
<tr>
<td>General</td>
<td>2 9 7</td>
<td>14 2 7</td>
<td>0 3 7</td>
</tr>
<tr>
<td>Unclear</td>
<td>1 5 4</td>
<td>10 1 11</td>
<td>3 5 3</td>
</tr>
<tr>
<td>Misc</td>
<td>3 1 6</td>
<td>4 13 2</td>
<td>0 6 7</td>
</tr>
<tr>
<td>Scoring</td>
<td>6 12 51</td>
<td>12 43 16</td>
<td>15 35 26</td>
</tr>
<tr>
<td>Technical</td>
<td>1 1 0</td>
<td>0 0 0</td>
<td>0 1 1</td>
</tr>
<tr>
<td><strong>Total segments</strong></td>
<td>121 131 198</td>
<td>89 152 99</td>
<td>80 123 108</td>
</tr>
<tr>
<td><strong>Total Utterances</strong></td>
<td>35 58 47</td>
<td>29 31 48</td>
<td>28 38 32</td>
</tr>
<tr>
<td><strong>Session 4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery</td>
<td>40 47 23</td>
<td>13 37 18</td>
<td>10 36 10</td>
</tr>
<tr>
<td>Use</td>
<td>15 35 34</td>
<td>8 44 19</td>
<td>21 9 23</td>
</tr>
<tr>
<td>Topic</td>
<td>32 32 19</td>
<td>41 36 17</td>
<td>26 39 28</td>
</tr>
<tr>
<td>General</td>
<td>0 0 5</td>
<td>16 0 4</td>
<td>5 4 1</td>
</tr>
<tr>
<td>Unclear</td>
<td>4 2 5</td>
<td>6 3 6</td>
<td>1 3 3</td>
</tr>
<tr>
<td>Misc</td>
<td>7 0 2</td>
<td>3 2 5</td>
<td>0 3 2</td>
</tr>
<tr>
<td>Scoring</td>
<td>6 17 41</td>
<td>13 46 27</td>
<td>12 16 33</td>
</tr>
<tr>
<td>Technical</td>
<td>0 0 0</td>
<td>0 0 1</td>
<td>0 1 0</td>
</tr>
<tr>
<td><strong>Total segments</strong></td>
<td>104 133 129</td>
<td>100 168 97</td>
<td>75 111 100</td>
</tr>
<tr>
<td><strong>Total Utterances</strong></td>
<td>31 58 38</td>
<td>33 34 46</td>
<td>26 37 30</td>
</tr>
</tbody>
</table>
REFERENCES


