Linguistic Laypersons’ Perspective on Second Language Oral Communication Ability

Takanori Sato

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

April 2014

Linguistics and Applied Linguistics
School of Languages and Linguistics
Faculty of Arts
The University of Melbourne
Abstract

Since the advent of the communicative movement, applied linguists or language specialists have theorized the ability to communicate in a second language (L2). Correspondingly, in general-purpose oral proficiency tests, language specialists (language testers or teachers) define L2 oral communication ability, determining the abilities or knowledge to be included in the criteria based on their professional perspective and/or theoretical models of communicative competence. Accordingly, ideal L2 oral communication, as conceived by communication theory and oral proficiency tests, reflects the perspectives of language specialists.

In contrast, the perspectives of linguistic laypersons (non-specialists in language) on L2 communication ability have not been incorporated into theories of communication ability and the construct definition of oral proficiency tests. This arguably seriously undermines the validity of the theories because, in the real world context, the ultimate arbiters of L2 speakers’ oral performance are largely not trained professionals in language. Furthermore, the construct definition relying on language specialists’ perspectives is flawed because a good performance as defined by language specialists is not necessarily identical to one perceived by linguistic laypersons. It is important for language tests to adopt, if possible, the perspectives of such laypersons on L2 communication ability in order to more accurately predict their impressions in real-life contexts. This study therefore investigates linguistic laypersons’ perspectives on L2 communication ability and aims to explore what features or behaviors of the speaker affect their intuitive judgments of this ability.

Twenty-three non-native and native English speaking graduate students from disciplines other than applied linguistics or TESOL participated in this study. First, they
watched video recordings of seven individual presentations on the College English Test-Spoken English Test and three paired interactions from the Cambridge English Examinations. Second, participants indicated their intuitive impressions of each speaker’s communication ability on a scale of 1 (Poor) to 7 (Excellent). Third, they provided justifications for their ratings. Fourth, they reviewed the performances and performed stimulated recall, verbalizing features of the performances that influenced their impressions. Finally, they responded to semi-structured interviews to provide supplementary information about their judgments. Features that had affected their impressions were explored in the verbal report data. The interview data were analyzed qualitatively to examine what features were perceived to most strongly influence their impressions.

The results showed that the outcome of communicative performance, including successful message conveyance and the quality of content, was one of the main criteria used to judge communication ability. Although linguistic features such as grammar, vocabulary, pronunciation, and fluency were recognized as factors impacting the outcome of communication, the participants neither considered them crucial nor penalized errors harshly unless comprehensibility was seriously impeded. Their impressions were also influenced by the test-taker’s non-verbal behavior and perceived level of confidence, which were found to be closely related. In addition, the participants frequently noted interactional features such as engagement and the size of contribution in the paired interactions.

The perspectives of linguistic laypersons can help define the features of oral communication ability valued in real-world contexts, which in turn can inform the assessment criteria used to measure this ability.
Declaration

This is to certify that:

i. the thesis comprises only my original work towards the PhD except where indicated in the Preface,

ii. due acknowledgement has been made in the text to all other material used,

iii. the thesis is fewer than 100 000 words in length, exclusive of tables, maps, bibliographies and appendices.

Takanori Sato
Acknowledgements

I would like to express my deepest gratitude to the following people for their support to my research. This thesis could not have been completed without their help.

I would like to thank my principal supervisor Professor Tim McNamara who warmly encouraged me and provided many professional insights throughout my candidacy. I also would like to thank my co-supervisor Associate Professor Catherine Elder who provided guidance even after her retirement. My thanks also go to my thesis committee panels who gave feedback on annual reports: Professor Gillian Wigglesworth, Dr. Carsten Roever, and Dr. Sally O’Hagan.

This thesis reports on research using examination data provided by the National College English Committee and Cambridge English Language Assessment. I am indebted to these two language testing organizations for providing invaluable video-recorded data for my research. Although they must remain anonymous, I also would like to thank all the participants for their interests in my research and cooperation.

I would like to acknowledge financial support from the Melbourne Scholarships at the University of Melbourne. I also would like to acknowledge the support of Educational Testing Service for providing me with a research grant.

I am also thankful to Dr. Kim Hyejeong, Naoki Ikeda, and Emi Fukasawa for their support, encouragement, and advice during my candidacy.

Finally, heartfelt thanks go to my parents and brother for their support in every way.
# Table of Contents

Abstract ........................................... i
Declaration ........................................ is
Acknowledgements ................................ iv
Table of Contents ................................... v
List of Tables ........................................ xi
List of Figures ....................................... xvi

**Chapter 1. Introduction** ............ 1
1.1 Background of the Study .............. 1
1.2 Aims of the Study .......................... 4
1.3 Outline of the Thesis .................. 6

**Chapter 2. Literature Review** ...... 8
2.1 Introduction .................................. 8
2.2 Theories of Communication .......... 8
   2.2.1 Evolution of Theories of L2 Communicative Competence .......... 8
   2.2.2 Interactional Competence .................. 16
   2.2.3 Empirically Based L2 Language Proficiency .................. 21
   2.2.4 L1 Communication Competence .................. 25
   2.2.5 Limitations of Theories of L2 Communication Ability .......... 32
2.3 The Need to Understand the Linguistic Laypersons’ Perspective ...... 34
   2.3.1 Non-linguists’ Views on L2 Communication .................. 35
   2.3.2 Language Specialists and Laypersons’ Perspectives on Communication ..... 39
Chapter 3. Methodology

3.1 Introduction

3.2 Methodological Considerations
   3.2.1 An Overview of Verbal Protocol Analysis
   3.2.2 Verbal Protocol Analysis Procedures in the Present Study

3.3 The CET-SET and the Cambridge English Examinations
   3.3.1 CET-SET
   3.3.2 Cambridge English Examinations

3.4 Methods
   3.4.1 Participants
   3.4.2 Instruments
   3.4.3 Data Collection Process
   3.4.4 Data Analyses
Chapter 4. Results

4.1 Introduction

4.2 Ratings of Rater Impressions
   4.2.1 Rater Measurement
   4.2.2 Test-taker Measurement

4.3 The Emerging Features Contributing to Raters’ Judgments

4.4 Demeanor
   4.4.1 Confidence
   4.4.2 Anxiety
   4.4.3 Other Demeanor Factors
   4.4.4 Differences among Proficiency Levels
   4.4.5 Differences among Raters

4.5 Non-verbal Behavior
   4.5.1 Body Movement
   4.5.2 Eye Contact
   4.5.3 Other Non-verbal Behavior
   4.5.4 Differences among Proficiency Levels
   4.5.5 Differences among Raters

4.6 Pronunciation
   4.6.1 Pronunciation
   4.6.2 Accent
   4.6.3 Prosody and Paralanguage
   4.6.4 Differences among Proficiency Levels
4.6.5 Differences among Raters 178

4.7 Linguistic Resources 178
  4.7.1 Vocabulary and Wording 179
  4.7.2 Grammar 185
  4.7.3 Other 190
  4.7.4 Differences among Proficiency Levels 191
  4.7.5 Differences among Raters 192

4.8 Fluency 193
  4.8.1 Overall Fluency 194
  4.8.2 Rate and Amount of Speech 195
  4.8.3 Pause Phenomena 197
  4.8.4 Repair Phenomena 203
  4.8.5 Differences among Proficiency Levels 204
  4.8.6 Differences among Raters 206

4.9 Content 206
  4.9.1 Ideas 207
  4.9.2 Framing of Idea 214
  4.9.3 Topical Knowledge 219
  4.9.4 Differences among Proficiency Levels 223
  4.9.5 Differences among Raters 224

4.10 Interaction 225
  4.10.1 Interaction and Engagement 226
  4.10.2 Interactional Pattern 231
  4.10.3 Differences among Proficiency Levels 234
  4.10.4 Differences among Raters 235
4.11 Overall Impression

4.11.1 Overall Performance and Global Ability
4.11.2 Overall Message Conveyance
4.11.3 Overall Comprehensibility
4.11.4 Overall English Proficiency
4.11.5 Differences among Proficiency Levels
4.11.6 Differences among Raters

4.12 Other

4.12.1 Miscellaneous Speech Features and Speaker Behaviors
4.12.2 Comments Unrelated to Speaker Behaviors
4.12.3 Rater’s General Perception
4.12.4 Fillers, Procedures, and Unclear Comments

4.13 Summary

Chapter 5. Discussion

5.1 Introduction
5.2 Restatement of Key Results
5.3 Factors in L2 Communication Ability

5.3.1 Overall Impression and Speech Content
5.3.2 Language Proficiency
5.3.3 Interactional Competence
5.3.4 Non-linguistic Factors
5.3.5 Other Factors

5.4 Transferability of the Findings
5.5 Summary
List of Tables

Table 2.1 Interactional Resources (Young, 2008, p. 71) 19
Table 2.2 Aspects of Interactional Competence 20
Table 2.3 Predictor Variables in De Jong et al. (2012) 24
Table 2.4 Core Criteria and Competences of Outcome- and Message-focused Approaches 28
Table 2.5 Components of Wiemann’s (1977) Communicative Competence 29
Table 2.6 Components of Spitzberg and Hurt’s (1987) Interpersonal Competence 31
Table 2.7 Statements Showing the Importance of Linguistic Laypersons’ Judgments 38
Table 2.8 Indigenous Assessment Criteria Applied by Physicists 45
Table 2.9 Indigenous Assessment Criteria Derived from Various Domain Experts 48
Table 2.10 Conceptual Differences Between EFL and ELF 52
Table 2.11 Innovative ELF Lexicogrammar (Cogo & Dewey, 2012; Seidlhofer, 2004, 2011) 53
Table 2.12 Research Approaches Employed by Previous Studies 75
Table 3.1 The CET-SET Assessment Criteria (Zhang & Elder, 2009, p. 303) 95
Table 3.2 Band Grade and Descriptors (Zhang & Elder, 2009, p. 304) 95
Table 3.3 Format of CET-SET (He & Dai, 2006, p. 376) 96
Table 3.4 The Characteristics of Individual Presentations on the CET-SET 97
Table 3.5 Cambridge English Examinations 98
Table 3.6 CAE Criteria (UCLES, n.d.) 99
Table 4.9  Frequency of Segments on Each Subcategory of Non-verbal Behavior 151
Table 4.10  Frequency of Segments on Non-verbal Behavior Given to the CET-SET Test-takers 162
Table 4.11  Frequency of Segments on Non-verbal Behavior Given to the Cambridge Exams Test-takers 163
Table 4.12  Number of Segments on Non-verbal Behavior Mentioned by Each Rater 164
Table 4.13  Frequency of Segments on Each Subcategory of Pronunciation 164
Table 4.14  Frequency of Segments on Pronunciation Given to the CET-SET Test-takers 177
Table 4.15  Frequency of Segments on Pronunciation Given to the Cambridge Exams Test-takers 177
Table 4.16  Number of Segments on Pronunciation Provided by Each Rater 178
Table 4.17  Frequency of Segments on Each Subcategory of Linguistic Resources 179
Table 4.18  Frequency of Segments on Linguistic Resources Given to the CET-SET Test-takers 191
Table 4.19  Frequency of Segments on Linguistic Resources Given to the Cambridge Exams Test-takers 192
Table 4.20  Number of Segments on Linguistic Resources Provided by Each Rater 193
Table 4.21  Frequency of Segments on Each Subcategory of Fluency 193
Table 4.22  Frequency of Segments on Fluency Given to the CET-SET Test-takers 205
Table 4.23  Frequency of Segments on Fluency Given to the Cambridge Exams Test-takers  
Table 4.24  Number of Segments on Fluency Provided by Each Rater  
Table 4.25  Frequency of Segments on Each Subcategory of Content  
Table 4.26  Frequency of Segments on Content Given to the CET-SET Test-takers  
Table 4.27  Frequency of Segments on Content Given to the Cambridge Exams Test-takers  
Table 4.28  Number of Segments on Content Provided by Each Rater  
Table 4.29  Frequency of Segments on Each Subcategory of Interaction  
Table 4.30  Frequency of Segments on Interaction Given to the CET-SET Test-takers  
Table 4.31  Frequency of Segments on Interaction Given to the Cambridge Exams Test-takers  
Table 4.32  Number of Segments on Interaction Provided by Each Rater  
Table 4.33  Frequency of Segments on Each Subcategory of Overall Impression  
Table 4.34  Frequency of Segments on Overall Impression Given to the CET-SET Test-takers  
Table 4.35  Frequency of Segments on Overall Impression Given to the Cambridge Exams Test-takers  
Table 4.36  Number of Segments on Overall Impression Provided by Each Rater  
Table 4.37  Frequency of Segments on Each Subcategory of Other  
Table 5.1  Features Mentioned by the Participants and Their Rankings in Terms of Frequency
Table 5.2 Feature that Raters Regarded as Most Influential

<table>
<thead>
<tr>
<th>Feature</th>
<th>Influence Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>92.5</td>
</tr>
<tr>
<td>Feature 2</td>
<td>87.8</td>
</tr>
<tr>
<td>Feature 3</td>
<td>91.6</td>
</tr>
<tr>
<td>Feature 4</td>
<td>89.7</td>
</tr>
<tr>
<td>Feature 5</td>
<td>90.1</td>
</tr>
</tbody>
</table>

Note: The table shows the features rated by raters and their respective influence scores.
List of Figures

Figure 2.1 Competence constellation (Munby, 1978, p. 21) 11
Figure 2.2 Chronological evolution of theories (Celce-Murcia et al., 1995, p. 11) 13
Figure 2.3 Components of language knowledge in CLA (Bachman, 1990, p. 87) 16
Figure 2.4 Core and periphery components of language proficiency (Hulstijn, 2011, p. 239) 25
Figure 2.5 Features heeded by raters according to test-takers’ proficiency (from Pollitt & Murray, 1996, p. 85) 71
Figure 3.1 Verbal protocol procedure in this study 92
Figure 3.2 Semantic differential scale for individual presentations 116
Figure 3.3 Semantic differential scale for paired conversations 117
Figure 3.4 Raters’ activities and data collected 120
Figure 4.1 Vertical map from the many-facet Rasch analysis 133
Figure 4.2 Relationships between confidence and other features 143
Figure 4.3 Relationships between anxiety and other features 145
Figure 4.4 Relationships between body movement and other features 155
Figure 4.5 Relationships between eye contact and other features 158
Figure 4.6 Relationships between other non-verbal behavior and other features 161
Figure 4.7 Relationships between pronunciation and other features 169
Figure 4.8 Relationships between accent and other features 172
Figure 4.9 Relationships between prosody/paralanguage and other features 176
| Figure 4.10 | Relationships between vocabulary/wording and other features | 184 |
| Figure 4.11 | Relationships between grammar and other features | 190 |
| Figure 4.12 | Relationships between the rate/amount of speech and other features | 197 |
| Figure 4.13 | Relationships between pause phenomena and other features | 203 |
| Figure 4.14 | Relationships between ideas and other features | 214 |
| Figure 4.15 | Relationships between framing of idea and other features | 219 |
| Figure 4.16 | Relationships between topical knowledge and other features | 223 |
| Figure 4.17 | Relationships between comprehensibility and other features | 247 |
| Figure 5.1 | Target and attributor loci of competency judgments (Spitzberg, 2000, p. 113) | 283 |
| Figure 6.1 | Theoretical implication of the present study | 287 |
Chapter 1 Introduction

1.1 Background of the Study

The ability to communicate is widely acknowledged as vital to a range of domains in our lives. Communication, however, is an extremely complex interpersonal act, and thus it is difficult to conceptualize what specific faculties it comprises. In the field of applied linguistics, scholars (i.e., language specialists) have conceptualized this ability for second language (L2) communication in particular, and have attempted to delineate the components of knowledge and ability it requires and developed various comprehensive theories to explicate it. These theories have provided “over-arching and relatively abstract theoretical descriptions of what it means to be able to communicate in a second language” (Fulcher & Davidson, 2007, p. 36) and played a crucial role in L2 teaching and assessment. With regard to the former, the theories have made substantial contributions to syllabus design, teaching methodology, teacher training, and materials development (Canale & Swain, 1980); in the latter, they have served as a framework of reference for defining the constructs of both specific- and general-purpose proficiency tests (Bachman & Palmer, 1996, 2010; Douglas, 2000).

Attempts to define L2 communication ability, or communicative competence (Hymes, 1972), began with Chomsky’s (1965) competence/performance distinction; its evolution continued as applied linguists proposed extended theories in order to provide more thorough explanations for elements of L2 communication (Purpura, 2008). The development of communication theories until the 1980s has been heavily driven by applied linguists’ criticisms of Chomsky’s linguistic theory (Spolsky, 1989). Subsequently, language specialists have modified and extended preexisting theories by adding new components or modifying terminology. In general then, the theories mainly
reflect applied linguists’ professional views on communication in their role as language specialists, normally consisting of language-related components with detailed explanations (e.g., grammatical, sociolinguistic knowledge) and paying less attention, if any, to non-linguistic cognitive, affective, and volitional factors. McNamara (1996) argues that the latter factors, which fall under what Hymes (1972) termed *ability for use*, have received less regard among linguists.

Some theories of L2 communication ability have been developed on the basis of empirical data, with the type of data most commonly incorporated being L2 speakers’ performances on language tests. During theory construction, test scores were statistically analyzed to find the relationships among different components of language knowledge and skills. Theories constructed through this approach are notable for having helped to establish the components of knowledge that have contributed to global language proficiency.

Accordingly, L2 communication theories exclusively reflect applied linguists’ professional views and/or L2 speakers’ actual language performance. Stated another way, linguists have not incorporated people’s subjective judgments of L2 communication in the process of theory construction, which may undermine the meaningfulness of theories thus developed. In essence, communication encompasses people’s attempts to share their thoughts and takes place among people (Hargie, 2011). Thus, it is an interpersonal act where people (unconsciously) judge others’ communication ability; it is the communicators themselves or observers who determine to what extent a speaker’s communication ability is adequate or satisfactory (Spitzberg, 2000). In other words, communication ability is not determined by observed behaviors or features themselves, but by people’s subjective and evaluative judgments of those behaviors or features. People’s subjective judgments should also determine what
behaviors or features comprise communication ability.

Another serious limitation of existing L2 communication theories is that theory construction to date has not invited the perspectives of linguistic laypersons (i.e., non-professionals in linguistics) regarding L2 communication ability. In fact, linguistic laypersons’ views on language and communication have generally been thought naïve and unimportant in the field of applied linguistics (Wilton & Stegu, 2011). However, their perspectives should be treated as a legitimate data source for defining L2 communication ability. One of the justifications for this impetus is that laypersons are in fact the ultimate arbiters of L2 speakers’ oral performance; L2 speakers are more likely to communicate with linguistic laypersons than with linguists and to be judged based on their unique perspectives. Another justification for investigating lay perspectives is that linguistic laypersons are likely to perceive L2 communication ability differently from language specialists; indeed, some empirical studies have demonstrated this (Brown, 1995; Elder, 1993; Galloway, 1980; Hadden, 1991). Training and experience as language specialists is considered to greatly influence and shape how such specialists judge L2 communication ability. Therefore, relying solely on linguists’ views regarding communication during theory construction limits the scope of the theory and undermines its value.

Despite the importance of factoring in linguistic laypersons’ perspectives on L2 communication, there is a dearth of studies investigating how they judge L2 oral performance. Instead, researchers have been concerned primarily with criteria contributing to language teachers’ and/or accredited raters’ overall judgments (e.g., Brown, Iwashita, & McNamara, 2005; McNamara, 1990; Sato, 2012). In particular, language-testing researchers have been investigating these language specialists’ judgments of oral performance in order to (a) check if raters insert construct-irrelevant
criteria into their ratings and/or (b) compare the evaluative judgments of non-native versus native English speakers. These studies do have important implications for how such raters choose to formulate global evaluations of test-takers’ oral performances. However, none of them touches on the perspectives of linguistic laypersons perspectives regarding L2 communication ability in general.

1.2 Aims of the Study

Given the problems outlined above about the character of existing theories of L2 communication and the process of their development, this study aimed to explore how linguistic laypersons conceive L2 communication ability. More specifically, the study investigated the features of oral performance and behaviors of speakers that influence their evaluative judgments of L2 speakers’ communication ability. Accordingly, oral communication ability is the ability that the study aimed to evaluate.

For the purposes of this study, linguistic laypersons are defined as those who do not have (a) specialized knowledge of applied linguistics, (b) experience of any training in language assessment and teaching, and (c) experience of rating and teaching L2 learners formally; that is, they are non-linguists without any evaluating or teaching experience. Specialized knowledge of applied linguistics entails knowledge of theories of L2 communication (e.g., communicative competence) and L2 pedagogical approaches (e.g., communicative language teaching). Metalinguistic knowledge is excluded from this specialized knowledge, which means that non-native English speakers who have learned a foreign language and can demonstrate basic linguistic knowledge (e.g., parts of speech) are selected as linguistic laypersons. The L2 performance the present study is concerned with is oral performance demonstrated by L2 speakers with a wide range of English proficiency from low-intermediate to
advanced levels. In other words, the communication ability that the present study aimed to explore does not necessarily include that of novice-level L2 speakers.

The investigation of linguistic laypersons’ evaluative judgments of oral performance contributes to the theory of L2 communication ability. In particular, this study delineates what behaviors of the speaker or features of communication are likely to be perceived positively or negatively by linguistic laypersons. Thus, unlike existing theories, the study provides an explanation of how L2 speakers’ performance is actually judged by interlocutors in the real world. In addition, the role of ability for use (Hymes, 1972) will be explored. As noted earlier, ability for use, which includes non-language-exclusive factors, has received little attention from language specialists (McNamara, 1996). However, by scrutinizing the view of linguistic laypersons, who are not necessarily concerned with language per se, it is possible to reveal the contribution of language-related ability and ability for use to their evaluative judgment of L2 communication. The salient components of the ability in their judgments are identified.

This study has some important implications for the assessment of L2 oral performance on a practical level. These implications are related to the crucial decisions to be made regarding assessment criteria: components and weights (Wigglesworth, 2008). First, the investigation of linguistic laypersons’ evaluative judgment of oral performance contributes to the construct definition of general-purpose oral proficiency tests. It is argued that the construct of oral proficiency tests assessing L2 communicative ability should be defined on the basis of the criteria that linguistic laypersons would actually use in the real-world context. Test developers should consider including among their rating criteria those factors that strongly influence linguistic laypersons’ judgment. If assessment criteria are developed in such a way as to align with real-world criteria, they are likely to result in more valid inferences regarding the ability to be tested
(Bachman & Palmer, 2010) by more closely reflecting interlocutors’ likely judgments of test-takers in the real world. The obtained score will thus represent how ordinary people would perceive L2 communicative ability of L2 English speakers.

Second, this study also provides useful information on the weights of rating criteria. Analytic assessment criteria used for oral tests are normally composed of multiple criteria, which tend to be weighted equally; this may be because current theories of communication do not specify what weighting should be accorded to the different components of communicative competence. However, since this study not only seeks to reveal the features that linguistic laypersons deem important but also their relative importance in overall judgment, it should be possible to identify how heavily laypersons weight individual features. Accordingly, it is argued that those salient features that powerfully influence their assessment should ideally be weighted more heavily than peripheral ones. This empirically driven determination of weightings potentially strengthens the validity of oral assessment and associated inferences about communicative ability, since the emphasis given to the criteria simulates the importance attached to these criteria in the real world.

1.3 Outline of the Thesis

There are six chapters in the thesis of the current study. Chapter 2 reviews how theories of communication ability have been constructed and evolved since the 1960s. Given some of the limitations of these theories, it provides an argument for the necessity of investigating linguistic laypersons’ perspectives on communication for theory construction. The chapter also reviews empirical studies on raters’ evaluative judgments of L2 speakers’ oral performance and considers the research design best suited to this study. The main research question is addressed at the end of this chapter.
Chapter 3 describes the methodological approach used in the study and its justifications. It also provides an overview of a data collection method—a verbal protocol analysis—and discusses what type of verbal protocol analysis procedure should be employed. The chapter then introduces the two oral proficiency tests used as stimuli to elicit the linguistic laypersons’ judgment of communication ability. Following this, the specific data collection methods and analysis procedures employed are described in detail.

Chapter 4 presents the perspectives on communication ability derived from linguistic laypersons. The results included the test-takers’ behaviors and performance features that affected the research participants’ impressions. The behaviors and features are classified into nine categories and discussed.

Chapter 5 provides a detailed discussion of key findings presented in the previous chapter in order to answer the guiding research question of this investigation. Factors explored are discussed in relation to previous findings and existing theories of communication ability.

Chapter 6 discusses implications for theories of communication ability and language testing.

Chapter 7 concludes the present thesis by evaluating the significance and importance of the study. It presents a summary of the key findings and the significance of the study. It addresses some limitations of the present study and recommendations for further research.
Chapter 2 Literature Review

2.1 Introduction

This chapter reviews the existing literature relevant to the present study and provides a justification for conducting the study. The first section reviews how communication ability has been conceptualized in and outside the field of applied linguistics, and the associated limitations of the related theories. The second section argues for the necessity of understanding linguistic laypersons’ perspectives on communication for theory construction and language assessment. The third section reviews empirical studies on rater judgments of L2 speakers’ oral performance and discusses methodological issues. The chapter concludes by addressing the research question of the study.

2.2 Theories of Communication

Linguists have attempted to elucidate the components of knowledge and ability required for L2 communication and developed theories of L2 communication ability. This section discusses some theories of communication ability, primarily reviewing (a) how theories have been developed and (b) what components of ability have been included and excluded. In addition, the last section reviews some theories of L1 communication proposed outside the field of applied linguistics. Different approaches to communication and limitations of theoretical models are discussed.

2.2.1 Evolution of Theories of L2 Communicative Competence

The evolution of notions of L2 communicative competence started from Chomsky’s (1965) distinction between linguistic competence and linguistic
performance. The former is defined as linguistic knowledge possessed by an ideal speaker-listener, whereas the latter is the actual, practical use of language. According to his theory, competence is narrowly restricted to grammatical knowledge, with performance covering all other factors related to communication, such as pragmatic competence. His strong claim treats competence as a basis underlying the actual language use (Munby, 1978). From the 1960s to the 1980s, a number of novel communicative competence theories rejected Chomsky’s linguistic theory, in particular his strong claim. In fact, theories of L2 communication can be conceived of as logical by-products of researchers’ reactions to his theory (Spolsky, 1989). Basically, linguists have been concerned with factors underlying L2 real-life communication, and thus have infused their theories with their professional perspectives.

Notably, Hymes (1972) pointed out an element missing in Chomsky’s theory and proposed an influential theory of communicative competence from a sociolinguistic perspective. He indicated a limitation of transformational generative grammar:

Such a theory of competence posits ideal objects in abstraction from sociocultural features that might enter into their description. Acquisition of competence is also seen as essentially independent of sociocultural features, requiring only suitable speech in the environment of the child to develop. The theory of performance is the one sector that might have a specific sociocultural content; but while equated with a theory of language use, it is essentially concerned with psychological by-products of the analysis of grammar, not, say, with social interaction. (Hymes, 1972, p. 271)

It is clear that Hymes’s concern is with sociocultural factors in actual language use, which motivated him to construct a broader theory. He contends that “a model of
language must design it with a face toward communicative conduct and social life” (p. 278). One of his major contributions to theory is the inclusion of sociolinguistic knowledge in communicative competence. He added an important question for communication: “whether (and to what degree) something is *appropriate* (adequate, happy, successful) in relation to a context in which it is used and evaluated” (Hymes, 1972, p. 281) [italics in original]; this interpretation is wider than concerns of only whether language is grammatically correct. He continues, “there are rules of use without which the rules of grammar would be useless” (p. 278). More specifically, in the process of language acquisition, children must learn not only the correct forms of language but also socially appropriate and acceptable ways of language use, such as what, when, where, and how to speak. Another important aspect of Hymes’s theory is the notion of *ability for use*. Communicative competence embraces not only the knowledge of language or language use but also ability to put it into practice in a given context. *Ability for use* includes non-cognitive, affective, and volitional factors (e.g., motivation, values, attitudes), which influence the performance of communicative tasks.

Hymes’s argument and definition brought about an enormous impact on a number of researchers in the field of L2 education. For example, Cooper (1968) emphasized Hymes’s concept of sociolinguistic competence by arguing that linguistic knowledge is not the only component of effective communication. Jakobovits (1970) also criticized Chomsky’s narrowly defined linguistic theory in light of L2 language assessment, arguing that discrete-point language testing focusing on assessing linguistic competence does not represent actual language performance. Savignon (1972) concurred with Jakobovits’s argument, stating that “linguistic accuracy in terms of pronunciation, grammar and vocabulary is but one of the major constituents in this complex interaction” (p. 9) [emphasis in original]. In the same way, Widdowson (1978)
underscored the importance of language use taking into account communicative purposes, and argued that learning correct forms of language is not the sole factor in successful communication.

As can be seen thus far, theories of communicative competence have evolved as applied linguists have proposed their views on the ability to communicate in reaction to Chomsky’s linguistic theory. As shown in Figure 2.1, Munby (1978) summarized many proposed communicative competence theories and illustrated their orientations according to researchers’ specialized fields. Some of the researchers’ theories in Figure 2.1 are not discussed in this thesis due to space limitations. Although their impetus for constructing theories of communication was commonly to suggest factors underlying communication, their foci and elements were vastly different.

![Competence constellation (Munby, 1978, p. 21).](image)

In 1980, Canale and Swain published a seminal article on communicative competence, particularly taking into account its implications for L2 pedagogy and
testing. They provided principles for communicative language teaching and proposed a theory of communicative competence building on pre-existing theories. In particular, they adopted Hymes’s (1972) sociolinguistic competence as an essential part of communicative competence. Later, Canale (1983a, 1983b) expanded their original theory, saying “This proposed framework is based on the research reported in Canale and Swain (1980) and other current work in this area” (Canale, 1983a, p. 6). It is thus clear that their theories are extended versions of previous theories that had made contributions to L2 education on a practical level.

Canale and Swain’s (1980) theory consists of three components: grammatical, sociolinguistic, and strategic competences. Although they followed Hymes’s theory, they decided not to include ability for use in communicative competence, deeming it too complex and concluding that its inclusion would not seem to have practical application to L2 education. Later, Canale (1983a, 1983b) expanded their original theory by adding another element, discourse competence, which is related to cohesion in grammatical forms and coherence in meaning. Furthermore, Canale expanded the meaning of strategic competence by adding verbal and non-verbal communication strategies that enhance meaning exchanges. Another difference between the versions of communicative competence outlined in Canale and Swain (1980) and Canale (1983a, 1983b) is that the latter put knowledge and skills under communicative competence underlying actual performance. As such, the theory outlined in the latter is closer to Hymes’s theory, since skills are equivalent to ability for use (McNamara, 1996).

The theoretical models of Canale and Swain (1980) and Canale (1983a, 1983b) impacted many researchers in the field of L2 teaching and assessment. Savignon (1983) provided a theoretical model based on Canale and Swain’s theory for L2 curriculum design and evaluation programs. She illustrated the relationship among Canale’s (1983a,
four components of communicative competence, claiming that “when an increase occurs in one area, that component interacts with other components to produce a corresponding increase in overall communicative competence” (2002, p. 8). However, the relative weight of strategic competence, or a set of coping strategy (e.g., paraphrasing, circumlocution), decreases as communicative competence develops.

Celce-Murcia, Dörnyei, and Thurrell (1995) also proposed a theoretical model of communicative competence by building on the theory developed by Canale and Swain. The purpose for constructing their theory was to provide detailed content specifications of communicative competence for L2 teaching. Their unique contribution was a set of practical and comprehensive lists of contents in each component with detailed explanations. They proposed a communicative competence composed of five competences: linguistic, strategic, sociocultural, actional, and discourse competences, as illustrated in Figure 2.2. They suggested that teachers need to place more emphasis on some components than others according to teaching and learning situations.

**Figure 2.2.** Chronological evolution of theories (Celce-Murcia et al., 1995, p. 11).
Bachman (1990) developed an influential theory called *communicative language ability* (CLA) and subsequently refined it with Palmer (Bachman & Palmer, 1996, 2010). Similar to other theories, the perspectives of CLA were built on the previous theoretical models (e.g., Canale, 1983a, 1983b; Canale & Swain, 1980; Hymes, 1972; Munby, 1978; Savignon, 1972, 1983; Widdowson, 1978). Their primary concern was with explaining factors affecting L2 performance on language tests:

Performance on language tests is affected by a wide variety of factors, and an understanding of these factors and how they affect test scores is fundamental to the development and use of language tests. … The frameworks … reflect my conviction that if we are to develop and use language tests appropriately, for the purpose for which they are intended, we must base them on clear definitions of both the abilities we wish to measure and the means by which we observe and measure these abilities. (Bachman, 1990, p. 81)

Their latest version of CLA includes six types of individual attributes related to language use: (a) language knowledge, (b) topical knowledge, (c) personal attributes, (d) affective schemata, (e) strategic competence, and (f) cognitive strategies. In this model, language knowledge is taken as a part of overall CLA, which influences language use by interacting with other components. Their definition of strategic competence is different from what Canale and Swain (1980) and Canale (1983a, 1983b) considered: “higher-order metacognitive strategies that provide a management function in language use, as well as in other cognitive activities” (Bachman & Palmer, 2010, p. 48). It involves three processes: goal setting, appraising required resources and one’s own knowledge, and planning how to use one’s knowledge to achieve the goal. This view of strategic competence was based on Sternberg’s (1985) model of intelligence.
The notion of Hymes’s (1972) *ability for use* is included in CLA in its original formulation (Bachman, 1990) but not satisfactorily discussed (McNamara, 1996). Strategic competence is considered to be equivalent to *ability for use* since it includes general cognitive factors that enable L2 speakers to use the language knowledge at their disposal. A new component relevant to *ability for use*, called affective schemata, was added in the modified version (Bachman & Palmer, 1996, 2010). Affective schemata refer to the speakers’ emotions associated with particular topics during language use, but their further discussion is restricted to an argument for “bias for best”, a practice that seeks to ensure elicitation of test-takers’ best performance by creating a comfortable environment for them.

Language knowledge in CLA consists of grammatical, textual, functional, and sociolinguistic knowledge, and these are further classified into two types of over-arching knowledge: organizational and pragmatic knowledge (Figure 2.3). This hierarchical relationship was derived from an empirical study (Bachman & Palmer, 1982). They conducted a multi-trait multi-method (MTMM) analysis in order to examine the relationship among three competences: grammatical competence (morphology and syntax), pragmatic competence (vocabulary, cohesion, and organization), and sociolinguistic competence (sensitivity to register, naturalness, and cultural references). These competences were tested by multiple methods including an oral interview, a writing test, multiple-choice questions, and student self-assessment. A confirmatory factor analysis extracted (a) a grammatical and pragmatic competence factor and (b) a sociolinguistic competence factor. This result suggests that grammatical knowledge and textual knowledge are clustered together and dissociated from sociolinguistic knowledge. Accordingly, CLA is a theory of communicative competence incorporating empirical data of test-takers’ performances, although the knowledge Bachman and
Palmer examined was adopted from previous theories, and the data were collected from performance on test items based on those theories.

![Components of language knowledge in CLA](image)

*Figure 2.3. Components of language knowledge in CLA (Bachman, 1990, p. 87).*

Bachman (1990) and Bachman and Palmer’s (1996, 2010) concept of CLA impacted some researchers and motivated them to develop extended theories of L2 communication. For example, concerned with the assessment of language for specific purposes (LSP), Douglas (2000) developed specific-purpose language ability, extending CLA and taking into account background knowledge. Purpura (2004) also provided accounts for the role of grammar using the framework of CLA. In conclusion, theories of communicative competence since the 1980s have evolved as applied linguists elaborate on earlier models suggested by Canale and Swain (1980) for L2 language teaching and testing.

### 2.2.2 Interactional Competence

Even though researchers have conceptualized communicative competence that L2 speakers need to possess in order to effectively communicate, their theories have not satisfactorily accounted for how L2 speakers actually interact with interlocutors in various social contexts. In other words, communicative competence is considered as a cognitive trait that is generalizable to different contexts (Chalhoub-Deville, 2003;
Chalhoub-Deville & Deville, 2005; Hall & Doehler, 2011; Luoma, 2004; McNamara, 1996, 2003; Purpura, 2008). However, real-life communication involves a range of social factors (e.g., settings, interlocutors) that affect language performance. McNamara (1997) argues that:

A danger of too exclusive a focus on defining the nature of candidate ability in cognitive terms is that the performance is seen as in some way a simple projection of the candidate’s ability. It is as if the candidate is exclusively responsible for the performance, and can be held accountable accordingly. But clearly a performance is not a simple projection of what is in the head of the candidate, even if that display is mediated by the candidate’s strategies for dealing with the interactional context in which it is to be achieved. (p. 453)

The central focus on individual’s cognitive trait was pointed out by Kramsch (1986), who proposed *interactional competence*, a concept that was further elaborated on by other researchers (He & Young, 1998; Jacoby & Ochs, 1995; Johnson, 2001; Young, 2000, 2002, 2008). Kramsch (1986) argued against an oversimplified perspective on interaction, stating that:

Whether it is a face-to-face interaction between two or several speakers, or the interaction between a reader and a written text, successful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or “sphere of inter-subjectivity” that is built through the collaborative efforts of the interactional partners. (p. 367)

The notion of *co-construction* underlies interactional competence. It is defined
as “the joint creation of a form, interpretation, stance, action, activity, identity, institution, skill, ideology, emotion, or other culturally meaningful reality” (Jacoby & Ochs, 1995, p. 171). They continue that “the co- prefix in co-construction is intended to cover a range of interactional processes, including collaboration, cooperation, and coordination” (p. 171) [italics in original]. Interactive practices are created collaboratively with interlocutors, and each participant has a responsibility for constructing interaction; a single person cannot be blamed for communication breakdown. Accordingly, Hall and Doehler (2011) state that “IC [interactional competence], that is the context-specific constellations of expectations and dispositions about our social worlds that we draw on to navigate our way through our interactions with others, implies the ability to mutually coordinate our actions” (pp. 1-2).

Interactional competence theorists posited that there is no universal competence generalizable across different contexts. Instead, following the view of Vygotsky’s sociocultural theory, they emphasized local competence applicable to specific contexts (Johnson, 2001). This is one of the most fundamental differences from communicative competence and CLA, which assume universal competence and aim to generalize ability across different language use settings. Therefore, interactional competence is considered as a behavioral construct rather than trait (Young, 2002), meaning that performance is determined by the summation of local contextual factors rather than the speaker’s traits (Chapelle, 1998).

It is argued that participants in interactional practices bring interactional resources and use them according to contexts. Young (2008) contends that participants, like architects, are required to apply resources at their disposal to construct interactional practices. He and Young (1998) and Young (2000, 2008) provided a list of resources including (a) identity resources, (b) linguistic resources, and (c) interactional resources,
as summarized in Table 2.1. All participants must employ these resources mutually and reciprocally in a given context. Resources employed by them vary according to the context.

### Table 2.1

*Interactional Resources (Young, 2008, p. 71)*

<table>
<thead>
<tr>
<th>Resource</th>
<th>Subcategory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td><em>Participation framework</em>: the identities of all participants in an interaction, present or not, official or unofficial, ratified or unratified, and their footing or identities in the interaction</td>
</tr>
<tr>
<td>Linguistic</td>
<td><em>Register</em>: the features of pronunciation, vocabulary, and grammar that typify a practice</td>
</tr>
<tr>
<td></td>
<td><em>Modes of meaning</em>: the ways in which participants construct interpersonal, experiential, and textual meanings in a practice</td>
</tr>
<tr>
<td>Interactional</td>
<td><em>Speech acts</em>: the selection of acts in a practice and their sequential organization</td>
</tr>
<tr>
<td></td>
<td><em>Turn-taking</em>: how participants select the next speaker and how participants know when to end one turn and when to begin the next</td>
</tr>
<tr>
<td></td>
<td><em>Repair</em>: the ways in which participants respond to interactional trouble in a given practice</td>
</tr>
<tr>
<td></td>
<td><em>Boundaries</em>: the opening and closing acts of a practice that serve to distinguish a given practice from adjacent talk</td>
</tr>
</tbody>
</table>

Specific aspects of interactional competence have been conceptualized by empirical studies conducting Conversation Analysis (CA) of L2 speaker interactions. An example is a study conducted by Galaczi (2013), who provided insights into interactional competence by examining the discourse of paired oral performances. She conducted CA to examine the discourses demonstrated by test-takers at different language proficiency levels and explored the interactional behaviors demonstrated during the interaction. It was found that advanced learners exhibited various types of interactional behaviors, such as the extensions of partner-initiated topics, turn-taking management similar to natural talk (e.g., overlaps, latches), listener support,
backchannels, and confirmation of comprehension (e.g., “absolutely”, “exactly”). In contrast, low-proficiency test-takers failed to exhibit these behaviors in their interactions. Their interaction was characterized by low mutuality and a number of short-lived topics without elaboration. Galaczi’s (2013) study contributed to the conceptualization of interactional competence that is required for interactional practices. She concluded that the development of interactional competence goes hand in hand with the development of language proficiency since higher automaticity of cognitive processes allows participants to engage in the interaction more efficiently.

Other studies have also contributed to the conceptualization of interactional competence by applying the CA approach or investigating the listener’s perception of interactional behaviors exhibited by L2 speakers (to be discussed in Section 2.4.2.2). Table 2.2 summarizes some specific aspects of interactional competence explored by empirical studies.

Table 2.2
Aspects of Interactional Competence

<table>
<thead>
<tr>
<th>Behaviors/Features</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Asking questions / responding to partner</td>
<td>Brown (2005); Lazaraton (2002); May (2011),</td>
</tr>
<tr>
<td></td>
<td>May (2011)</td>
</tr>
<tr>
<td>● Cooperation (e.g., help partner)</td>
<td>May (2011)</td>
</tr>
<tr>
<td>● Interactional pattern (e.g., collaborative)</td>
<td>Galaczi (2008, 2013); May (2011); Nakatsuhara</td>
</tr>
<tr>
<td></td>
<td>(2011)</td>
</tr>
<tr>
<td>● Interactive listening (e.g., backchannel)</td>
<td>Ducasse (2010); Galaczi (2013)</td>
</tr>
<tr>
<td>● Non-verbal interpersonal communication</td>
<td>Ducasse (2010); May (2011)</td>
</tr>
<tr>
<td>● Topic development (engagement)</td>
<td>Brown (2005); Ducasse (2010); Galaczi (2013);</td>
</tr>
<tr>
<td></td>
<td>Gan (2010)</td>
</tr>
<tr>
<td>● Turn-taking management</td>
<td>Ducasse (2010); Galaczi (2013); Nakatsuhara</td>
</tr>
<tr>
<td></td>
<td>(2011)</td>
</tr>
<tr>
<td>● Understanding partner</td>
<td>May (2011)</td>
</tr>
</tbody>
</table>
2.2.3 Empirically Based L2 Language Proficiency

As discussed above, researchers have constructed theories of L2 communication by extending previous theories and infusing them with their own professional beliefs. However, some theoretical models have actually been developed based on empirical studies. This section reviews (a) the Common European Framework of Reference (CEFR) and (b) empirically based L2 speaking proficiency developed by Hulstijn, Schoonen, Van Gelderen, De Jong, and their colleagues.

The most widely used empirically based L2 language proficiency model is the CEFR, which “describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively” (Council of Europe, 2001, p. 1). Thus, the CEFR is a model describing language use and defines language proficiency for communication. Currently, it widely serves as an influential standard for syllabus design or assessment at all educational levels, even beyond Europe (McNamara & Roever, 2006). The CEFR adopts an action-oriented approach that views users and learners of a language primarily as ‘social agents’, i.e. members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action. (Council of Europe, 2001, p. 9)

Following this approach, the developers made behavioral statements describing what learners at each proficiency level (six levels from A1 to C2) can do using language (i.e., can-do statements). Although the Council of Europe (2001) attempted to describe language use taking into account contexts of language use and language competence, it
has been claimed that the ability-related statements (i.e., what speakers can or cannot do) fail to refer to contextual factors and competences underlying the actions learners are required to do (Fulcher, 2010; Fulcher & Davidson, 2007; Weir, 2005). It is the users of the framework that consider the language use contexts and competences learners should be equipped with to execute the behavioral statements.

The CEFR was developed based on existing rating scales of language proficiency and language teachers’ perceptions of language proficiency (Council of Europe, 2001; Fulcher, 2010; North, 2000). First, 2,000 descriptors were collected from existing rating scales, and provisional categories and levels were identified from the pool of descriptors. Second, language teachers were asked to categorize the descriptors and divide them into proficiency levels in order to exclude unnecessary descriptors from the pool. Third, questionnaires were developed to collect another group of teachers’ perceptions of difficulties of the descriptors. Finally, the perceived difficulties were calibrated using Rasch analysis and divided into six proficiency levels. Thus, the components of the CEFR originated from existing rating scales, which suggests that the CEFR “is a reflection of the consensus over achievement in foreign language learning at different levels of competence” (North, 2000, p. 181). However, the CEFR is a reflection of teachers’ perceptions of pre-existing descriptors, rather than teachers’ evaluative judgments of L2 performance. Some researchers have thus claimed that the CEFR does not necessarily reflect actual performance or second language acquisition theory (Fulcher, 2010; Hulstijn, 2007).

The CEFR includes six qualitative aspects of spoken language use: (a) range, (b) accuracy, (c) fluency, (d) interaction, and (e) coherence. It also describes each aspect’s quality of performance that learners at different proficiency levels (from A1 to C2) can demonstrate.
In an approach quite different from the development of the CEFR, a group of researchers including Hulstijn, Schoonen, Van Gelderen, De Jong, and their colleagues has investigated the componential structure of L2 ability by analyzing actual performance data from L2 and L1 speakers. Their main concern is expressed in the following questions:

Which capacities should be included in a construct of language ability, and which should not and should language ability be considered as one monolithic ability or not? And if not, how should we value the contributions of constituent component abilities? (Schoonen, 2011, p. 702)

The relative contributions of multiple components to speaking performance have been investigated in the project called “What Is Speaking Proficiency? (WiSP)” (Hulstijn, Schoonen, & Van Gelderen, 2007). Assuming that speaking proficiency is multicomponential, De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012) investigated what skills (linguistic knowledge and processing skills) underlie oral proficiency. They administered a Dutch speaking test to measure Functional adequacy of test-takers’ monologic performance on computer-delivered role-play tasks. Functional adequacy includes features such as “both the amount and detail of information conveyed, relevant to the topic, setting (formal or informal), and discourse type (descriptive or argumentative) and the ease with which the description could be followed” (p. 15). Additionally, L2 Dutch students undertook eight tests that measured (a) linguistic knowledge, (b) processing skills, and (c) pronunciation (see Table 2.3).

1 Similar studies were conducted on writing proficiency (Schoonen et al., 2003) and reading proficiency (Schoonen, Hulstijn, & Bossers, 1998; Van Gelderen et al., 2004; Van Gelderen et al., 2003). This section only reviews studies of speaking proficiency.
Structural equation modeling (SEM) was applied to reveal the componential structure of the factors. The result showed that the measured components explained 75.7% of variances of *Functional adequacy*. It was significantly correlated with all the measures except for speed of articulation and significantly predicted by vocabulary and intonation. These findings indicate that linguistic knowledge and processing skills as well as pronunciation can be considered as components of oral proficiency. Although processing skills are not explicitly explained in theories of communicative competence (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980), this research has provided evidence that both linguistic knowledge and processing skills are included in language proficiency.

<table>
<thead>
<tr>
<th>Knowledge and Skill</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic knowledge</td>
<td>- Knowledge of vocabulary</td>
</tr>
<tr>
<td></td>
<td>- Knowledge of grammar</td>
</tr>
<tr>
<td>Processing skill</td>
<td>- Speed of lexical retrieval</td>
</tr>
<tr>
<td></td>
<td>- Speed of articulation: response latency</td>
</tr>
<tr>
<td></td>
<td>- Speed of articulation: response duration</td>
</tr>
<tr>
<td></td>
<td>- Speed of sentence building</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>- Speech sound</td>
</tr>
<tr>
<td></td>
<td>- Word stress</td>
</tr>
<tr>
<td></td>
<td>- Intonation</td>
</tr>
</tbody>
</table>

Based on the studies summarized in this section and several empirical studies using MTMM analysis (Bachman & Palmer, 1982; Harley, Cummins, Swain, & Allen, 1990), Hulstijn (2011) proposed that language proficiency is composed of core and periphery, as shown in Figure 2.4 (see also Hulstijn, 2007). The core consists of linguistic knowledge and fluency, whereas the periphery includes non-linguistic
components explained by some theories, such as strategic competences (e.g., Canale & Swain, 1980) and metalinguistic knowledge (e.g., Bachman, 1990). This categorization was proposed since the knowledge in the periphery cannot exist without linguistic knowledge. In addition, grammatical knowledge turned out to be the primary component of general language proficiency factor in the studies of Bachman and Palmer (1982) and Harley et al. (1990). Hulstijn (2011) not only included less purely linguistic competences in language proficiency, but also put forward the hierarchical relationship between linguistic and non-linguistic knowledge. However, as Figure 2.4 implies, the list of knowledge in the periphery is not exhaustive and needs to be explored further.

<table>
<thead>
<tr>
<th>Language Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CORE</strong></td>
</tr>
<tr>
<td>Linguistic cognition (knowledge and speed) in the phonetic-phonological, morphonological, morphosyntactic, and lexical domains:</td>
</tr>
<tr>
<td>* Basic language cognition</td>
</tr>
<tr>
<td>* Higher language cognition</td>
</tr>
<tr>
<td><strong>PERIPHERY</strong></td>
</tr>
<tr>
<td>Metacognitive competences</td>
</tr>
<tr>
<td>* Metalinguistic knowledge</td>
</tr>
<tr>
<td>* Knowledge of various types of oral and written discourse with their own characteristics</td>
</tr>
<tr>
<td>* Strategic competences</td>
</tr>
<tr>
<td>* …</td>
</tr>
</tbody>
</table>

*Figure 2.4. Core and periphery components of language proficiency (Hulstijn, 2011, p. 239).*

### 2.2.4 L1 Communication Competence

Since interpersonal communication is vital to many aspects of human life—including marriage, health, academic achievement, workplace success, and psychological well-being—communication competence and interpersonal communication skills are also widely studied outside the field of applied linguistics (Spitzberg, 2003; Spitzberg & Cupach, 2002). In the field of communication studies,
most researchers concerned with L1 communication skills or interpersonal skills have investigated the competence to communicate. Wilson and Sabee (2003) claim that there are three reasons underlying the large number of studies on communication competence:

(a) within any situation, not all things that can be said and done are equally competent; (b) success in personal and professional relationships, in no small part, on communicative competence; and (c) most people display incompetence in at least a few situations, and a smaller number are judged incompetent across many situations. (pp. 3-4)

Nevertheless, they acknowledged the difficulty in defining what communication competence consists of. Researchers in this field thus have attempted to develop various theoretical models of communication, which appear to be quite different from those developed for L2 teaching or testing reviewed in the previous sections.

Researchers have developed a wide range of competence related to communication, including fundamental competence, social competence, interpersonal competence, linguistic competence, and communicative competence. Spitzberg and Cupach (1984) conceptualized these competences by categorizing them into two approaches: (a) the outcome-focused approach and (b) the message-focused approach. The former includes fundamental, social, and interpersonal competences, which underscore the effectiveness of communication (e.g., outcomes of interaction, successful goal achievement) as the core criterion. The product of interaction is the primary concern in this approach. Accordingly, the outcome-focused competences are generally defined as the ability for speakers to adapt in order to achieve goals (fundamental competence), to facilitate competent interactions (social competence), and
to accomplish tasks through interactions (interpersonal competence). Although message conveyance is one of the communication outcomes to be evaluated, it does not guarantee successful goal achievement, as argued by Rickheit, Strohner, and Vorwerg (2010):

Even though the addressee might understand what was intended by an utterance, he or she doesn’t necessarily act according to this intention. A request or an offer may be rejected; the child may be leaving the house without taking the jacket suggested by a caretaker, etc. (p. 28)

The effectiveness of outcome depends on whether (a) the intended message is conveyed and (b) the communicative goal is achieved as a result of successful message conveyance.

Linguistic and communicative competences belong to the message-focused approach, whose main concern is with issues of language and message behaviors. The core criterion of these competences is the appropriateness of messages, which is related to the extent to which performance is legitimate in a given context. Spitzberg and Cupach (1984) state that “linguistic and communicative competence constructs … generally focus on the appropriateness of messages, either grammatically or contextually, with a corresponding deemphasis of the functional outcomes of the communication observed” (p. 67). These competences are characterized as linguistic knowledge underlying the use of language and the ability to convey messages appropriately in a given context, respectively (Spitzberg & Cupach, 1984). Thus, the competence broadly parallels Chomsky’s (1965) linguistic theory and Hymes’s (1972) communicative competence. Table 2.4 summarizes the core criteria and competences of
both approaches to communication.

Table 2.4

<table>
<thead>
<tr>
<th>Core Criteria and Competences of Outcome- and Message-focused Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core criterion</td>
</tr>
<tr>
<td>Competences</td>
</tr>
<tr>
<td>- Fundamental competence</td>
</tr>
<tr>
<td>- Social competence</td>
</tr>
<tr>
<td>- Interpersonal competence</td>
</tr>
</tbody>
</table>

In general, both effectiveness and appropriateness are considered as the primary criteria of overall communication competence (Morreale, Spitzberg, & Barge, 2013; Rickheit et al., 2010; Spitzberg, 2000, 2003; Spitzberg & Cupach, 2002). Morreale et al. (2013) claim that “Competent communication is both appropriate and effective” (p. 5). Accordingly, communication competence consists of components or skills that enable a person to achieve his or her desired outcomes in a socially acceptable manner.

Rickheit et al. (2010) claim that the contributions of Wiemann, Spitzberg, and Cupach should be regarded as important milestones in the development of communicative competence, in addition to those of Chomsky and Hymes. The former researchers’ particular interests are in how communication forms interpersonal relationships and factors underlying social interactions. Wiemann (1977) proposed a model of communicative competence composed of five components: affiliation/support, social relaxation, empathy, behavioral flexibility, and interactional management. Their associated specific skills and behaviors are given in Table 2.5. His theoretical model is

---

2 Spitzberg (2003) lists seven criteria of competence (dialogical criteria, clarity, understanding, efficiency, satisfaction, effectiveness, and appropriateness). Among them, effectiveness and appropriateness are the criteria most commonly used.
built on previous theories including Argyle’s (1969) social skill approach, which emphasizes empathic, supportive, and relaxed interaction as competent communication (called other orientation). He defined communicative competence as

the ability of an interactant to choose among available communicative behaviors in order that he may successfully accomplish his own interpersonal goals during an encounter while maintaining the face and line of his fellow interactants within the constraints of the situation. (p. 198)

In other words, communicative competence is realized as a person’s ability to apply communicative skills to achieve his or her interpersonal goals.

Table 2.5

<table>
<thead>
<tr>
<th>Component</th>
<th>Skill/Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affiliation/support</td>
<td>eye behavior, marking relationships, head nods, speech rate, smiling, owning of one’s behavior, proximity</td>
</tr>
<tr>
<td>Social relaxation</td>
<td>relaxing cues, rate of speech, speech disturbance, object manipulations</td>
</tr>
<tr>
<td>Empathy</td>
<td>reciprocity of affect displays, verbal immediacy, perceived active listening</td>
</tr>
<tr>
<td>Behavioral flexibility</td>
<td>verbal immediacy cues, speech choices</td>
</tr>
<tr>
<td>Interaction management</td>
<td>speaker interruptions, turn-taking, pauses, attention to the encounter</td>
</tr>
</tbody>
</table>

Wiemann (1977) collected empirical data on the interrelationship among the components of communicative competence and concluded that interaction management is a central factor of communicative competence judged by research participants (students). In addition, other components of his theoretical model are also interdependent on communicative competence. Wiemann underscored the nature of
dyadic interaction based on the finding that a speaker was not regarded as competent unless he or she gave attention to the interaction and contributed to smooth conversation. This concept appears to be equivalent to the co-construction emphasized in interactional competence (Jacoby & Ochs, 1995), although Wiemann’s theory provides more specific behaviors and skills related to competent interaction.

Spitzberg and his colleagues (Spitzberg & Cupach, 1984, 1989; Spitzberg & Hurt, 1987) conceptualized communicative competence by identifying observable behaviors related to people’s judgments of a competent communicator. Spitzberg and Hurt (1987) thus asked “what behaviors are most likely to be viewed as competent” (p. 30) [italics in original]. They proposed four skill-clusters: interactional management, altercentrism, expressiveness, and composure, as listed in Table 2.6. Their main purpose was to establish a valid measurement of interpersonal competence. Specific behaviors related to the competence were derived from their participants’ open-ended descriptions of competent/incompetent interactions and chosen from previous theories. These behaviors were reduced and clustered into four components. Based on empirical studies, 25 items were finally included in the Conversational Skills Rating Scale for measuring interpersonal communication competence (Spitzberg, 1995; Spitzberg & Hurt, 1987).
Table 2.6

*Components of Spitzberg and Hurt’s (1987) Interpersonal Competence*

<table>
<thead>
<tr>
<th>Component</th>
<th>Skill/Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction management</td>
<td>topic initiation, question, topic follow-up, encouragements, interruptions, talk time, fluency, vocalized pauses, restatements, response time</td>
</tr>
<tr>
<td>Altercentrism</td>
<td>supportiveness, seek clarification, self- and other-references, approval indication, head nods, body lean, smiling, eye contact, body orientation</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>personal opinion expression, appropriate use of humor, monotone voice, vocal variety, brevity of speaking turns, gestures, appropriate facial expressions, smiling, laughing, body posture and orientation, illustrators</td>
</tr>
<tr>
<td>Composure (Relaxation / Confidence)</td>
<td>response shortness, speech rate, speech blockages, vocal tension, response relevance, object manipulation, eye contact, extraneous movement, fidgeting, postural rigidity</td>
</tr>
</tbody>
</table>

More recently, Morreale et al. (2013) defined communication competence as “the extent to which speakers achieve desired outcomes through communication behavior acceptable to a situation” (p. 25). They claim that people are more likely to be perceived as a competent communicator if they have (a) motivation to communicate, (b) content knowledge (related to topics, words, meanings, etc.) and procedural knowledge (related to how to perform), and (c) skills to apply their motivation and knowledge. People need to apply these three components effectively and appropriately in actual communicative contexts.

A common feature of these competence theories (Morreale et al., 2013; Spitzberg & Cupach, 1984, 1989; Spitzberg & Hurt, 1987; Wiemann, 1977) is an emphasis on successful communication or goal achievement for all interaction participants. Linguistic features were unlikely to be assumed, because these theories were concerned with L1 communicative or interpersonal competence and thus high language proficiency was presumed. Another commonality among the communicative competence theories above is inclusion of non-linguistic features or skills. In particular,
the importance of non-verbal behavior in communication has been emphasized. When
verbal and non-verbal cues conflict, the latter are more likely to be trusted (Puccinelli,
2010).

2.2.5 Limitations of Theories of L2 Communication Ability

As discussed above, the process of L2 theory construction has been based on
(a) reaction to or elaboration of previous theories and/or (b) empirical data of
performance on language tests. It was a theoretical debate, not empirically grounded
except through test data from items, which were based on various theoretical positions
within this debate.

A crucial limitation of theories of L2 communication ability is that their
development processes have not included people’s subjective judgments of
communicative performance. Although some theories have been developed using
empirical data of performance on language tests and investigating how various factors
affect communicative performance (e.g., De Jong et al., 2012), the factors they
examined (e.g., Table 2.3) were not empirically derived but from existing theories of
communication ability. Thus, they have not taken into account how a person in the real
world actually judges communicative performance; that is, the factors to be included in
the theory have not been derived from a person’s actual evaluative judgment.
Accordingly, theories do not necessarily explain what features or behaviors of speakers
are likely to be perceived as constituting competence in communication by general
people. This could potentially undermine the meaningfulness of the theories.

Arguably, components of communication ability should be determined based
on people’s subjective judgments. Since communication is an interpersonal act, how
people actually judge or perceive others’ communicative performance is crucial; in the
real world, other communicators subjectively judge people’s communication ability. Morreale et al. (2013) assert that “how we actually behave in most instances is less important than how others perceive us to have behaved” (p. 25) [emphasis in original]. In fact, some have stated that competence in communication is determined by a person’s subjective judgment of communicative performance (McFall, 1982; Spitzberg, 2000, 2003; Spitzberg & Cupach, 1984, 2002). Thus, the components of communication ability should be composed of the features or behaviors that lead to people’s positive judgment of communicative performance. With regard to this, Spitzberg (2000) contends that “All the available candidates for competent communication criteria [e.g., effectiveness, appropriateness] are deeply infected with subjectivity. If the criteria of competence are subjective, then any skill claimed to comprise competence is itself subject to subjective inclusion, exclusion, or redefinition” (p. 109).

Another limitation of existing theories of L2 communication ability is a heavy emphasis on language proficiency. Since applied linguists have developed and elaborated on theories of L2 communication, it is not surprising that the theories meticulously describe linguistic knowledge or interactional resources. Unlike some L1 communication theories (e.g., interpersonal competence), theories of L2 communication ability tend to underscore language use rather than communication outcome. In other words, applied linguists tend to explicate language knowledge with detailed definitions and avoid specifying other components of communication (Luoma, 2004). In particular, Hymes’s (1972) ability for use, which encompasses various non-linguistic factors affecting communicative performance, has rarely been attended to by applied linguists. McNamara (1996) states that:

*Ability for use, on the other hand, is more difficult to grasp, because we need to consider*
here a range of underlying language-relevant but not language-exclusive cognitive and affective factors (including general reasoning powers, emotional states and personality factors) which are involved in performance of communicative tasks. Because these factors are not exclusive to the domain of language use they are not the preserve of language specialists; that [sic] have therefore been less often discussed in the language field and, consequently, their role in communication is less clearly understood. (p. 59) [italics in original]

Conversely, interpersonal and communicative competence theories in the field of communication studies (Morreale et al., 2013; Spitzberg & Hurt, 1987; Wiemann, 1977) have stressed the importance of non-linguistic features or behaviors for achieving the goal of interaction participants. Most features explained in the theories have been excluded from L2 communication theories. Although the role of non-language-specific factors in communication has been acknowledged by many applied linguists (e.g., Bachman, 1990; Hulstijn, 2011; Hymes, 1972; Young, 2008), little is known about what features contribute to successful L2 communication vis-à-vis linguistic knowledge. Moreover, as reviewed above, the role of non-verbal behavior in communication has not been emphasized in L2 communicative theories. Detailed explanation of this aspect (e.g., specific non-verbal behavior) rarely appears in the existing L2 literature, with a few exceptions (Celce-Murcia et al., 1995; Savignon, 1983).

2.3 The Need to Understand the Linguistic Laypersons’ Perspective

As discussed above, the development of L2 communicative theories have mostly involved language specialists’ views on L2 communication. However, L2 speakers’ oral performance is normally judged by language non-specialists (linguistic
laypersons) in real-world domains. Accordingly, understanding their unique evaluative judgments is considered important for building a theory of L2 communication ability. This section focuses on linguistic laypersons’ views on L2 oral communication and reviews empirical studies on various non-linguistic domain experts’ indigenous assessment criteria. This section also discusses who should be considered “expert judges” in general communication in the current globalized world.

### 2.3.1 Non-linguists’ Views on L2 Communication

In the field of linguistics, linguistic laypersons’ views on language have been dismissed as “uninteresting, unqualified, uninformed or even dangerous” (Wilton & Stegu, 2011, p. 1). Nevertheless, it has been claimed that the non-linguists’ views on language and communication are worth investigation. In the book *Folk Linguistics*, Niedzielski and Preston (2000) state that “We have sought out and even encouraged stankos [the language beliefs of non-linguists], for we believe that what the folk believe about language deserves careful consideration” (p. vii). They defined *folk* as “those who are not trained professionals in the area under investigation” and did not mean “rustic, ignorant, uneducated, backward, primitive, minority, isolated, marginalized, or lower status groups or individuals” (p. viii). They argue that non-linguists’ views on language should be focused on in linguistics for three reasons: (a) their views represent their culture of language use; (b) it is hard for linguists to influence laypeople without recognizing their views; and (c) their views may influence the shape of language per se.

Wilton and Stegu (2011) argue that applied linguists should pay close attention to non-linguists’ views on language and communication. They state that:

> Applied linguistics is a science predominantly conducted for the benefit of laypeople.
Therefore, we consider the detailed investigation of lay views of language(s) and communication as very important, if not essential, for applied linguists. On the one hand, linguists can learn a lot from laypeople for their own understanding of linguistic issues; on the other hand, in order to act as experts and advisors on linguistic issues, linguists need to know what non-linguists think of the issues at hand, how their knowledge is generated, and how it can be enriched in order to help them solve language-related problems or change negative attitudes. (p. 12)

Accordingly, a fundamental characteristic of folk linguistics is that non-linguists’ views on language and communication must not be considered incorrect but rather treated “with respect” (Pasquale, 2011, p. 97). For example, Niedzielski and Preston (2000) collected non-linguists’ beliefs about second language acquisition and found that their beliefs were not necessarily far from professional opinions. Pasquale (2011) also investigated non-linguists’ beliefs about L2 teaching and learning by using discourse approaches and concluded that course designers can develop better materials and syllabi suitable to students by leveraging knowledge about their beliefs and attitudes. He argues that “folk linguistics can provide applied linguistics with the kind of information that aids them in their work by giving them access to a wide range of data concerning the beliefs and attitudes of students and teachers” (p. 97).

Although it has been acknowledged that linguistic laypersons’ views might make a contribution to the field of applied linguistics, the actual process of theory construction has seldom included their views, as has been discussed above. Likewise, in the field of language testing, language specialists (language testers or teachers) primarily define the construct of general-purpose oral communication ability. They often rely on their intuition, referring to theoretical models of L2 communication ability
in determining the criteria to be included in construct definitions (Bachman & Palmer, 2010). In this sense, communication ability, as conceived by general oral proficiency tests, reflects language specialists’ perspectives. In contrast, the perspectives on the L2 communication ability of linguistic laypersons have seldom been incorporated into language assessment. Thus, the ideal performance defined by the test may not be identical to the ideal performance perceived by linguistic laypersons. Since linguistic laypersons are in most cases the ultimate arbiters of L2 speakers’ oral performance in real-world contexts, it is necessary for language tests to investigate their perspectives on L2 communication ability as a basis for more accurately predicting laypersons’ impressions in real life.

In fact, some researchers have already claimed that investigation of linguistic laypersons’ evaluative judgments is important for general-purpose oral performance testing, as summarized in Table 2.7. Two justifications are often provided. First, linguistic laypersons are considered to have different views on communication from language teachers, which will be discussed in the following section in detail (Chalhoub-Deville, 1995, 1996; Pollitt & Murray, 1996). Second, linguistic laypersons are the ultimate arbiters of L2 speakers’ oral performance (Barnwell, 1986; Brindley, 1991; Chalhoub-Deville, 1996). L2 speakers are more likely to communicate with non-linguists than with language specialists in real-life settings. In this sense, linguistic laypersons are the “eventual receivers” of L2 speakers’ oral communication (Hadden, 1991, p. 4) and should be regarded as an “expert audience” (Brindley, 1991, p. 149) once speakers begin engaging in communication outside classrooms and testing centers. Furthermore, how they judge L2 communication ability should be regarded an important source for the assessment criteria used in speaking tests since test developers need to ensure that “the criteria and procedures for recording the responses to the
assessment tasks correspond closely to those that are typically used by language users in assessing performance in TLU tasks” (Bachman & Palmer, 2010, p. 236).

Table 2.7

<table>
<thead>
<tr>
<th>Researcher(s)</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnwell (1986)</td>
<td>Surely we cannot continue to claim to gauge proficiency without seeking input from the “naïve” native speakers with whom we hope our students will one day interact. ... Other people should be the judges of how well we speak, not other language teachers or testers. (p. 42) [emphasis in original]</td>
</tr>
<tr>
<td>Brindley (1991)</td>
<td>So-called “naïve” native speakers constitute another “expert” audience whose perceptions could profitably be drawn on in establishing performance criteria. (p. 149)</td>
</tr>
<tr>
<td>Hadden (1991)</td>
<td>Researchers have begun to study the perceptions of nonnative communication of the so-called “linguistically naïve” native speaker, who lacks a linguistic background and extensive exposure to student interlanguage. (p. 4)</td>
</tr>
<tr>
<td>Chalhoub-Deville</td>
<td>In conclusion, because the NS teachers, who usually evaluate learners’ L2 oral proficiency, are acting as surrogates for the nonteaching NSs, it is necessary to validate these teachers’ criteria with those of nonteaching NSs. (p. 258)</td>
</tr>
<tr>
<td>Chalhoub-Deville</td>
<td>Teacher training, however, may influence teachers’ assessment and render their judgement different from non-teaching native raters … with whom many L2 learners wish to communicate. (p. 56)</td>
</tr>
<tr>
<td>Pollitt &amp; Murray</td>
<td>The next phase of research should perhaps be to repeat this study using naïve judges, that is judges who are not trained in linguistics, to see if the order of proficiency they produce is similar to that produced by teachers. (p. 89)</td>
</tr>
</tbody>
</table>

Although some studies have touched on an aspect of non-linguists’ judgments of L2 communication for general purposes (Barnwell, 1989; De Jong et al., 2012; Gullberg, 1998; Hadden, 1991), they have not necessarily elicited the criteria used for assessing communicative performance or information on how such judges make evaluative judgments of L2 communication ability. An exception is Galloway (1980), who elicited both language teachers’ and non-teachers’ ratings of L2 Spanish oral
performances and analyzed their written comments on the ratings. Although this research was concerned primarily with the difference in severity between the two groups of raters, it also explored the features heeded by them while judging the performances. The results showed that non-teaching native speakers (linguistic laypersons) commented mostly on the content of speech, whereas non-native teachers focused solely on grammatical accuracy (see also Section 2.3.2). The assessment criteria underlying linguistic laypersons’ evaluative judgments of L2 oral performances are under researched.

2.3.2 Language Specialists and Laypersons’ Perspectives on Communication

Empirical studies have found that linguistic laypersons hold unique views on oral communication ability and judge oral performances differently from language specialists. This difference is one of the present study’s primary justifications for investigating the perspectives of linguistic laypersons.

Some studies demonstrated that language specialists are more critical of L2 speakers’ linguistic features. Hadden (1991) compared judgments made by ESL teachers and non-teachers. Language teachers and linguistic laypersons watched video recordings of Chinese university students’ oral performances and judged 24 aspects of the performances, such as comprehensibility, grammatical accuracy, and so forth. A factor analysis of their judgments revealed that the factor structures derived from the two groups of raters were similar. However, the ESL teachers were significantly stricter on language ability (including vocabulary, grammar, and fluency) than the non-teachers. There was no statistically significant difference in their judgments of comprehensibility, social acceptability, personality, or body language. Hadden’s (1991) study thus suggested that language teachers may be more sensitive to linguistic features than
non-teachers. However, the results did not necessarily indicate that language teachers attended to linguistic ability more strongly than to other features of performances.

The notion that language teachers harshly judge linguistic features was supported by Brown (1995), who investigated language teachers’ and professional tour guides’ assessments of performances on the Japanese Test for Tour Guides. These raters were asked to assess oral performances on simulated tasks using analytic rating criteria composed of (a) *Resources of grammar and expression*, (b) *Fluency*, (c) *Pronunciation*, (d) *Vocabulary*, (e) *Use of polite forms*, and (f) *Comprehension*. Overall harshness was not significantly different between the two groups of raters. Nevertheless, it was found that the language teachers judged *Resources of grammar and expression*, *Fluency*, and *Vocabulary* more severely than the tour guides did. This finding is in accord with that of Hadden (1991). In contrast, the language teachers were more lenient on the criterion of *Pronunciation*. Brown (1995) argues that the language teachers were accustomed to learners’ pronunciation after being exposed to their Japanese for a long time. This research added evidence that linguistic laypersons assess oral performances differently from language professionals. Additionally, it was suggested that language teachers’ experience might influence their assessments of linguistic forms. Similar to Hadden (1991), however, the difference in severity of criteria was the primary focus in the research.

Unlike the studies above, some research has investigated differences in the factors influencing overall judgments made by language specialists and linguistic laypersons. For investigating what factors contribute to different raters’ perceptions of oral proficiency in Spanish, Galloway (1980) compared the evaluative judgments made by four groups of raters: (a) non-native Spanish teachers, (b) native Spanish teachers, (c) native Spanish non-teachers with high English proficiency, and (d) native Spanish
non-teachers without English proficiency. Oral performances of 10 university-level students of Spanish were videotaped and rated by the four groups of raters using the analytic criteria of *Quantity of communication, Efforts to communicate, Comprehensibility, Effects of paralanguage*, and *Overall impression*. They were also asked to provide written comments on the students’ performances. The results showed that non-teacher groups evaluated students’ efforts to communicate more positively, even though grammatical errors were evident in the performances. Their comments also indicated that visible efforts and desire to communicate positively affected non-teachers. In addition, they commented mostly on the content of performance and expressed boredom with the students’ presentations. In contrast, the teacher groups did not pay attention to paralinguistic features but focused solely on grammatical forms. These results seem to support Ludwig’s (1982) assertion that:

> Perhaps as a result of contact with too much “bad French” (“bad German,” “bad Spanish”) over the years, FL [foreign language] teachers are overly sensitive to errors of pronunciation and grammar, which reflect form rather than content. On the other hand, NS’s who are not involved in teaching have no personal stake in the form of the message. They see the production as a communicative attempt and immediately begin to assign it meaning within the context of the utterance. (p. 281)

Accordingly, professional interests as language teachers are considered to make them strongly attend to language forms. Conversely, linguistic laypersons may be concerned more with content or attitudes than linguistic forms. Accordingly, raters with different occupational backgrounds tend to judge specific features of performance differently. Language specialists are generally sensitive
to linguistic forms and more severe on linguistic errors. Furthermore, language teachers tend to focus on language proficiency or grammatical accuracy when making an overall judgment of communicative ability. However, these features seem to play a minor role in linguistic laypersons’ evaluative judgments. Research on the raters’ judgments of test-takers’ writing performance has also suggested that raters’ occupational background (language teachers or otherwise) and experience as professional raters influence how they make evaluative judgments (Cumming, 1990; Cumming, Kantor, & Powers, 2001, 2002; Huot, 1993; O’Loughlin, 1992; Shohamy, Gordon, & Kraemer, 1992; Song & Caruso, 1996; Weigle, Boldt, & Valsecchi, 2003).

Professional experience as language teachers is considered to make language specialists assess L2 performances differently from linguistic laypersons. Since language teachers’ main duty is to teach an L2, they are inclined to concentrate on L2 learners’ linguistic aspects of performance more strongly than linguistic laypersons do; this notion is supported by empirical studies (Galloway, 1980; Ludwig, 1982). It might be that language teachers are comfortable with attending to and pointing out grammatical errors while judging communicative performance (Brown, 2007). However, language teachers do not always underscore grammatical accuracy in oral performances. Their practice of meaning-focused approaches to language teaching (communicative language teaching and task-based language teaching) may also influence their judgments (Sato, 2012). Although the role of grammatical accuracy in language teachers’ overall judgment is complex as such, L2 teachers’ duties and prevailing pedagogical approaches typically shape their beliefs about which features should be emphasized in assessment of oral performances.

The teachers’ experience and familiarity with oral performances produced by particular L2 learners also make their judgment unique and different from how
linguistic laypersons perceive L2 oral performance. For example, Brown (1995) notes that “Teachers, through their experience of dealing with language at a very detailed level may be no longer able to make intuitive, global evaluations such as a ‘naive’ language user would” (p. 13). In a similar vein, with regard to familiarity with L2 performance, Hadden (1991) argues that “after years of becoming accustomed to students’ language, teachers may no longer be able to determine which errors impede comprehension” (p. 3).

Furthermore, experience as trained raters is considered to affect raters’ judgment. Raters’ intuitive judgments are inevitably influenced by rating criteria that they have previously used. This phenomena was reported in research on judgments of writing performance (Cumming et al., 2001, 2002), but it may also be applicable to oral assessment. In their research, participants with extensive experience in teaching and rating L2 writings were asked to assess L2 compositions without referring to any rating criteria and to provide verbal reports on their thought process. Their verbal reports indicated that most participants acknowledged that their experience as raters influenced their evaluative judgments. As suggested by Brown (1995) and Hadden (1991), Cumming et al. argue that highly professional raters may not be able to assess L2 performance in such a way that linguistic laypersons would do; “Indeed, it may be relatively difficult for experienced raters to ‘unlearn’ or shift themselves away from particular criteria or procedures that they have become skilled at using” (Cumming et al., 2001, p. 13).

In conclusion, language specialists who have experience in teaching or rating L2 learners construe L2 oral communicative ability in a unique way, influenced by their professional views and experiences. Thus, they may not be able to judge L2 performance as linguistic laypersons do, even though they are allowed to rely on their
2.3.3 Indigenous Assessment Criteria

It has been shown that language teachers’ perspectives on communication reflect their experience and professional training and differ from linguistic laypersons’ perspectives. Therefore, research has been conducted on the unique perspectives of various domain experts who are also linguistic laypersons, who become the ultimate arbiters of communication ability in real-world domains, including medical doctors, non-linguistics-related subject teachers, and professionals in various academic fields. Their unique perspectives on communication form indigenous assessment criteria, which are “used by subject specialists in assessing the communicative performances of apprentices in academic and vocational field” (Douglas, 2000, p. 68).

Jacoby (1998) explored the indigenous criteria used by physicists in assessing oral presentations of post-doctoral researchers and PhD candidates (see also Jacoby & McNamara, 1999). She analyzed a senior physicist’s feedback on the presenters’ (both non-native English speakers (NNESs) and native English speakers (NESs)) rehearsals for conference presentations. Table 2.8 summarizes 11 criteria used by the physicist while giving feedback. The intriguing finding of the research is that the senior physicist appeared to adopt non-linguistic criteria for both NNES and NES presenters and paid little attention to linguistic errors made by NNES presenters. Linguistic errors pointed out in the sessions were mainly those in spelling, preposition choices, and irregular past tense found in presenters’ overhead transparencies. One of the reasons for lack of comments on linguistic aspects may be the relatively high English proficiency of the NNES presenters. However, the criteria are quite different from conventional linguistically oriented assessment criteria employed by the majority of English
Abdul Raof (2011) also investigated indigenous assessment criteria employed by Civil Engineering specialists for evaluating conference presentations. In this research, experienced engineering specialists evaluated presentations given by Malaysian engineering specialists at international conferences without rating scales. Afterwards, indigenous criteria were elicited through semi-structured interviews, and rating scales were developed based on their criteria. The criteria consisted of (a) Delivery (legibility of visual materials, confidence, and communication with audience), (b) Organization (flow of points and outlining), and (c) Subject matter (mastery of knowledge, clarity of explanation, and topic of presentation). These indigenous criteria were similar to those found by Jacoby (1998) in which how to deliver a speech and the content of a speech were more strongly highlighted than linguistic accuracy or complexity. Nonetheless, it was likely that the presenters surveyed had already achieved high levels of English proficiency.

In an educational setting, Elder (1993) investigated how subject teachers (math and science) judge teacher trainees’ teaching practice and compared the judgments made by language teachers. Math and science teachers observed L2 speakers’ teaching

---

**Table 2.8**

*Indigenous Assessment Criteria Applied by Physicists*

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>i Timing</td>
<td>vii Argumentation/Persuasion</td>
</tr>
<tr>
<td>ii Newsworthiness/Significance of research</td>
<td>viii Content accuracy</td>
</tr>
<tr>
<td>iii Visual coherence</td>
<td>ix Technical delivery</td>
</tr>
<tr>
<td>iv Clarity</td>
<td>x Overall quality</td>
</tr>
<tr>
<td>v Explicitness/Completeness</td>
<td>xi Linguistic error</td>
</tr>
<tr>
<td>vi Economy of expression</td>
<td></td>
</tr>
</tbody>
</table>
practice, indicating their global impressions and assessing multiple criteria (Intelligibility, Fluency, Accuracy, Comprehension, Subject-specific language, and Interaction). Correlational and multiple regression analyses were conducted to examine which analytic criteria predicted their global impressions of teaching practice. The subject teachers’ global impression was significantly correlated with interactional behaviors such as posing questions to check students’ understanding or giving clear instructions. These behaviors were also the strongest predictors of their impressions. This finding indicates that the main concern of subject teachers was with the speakers’ behaviors making students’ learning more successful rather than the form of language.

In a similar context, Plough et al. (2010) investigated how faculty evaluators assess prospective graduate student instructors’ (GSIs) teaching practice. The evaluators were from non-linguistics-related faculties such as mathematics, chemistry, economics, and so forth. While evaluation GSIs, these faculty members seemed to strongly focus on (a) listening comprehension, (b) pronunciation, and (c) responses to questions. Thus, it was considered that comprehending students’ questions and responding to them with clear pronunciation were perceived crucial for GSI duties. Similarly to Elder (1993), they found that the faculty members were concerned with features that made the classroom interaction successful, rather than with linguistic accuracy.

Douglas and Myers (2000) studied the communication skills of veterinary students by investigating the indigenous criteria of veterinary students, veterinary professionals, and applied linguists (see also Douglas, 2001). These groups of participants watched videotapes of simulated interactions between clients and veterinary students. First, the students commented on any aspects that they thought intriguing or problematic in their own performances. Then, veterinary professionals and applied linguists were invited to discuss each student’s performance and evaluate it based on
their impressions. Their comments showed that all three groups of informants evaluated the performances using similar criteria overall, such as Demeanor, Knowledge base, Timing, Coverage, Phraseology, and Appearance. This finding conflicts with research that has demonstrated differences in judgments between linguistic laypersons and language specialists. Nevertheless, this might be attributable to the test-takers’ high language proficiency, which can be inferred from the fact that there were no linguistic-related features in the comments provided.

These studies suggest that assessment criteria derived from domain experts are considerably different from conventional linguistic-oriented criteria for L2 oral proficiency tests, as shown in Table 2.9. Compared with weak performance tests, domain experts appear to orient to a strong approach to performance testing, focusing on real-life criteria or task fulfillment. Ubiquitous linguistic-related features—grammar, vocabulary, pronunciation, and fluency—are not evident in their indigenous assessment. However, this does not necessarily mean that speakers’ linguistic features are unimportant. Domain experts’ major focus on non-linguistic criteria or task fulfillment might be attributable to the high language proficiency of speakers. Participants in the research could generally be considered highly proficient, and to not have serious linguistic problems in their performances. This issue will be discussed in Section 2.4.3. It also might be that domain experts, as linguistic laypersons, did not have the meta-language necessary to describe linguistic problems in the performances, even though they may have attended to language.
Table 2.9

*Indigenous Assessment Criteria Derived from Various Domain Experts*

<table>
<thead>
<tr>
<th>Researcher(s)</th>
<th>Domain experts</th>
<th>Example assessment criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elder (1993)</td>
<td>Subject teachers</td>
<td>Classroom interactions</td>
</tr>
<tr>
<td>Jacoby (1998)</td>
<td>Physicists</td>
<td>Timing, Significance of research, etc.</td>
</tr>
<tr>
<td>Douglas &amp; Myers (2000)</td>
<td>Vet students/professionals</td>
<td>Demeanor, Knowledge base, Timing, Coverage, Appearance, etc.</td>
</tr>
<tr>
<td>Plough et al. (2010)</td>
<td>Faculty evaluators of GSIs</td>
<td>Listening comprehension, Pronunciation, Responding to questions</td>
</tr>
<tr>
<td>Abdul Raof (2011)</td>
<td>Engineering specialists</td>
<td>Delivery, Organization, Subject matter</td>
</tr>
</tbody>
</table>

These studies on domain experts’ (i.e., linguistic laypersons’) perspectives contributed to defining specific-purpose communication ability required in particular domains. Jacoby and McNamara (1999) argue that “studies of naturally occurring ‘indigenous’ socialization and assessment practices in professional settings, can provide more direct access to what counts as communicative competence in particular contexts” (p. 214).

2.3.4 Experts in English Communication in the Globalized World

A crucial point to be considered for this study is who should be regarded as “experts” and whose views should be incorporated into the theory of L2 communication ability. The previous sections have discussed that linguistic laypersons should be the ultimate arbiters of L2 communication ability in real-life domains and that L2 communication theory should adopt their perspectives. Another key issue for this study is whether non-native English speaking (NNES) laypersons should be included as experts of English communication. This issue should be considered carefully in light of the current prevalence of English use in the globalized world.

English has spread all over the globe and is used as an international language.
It differs from any other languages in that English is used by NNESs more than by NESs. Kachru (1988) described the following three distinct groups of English users: the Inner Circle, the Outer Circle, and the Expanding Circle. His description is considered as a useful and influential model depicting the spread of English as well as the patterns of acquisition and the function of English, but several limitations of it have been acknowledged (Jenkins, 2009b).

The Inner Circle consists of countries using English as native language (ENL) such as the U.S., the U.K., Australia, and so forth. The Outer Circle consists of countries where English is used as a second language (ESL), such as India, Singapore, Philippines, and more than fifty other territories. In these multilingual nations, English is used in chief institutions as an important intra-national lingua franca. The Expanding Circle consists of English as a foreign language (EFL) countries such as Japan, Germany, Indonesia, and an increasing number of other countries. Although English does not have any special administrative status, it is used for a wide variety of purposes inside and outside the countries. Crystal (2003) estimated the population of the three circles as 320-380 million in the Inner Circle; 300-500 million in the Outer Circle; and 500-1,000 million in the Expanding Circle. According to his estimation, the ratio of NNESs to NESs is approximately 3:1. It has been acknowledged that NNESs are major users of English, and ENL speakers are the minority in various international domains.

A considerable number of speakers of English are users of English as a lingua franca (ELF). A lingua franca is known as “a contact language used among people who do not share a first language, and is commonly understood to mean a second (or subsequent) language of its speakers” (Jenkins, 2007, p. 1). Some researchers used to regard ELF communication as English used among NNESs with different L1s and as excluding NESs (Firth, 1996; House, 1999). Currently, however, most ELF researchers
take a broader view, and include all English users in the definition of ELF (Jenkins, 2009a). Seidlhofer (2011), for example, defined ELF as “any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option” (p. 7). She included those from the Inner and Outer Circles in ELF communication. Nonetheless, ELF communication frequently occurs among speakers from the Expanding Circle without NESs’ presence. Jenkins (2009b) notes that one of the fundamental characteristics of ELF communication is that “it [English] is used in contexts in which speakers with different L1s (mostly, but not exclusively, from the Expanding Circle) need it as their means to communicate with each other” (p. 143). This study adopts the definition of Seidlhofer (2011) when discussing ELF.

The term lingua franca can connote low-level language use (Seidlhofer, 2011). However, ELF communication frequently occurs in highly specialized domains or influential frameworks in House’s (1999) term, including “global business, politics, science, technology and media discourse” (p. 74). House adds that

> English has gained a dominant role in the world in politics, business negotiations, cultural and scientific congresses, workshops and symposia, international student meetings, and wherever members of smaller, ‘exotic’, ‘less commonly taught’ languages have occasion to talk to members of larger language communities, who see no reason to acquire, and converse in, the ‘smaller language’. (p. 73)

This shows an important role of ELF in terms of its functions and the domains in which the language is spoken. English even plays a crucial role within the borders of Expanding Circle countries, such as in the realm of higher education (Canagarajah,
2006; Jenkins, 2007). More and more NNESs thus use English because they wish to join these specialized communities of practice and to be acknowledged as legitimate members.

ELF is different clearly from English as a foreign language (EFL). The fundamental assumption of EFL is that the goal of L2 speakers (often regarded as learners) is to communicate with NESs and assimilate into NES communities. Therefore, acquiring NES-base communicative competence (e.g., Canale & Swain, 1980) is required. Deviance from NES norms is regarded as deficit or errors; for example, code-mixing and code-switching are taken as interference errors demonstrating a lack of competence (Jenkins, 2006a). In this sense, the notion of standard (NS) language ideology—a belief that the standard variety should be regarded as the only legitimate version of language—underpins EFL. In contrast, ELF does not require conformity to NES norms since ELF communication often occurs without NESs’ presence. Instead, the essential objective of ELF is achieving mutual understanding and intelligibility. ELF lingua-cultural norms are co-constructed among the participants. Seidlhofer (2011) claims that

these norms [ELF norms] are tacitly understood to be established during the interaction, within the current possibilities, and … they are primarily regulated by interactional exigencies, rather than by what native speakers would say, or would find correct, or ‘normal’, or appropriate. (p. 18)

Deviance from NES norms is considered as difference rather than deficit, and code-mixing and code-switching are taken as natural phenomena that occur when utilizing bilingual resources (Jenkins, 2006a). Table 2.10 summarizes the fundamental
conceptual differences between EFL and ELF (see also Jenkins, 2006a, 2009a; Kirkpatrick, 2007; Seidlhofer, 2011).

Table 2.10

<table>
<thead>
<tr>
<th></th>
<th>Foreign language (EFL)</th>
<th>Lingua franca (ELF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary objectives</td>
<td>integration, membership in NES community</td>
<td>mutual understanding, intelligibility</td>
</tr>
<tr>
<td>Target interlocutor</td>
<td>exclusively NESs</td>
<td>primarily NNESs, NESs</td>
</tr>
<tr>
<td>Linguacultural norm</td>
<td>pre-existing NES norms</td>
<td>co-constructed among participants</td>
</tr>
<tr>
<td>Deviance from NES norm</td>
<td>deficit, interlanguage, broken English</td>
<td>difference, variance, natural phenomena</td>
</tr>
</tbody>
</table>

ELF communication yields some unique features that cannot usually be observed in communication among monolingual English speakers. One of such features is *let it pass* (Firth, 1996). It is an interactional strategy to ignore unclear or unknown expressions that are non-fatal or redundant in the interaction. Accordingly, ELF communication that includes linguistic forms deviant from NES norms is interactionally robust, and communication breakdown does not always occur due to these forms. Furthermore, several ELF corpora have revealed various lexico-grammatical features deviant from NES norms but widely prevalent in ELF communication (see Table 2.11). These lexico-grammatical expressions are not likely to hamper message conveyance. These are regarded as *innovative* rather than idiosyncratic or errors (Cogo & Dewey, 2012; Seidlhofer, 2004, 2011).
Table 2.11

<table>
<thead>
<tr>
<th>ELF Lexicogrammar</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropping the third person present tense –s</td>
<td>If someone take a disadvantage ...</td>
</tr>
<tr>
<td>Confusing the relative pronouns</td>
<td>children which need a family</td>
</tr>
<tr>
<td>Omitting definite and indefinite articles, and inserting them</td>
<td>There are lot of foreigner.</td>
</tr>
<tr>
<td>Failing to use correct forms in tag questions</td>
<td>They should study hard, isn’t it?</td>
</tr>
<tr>
<td>Inserting redundant prepositions</td>
<td>We have to study about ...</td>
</tr>
<tr>
<td>Overusing certain verbs of high semantic generality (e.g., do, have, make, put,</td>
<td>took drive license, do university</td>
</tr>
<tr>
<td>take)</td>
<td></td>
</tr>
<tr>
<td>Replacing infinitive-constructions with that-clauses</td>
<td>I want that ...</td>
</tr>
<tr>
<td>Overdoing explicitness</td>
<td>black color</td>
</tr>
</tbody>
</table>

Given the current use of English all over the world, adopting and conforming only to NES norms and NESs’ perspectives would appear inappropriate. As discussed above, a large part of English spoken in the world is ELF, often where NESs are not present, and the use of forms deviant from NES norms does not always cause problems among NNESs. Therefore, there is no necessity to adopt the NES norms. Seidlhofer (2011) claims that “it would be interactionally counter-productive, even patently absurd in most cases, for speakers to (strive to) adhere to ENL linguacultural norms when no ENL speakers may even be present” (p. 18). Moreover, one of the important purposes for learning English in the current globalized world is acquiring registers used in professional communities that L2 speakers wish to participate in (Widdowson, 2003). Since many communities do not always entail NESs, conformity to NES norms is not necessary. Instead, speakers need to learn how to communicate in such a way that the community members accept them as a legitimate member. If the members of the community are NNESs, their norm should be adopted for better mutual understanding. As Widdowson (2003) states, “How English develops in the world is no business
whatever of native speakers in England, the United States, or anywhere else. They have no say in the matter, no right to intervene or pass judgement. They are irrelevant” (p. 43).

In fact, some L2 speakers might wish to achieve native-like proficiency by adhering to NES norms. However, deferring to NES norms is not strictly necessary unless English speakers only aim to be a member of a particular NES community. It is claimed that native-like English serves as a kind of shibboleth in order for NNESs to be accepted in the NES community (Seidlhofer, 2011; Widdowson, 2003). In this sense, NES norms play a minor role in intelligibility, only playing a major role in integration to a NES community. Instead of acquiring lexico-grammatical accuracy based on NES norms only, L2 speakers also have to acquire what is important for effective ELF communication. It is thus important to know what is perceived as “good communication”, not only from the perspective of NESs but also from NNES norms and perspectives.

Conforming only to NES perspectives is also problematic for effective ELF communication. NESs are asserted to be not so adept at ELF communication or accommodating their speech to NNESs. Phillipson (2003) argues that:

Native speakers have greater facility in speaking the language, but not necessarily greater sensitivity in using it appropriately. In many international fora, competent speakers of English as a second language are more comprehensible than native speakers, because they can be better at adjusting their language for people from different cultural and linguistic backgrounds. (p. 167)

This suggests that NESs might not necessarily possess skills for effective ELF
communication, including “sociolinguistic skills of dialect differentiation, code switching, style shifting, interpersonal communication, conversation management, and discourse strategies” (Canagarajah, 2006, p. 233). Their lack of sensitivity and skills in using language in ELF settings may frequently cause communication problems. In this sense, NESs’ ability to communicate in ELF settings is not necessarily higher than NNESs’ ability.

Seidlhofer (2011) also argues that NESs may make evaluative judgments of communicative ability without taking account of ELF communication:

> What ENL speakers find easy to process and comfortable or pleasant to read or listen to is not an a priori consideration for ELF interactions, not one that is more important than the perceptions and communicative preferences of non-native ELF users. (p. 40)

This is because ENL speakers as well as ESL speakers use English primarily for domestic purposes and do not necessarily consider appropriate intercultural ELF communication. Some empirical studies have demonstrated that NNESs and NESs make evaluative judgments of L2 speakers’ oral performance differently (Kim, 2009a, 2009b; Riney, Takagi, & Inutsuka, 2005; Zhang, 2010; Zhang & Elder, 2011). They have indicated that NNESs and NESs attend to different features of L2 oral performance while judging language proficiency or accent. Someone judged as a competent communicator by NESs may not receive the same evaluation from NNESs. Seidlhofer (2011) contends that:

> What it means to be communicatively competent in English can no longer be described with reference to norms of linguistic knowledge and behaviour that are relevant only to
particular native-speaker communities. Conformity to these norms is neither necessary nor sufficient to meet the international demands for the effective use of English as a lingua franca. (p. 92)

In summary, communicative performances positively judged from NES norms and perspectives cannot be guaranteed to be positively judged by the majority of members in ELF communication. Adopting NES perspectives per se is unnecessary and potentially problematic for effective ELF communication. This makes it important to understand NNES perspectives on communication ability and ideal communication.

2.3.5 Summary

This section has provided an argument that it is necessary to investigate and understand linguistic laypersons’ perspectives on communication. The main rationales are that (a) they are the ultimate arbiters of L2 communication ability in real-world contexts and (b) linguistic laypersons have unique perspectives different from language specialists. This section has also discussed who should be regarded as “expert English users” in the globalized society. This study has proposed that the perspectives of proficient NNESs should be included as a legitimate source contributing to theory construction and language testing. Investigation of perspectives held by linguistic laypersons including NNESs is believed to broaden the scope of theory of communication ability required for ELF communication in real-life contexts.

2.4 Previous Studies on Raters’ Judgment of Oral Performance

In the field of language testing, researchers have been concerned with raters’ thought processes during rating oral performances and the assessment criteria
contributing to their overall judgments. Since the main purpose of these studies has been to validate raters’ judgments in the tests, most of the rater participants were accredited raters and language teachers (i.e., language specialists). Nevertheless, this line of empirical studies is of relevance to the present study in the sense that they investigated the features contributing to people’s holistic judgments of oral performances. This section reviews such studies and discusses the methodological issues concerned.

2.4.1 Raters’ Focus on Language Proficiency

Some empirical studies have demonstrated that components of language proficiency—grammar, vocabulary, and pronunciation—strongly contributed to raters’ judgments of oral performances. McNamara (1990) investigated how individual analytic criteria are interpreted by accredited raters of the Occupational English Test (OET), which is an English test used to screen immigrant health professionals in Australia. In its speaking sub-test, test-takers are given role-play tasks simulating patient consultations. The raters awarded scores for five analytic criteria: (a) Intelligibility, (b) Fluency, (c) Comprehension, (d) Appropriateness of language, and (e) Resources of grammar and expression. In addition, they indicated their overall impressions of the performances labeled as Overall communicative effectiveness. Since this holistic criterion did not specify which features to gauge, it broadly reflected the raters’ perspective of effective communication in the task. Through a many-facet Rasch analysis and multiple regression, McNamara revealed that the scores for Overall communicative effectiveness and Resources of grammar and expression were highly interdependent, suggesting that the raters closely associated the effectiveness of communication with the grammar and vocabulary used by the test-takers. This finding
was unexpected since the raters were trained not to overemphasize linguistic accuracy as long as it did not hamper message conveyance. Despite the training, they seemed to allocate a heavier weight to linguistic accuracy than any other criteria. In this regard, this study touched on the raters’ unconscious orientation that cannot be revealed by other methods (e.g., self-reporting).

Several studies employed a retrospective verbal protocol analysis and found that grammatical accuracy contributes to raters’ overall judgments of oral performances (Brown, 2007; Brown et al., 2005; Kim, 2009a, 2009b; Zhang, 2010; Zhang & Elder, 2011). These studies all asked the raters to (a) rate oral performances holistically, (b) provide the reasons for their ratings, and (c) elaborate on the reasons reviewing the performances. Based on this method, Brown (2007) explored the features heeded by eight accredited IELTS examiners when awarding scores to test-takers’ oral performances. She found that the largest groups of comments (about 40% of all comments) were on syntax (accuracy, maturity, and range) and vocabulary. This finding might indicate that syntax was highly salient to raters, although Brown acknowledged that it should be interpreted carefully. The large number of rater comments on syntax was not surprising because the raters were asked to rate the performance using the IELTS scales, which include Grammatical range and accuracy. Thus, the verbal reports reflected the raters’ interpretations of the rating scales rather than their personal impressions of performances, as was the case in studies on the OET (McNamara, 1990).

With the same method, Brown et al. (2005) discovered what features contributed to 10 English-for-academic-purposes (EAP) specialists’ overall evaluations of oral performances on the TOEFL iBT (monologic independent and integrative tasks). The impetus for this study was to identify the salient criteria for testing performance rather than confirming how raters apply its rating scales. Accordingly, the raters were
asked to rate the oral performances based on their overall impressions. The result showed that the criteria they commented on most frequently by far were linguistic resources such as grammar, vocabulary, expression, and textualization; they attended to the accuracy, sophistication, and range of structures of language produced. This finding, as Brown (2007) further explored, indicates that the raters might strongly attend to test-takers’ linguistic resources and accuracy. In contrast to Brown’s (2007) study, however, the 2005 study revealed raters’ impressions that were not related to given assessment criteria, and the comments reflect the raters’ pure judgments of communication ability during the test performance. The raters’ attention to linguistic features may be attributable to the occupational background of the raters; linguistic aspects drew their attention since assessing language is one of the main duties of EAP specialists.

Iwashita, Brown, McNamara, and O’Hagan (2008) investigated the features differentiating the five proficiency levels of oral performances on the TOEFL iBT. They examined oral performances sorted into five proficiency levels by trained raters using a draft holistic scale (Lee, 2005). Specific linguistic features of the oral performances were measured through discourse analysis of three features chosen based on the results of Brown et al. (2005): (a) linguistic resources (accuracy, complexity, and vocabulary), (b) phonology (pronunciation, intonation, and rhythm), and (c) fluency of the performances at each level. This study thus examined the relationship between subjective holistic ratings and objectively measured linguistic features. ANOVA was performed to reveal what features significantly differentiated proficiency levels. The main finding of the study was that most of the linguistic features significantly differentiated the performances. More specifically, vocabulary (type and token) and fluency (pauses and speech rate) strongly discriminated the proficiency levels. This
indicated that raters’ holistic judgments were not determined by a single linguistic aspect but rather by multiple categories. At the same time, the strong contributions of vocabulary and fluency support the findings of the studies above (Brown, 2007; Brown et al., 2005; McNamara, 1990). However, the holistic judgments in this study were derived from the raters using a holistic scale, which included statements on linguistic resources, pronunciation, and fluency. In this regard, this study revealed the raters’ interpretation of the rating scale as in Brown’s (2007) study.

De Jong et al. (2012) investigated the features contributing to the global assessments (labeled as Functional adequacy) of L2 Dutch oral performances (reviewed in Section 2.2.3). They examined the relationship between the global assessments and three areas of skills measured by eight independent tests, which were (a) linguistic knowledge skills (vocabulary and grammar), (b) speed-of-processing skills (lexical retrieval and sentence building), and (c) pronunciation skills (speech sounds, word stress, and intonation). Four linguistic lay judges were asked to rate learners’ monologic oral performances on computer delivered role-plays by applying rating criteria that assessed functional (informational) adequacy: amount and detail of information, setting, relevancy to the topic, discourse type, and how easy it is to follow the description. A statistical analysis of the scores using SEM revealed that all the three areas of skills were significantly correlated with the raters’ judgments of oral proficiency, except for articulation speed (a subcategory of processing skills). Among the knowledge and skills measured by the tests, vocabulary knowledge and intonation significantly predicted proficiency judged by the raters. This study showed that speaking proficiency in terms of message conveyance was determined by linguistic knowledge (vocabulary, grammar, and pronunciation) and cognitive fluency. It should be noted that this study differed from the studies reviewed above (Brown, 2007; Brown et al., 2005; Iwashita et al.,
2008; McNamara, 1990) in that (a) the global assessment criterion did not include linguistic features and (b) the performances were judged by linguistic laypersons.

In summary, the studies reviewed in this section demonstrated that linguistic resources (grammar and vocabulary), pronunciation, and processing skills (fluency) affected or were correlated with raters’ judgments of performance on various speaking tests. Most of the studies (Brown, 2007; Brown et al., 2005; De Jong et al., 2012; Iwashita et al., 2008) suggest that linguistic accuracy is not the sole factor related to oral performances, although its impact turned out to be relatively larger than other components. Raters appear to believe that a competent performance entails the accurate use of grammar and vocabulary, clear pronunciation, and utterance fluency, as many L2 theories have argued.

2.4.2 Raters’ Focus on Non-linguistic Features

Applied linguists have not satisfactorily specified and theorized non-linguistic features underlying L2 communication (McNamara, 1996). Nevertheless, empirical studies have demonstrated that raters of speaking tests consider not only linguistic aspects of test-takers’ performance but also features that have not been proposed in theories. These features are not usually included in rating criteria used for tests and are thus called *implicit criteria* (Brown, 2007), *non-criterion features* (Orr, 2002), or *test criterion-independent category* (Zhang & Elder, 2013) in the language testing literature.

2.4.2.1 Content

The content or elaboration of speech has been found to affect raters’ overall judgment. Eckes (2009) investigated the criteria perceived to be important by raters of TestDaF, a German test used for screening applicants for higher education. In a
questionnaire survey, 53 trained raters (experienced teachers of German) were asked to indicate the importance they attached to the following criteria: (a) Content, (b) Vocabulary, (c) Correctness, (d) Adequacy, (e) Completeness, (f) Description, (g) Discussion, and (h) Standpoint. Rasch analysis of the perceived importance revealed that Content—the degree to which the message of the oral performance is meaningful—was believed to be the single most important criterion. In contrast, Correctness (related to morphosyntactic and lexical errors) was perceived to be the least important. This study employed self-reports that represent participants’ general beliefs rather than actual behaviors. However, the finding suggested that at least on conscious level the raters highlighted the importance of content than the other linguistic criteria.

The emphasis on the content in the academic context was also confirmed by Brown (2007) and Brown et al. (2005). These studies, as reviewed earlier, employed a retrospective verbal protocol analysis and investigated the criteria affecting accredited raters of two EAP speaking tests: IELTS and the TOEFL iBT. Although both studies found that linguistic resources were most frequently mentioned, the content (including the sophistication of ideas, task fulfillment, and discourse organization) was the second largest category. Brown (2007) argues that the test purpose (i.e., assessing languages for academic purposes) might have led raters to emphasize the content of speech, as intellectual maturity is relevant to success in higher education. Furthermore, Brown et al. (2005) claim that the finding “certainly indicates the need for content to be taken into account when developing scales for the assessment of cognitively complex speaking tasks” (p. 33).

Research has also discovered that raters of general-purpose English tests orient to the content of speech. Sato (2012) investigated the relative contribution of four linguistic-related criteria (Grammatical accuracy, Vocabulary range, Fluency, and
Pronunciation) and the elaboration of content (Content elaboration/development) to the raters’ overall impressions of performance, labeled Overall communicative effectiveness. Nine English teachers first rated university students’ monologic performances on a general-purpose speaking test based on their impressions. They were then asked to assess the four linguistic features and the elaboration of content. Their scores were statistically analyzed using the many-facet Rasch model and multiple regression. The results showed that scores for Content elaboration/development most strongly predicted those for Overall communicative effectiveness. This strong focus on speech contents reflected raters’ pure impressions because the scores for Overall communicative effectiveness were not dependent on any pre-determined rating criteria. One of the plausible reasons for this finding is the influence of message- and content-based pedagogical approaches. All the raters were English teachers with master’s degrees, and must have learned and practiced communicative approaches to language teaching. The practice of communicative models of language teaching may affect language teachers’ perception of “good oral performances” and they might devalue the importance of linguistic resources in communicative tasks accordingly.

Using a similar approach, Hinofotis, Bailey, and Stern (1981) investigated what criteria could be used to predict the global score on an English examination for screening applicants for foreign teaching assistantships. In the test, test-takers engaged in role-play that required them to explain a concept that students had difficulty in understanding. Six English teachers assessed 10 applicants’ videotaped oral performance based on their overall impressions and three subcategories: Language proficiency (e.g., grammar, vocabulary), Delivery (e.g., enthusiasm, eye contact), and Communication of information (e.g., development of explanation). Multiple regression analysis revealed that Communication of information most strongly predicted the raters’
impressions of the oral performance. This result suggests that the ability to give clear explanations or use supporting evidence is considered more essential than the correct use of grammar and vocabulary in the context of the duties of teaching assistants.

2.4.2.2 Interactional resources and co-construction

Interactional competence theorists posit that participants bring interactional resources and use them during discursive practices (He & Young, 1998; Young, 2000, 2008). Empirical research has supported the idea that interactional resources contribute to raters’ judgments in interactive tasks. It has also been discovered that raters take into consideration the performance co-constructed by paired test-takers, focusing on the mutual achievement. Accordingly, listening comprehension has also been found to be a factor underlying raters’ judgments of performance.

May (2009, 2011a, 2011b) reported on how raters of a paired EAP speaking test interpret the interactive effectiveness by analyzing their verbal protocols. Four experienced EAP teachers were asked to rate paired oral performances on structured discussion tasks, using analytic rating scales composed of Fluency, Accuracy, Range, Interactional effectiveness, and Overall. This study was concerned with what features the raters heeded when awarding scores for Interactional effectiveness, which entailed (a) understanding the interlocutor’s message, (b) responding appropriately to the interlocutor, and (c) using appropriate communicative strategies. It was found that the raters commented on all the three features stated in the category of Interactional effectiveness. The raters also mentioned the mutual achievement co-constructed by both test-takers and their contributions to an authentic interaction. The features most frequently commented on were: understanding of the partner, responding to the partner, working together cooperatively, contribution to an authentic interaction, and quality of
the interaction. When they encountered asymmetric interactions (unbalanced contributions from paired test-takers), the raters compensated for or penalized individual test-takers by speculating on their potential performance with different partners. This study thus revealed that raters pay attention not only to individual test-takers’ performances but also to co-constructed interactions.

Raters’ attention to co-constructed performance was also reported by Galaczi (2008), who investigated the relationship between interactional patterns and scores for Interactional communication (IC) in the Cambridge First Certificate in English Examination (FCE). This criterion “refers to the candidate’s ability to take an active part in the development of the discourse” (UCLES, 2007, p. 86). Using CA, she identified four qualities of interactional patterns in peer-peer interactions exhibited by the test-takers: (a) collaborative (high mutuality and equality), (b) parallel (low mutuality and high equality), (c) asymmetric (moderate mutuality and low equality), and (d) blended. These patterns were related to how interactions were constructed by each test-taker’s contributions. The analysis of test scores showed that test-takers who exhibited the collaborative pattern received the highest IC score. The parallel interaction pattern received the lowest mean. Similar to May (2011a), Galaczi examined scores for one specific criterion related to interaction, not scores for holistic judgment. In this sense, their studies did not show the significance of co-constructed interactions in raters’ overall judgments of oral performances. Nor did they demonstrate the relative importance of interactive resources compared with other linguistic criteria. Importantly, however, their studies have suggested that raters’ judgments of peer-peer interaction are affected by the mutually achievement of interaction, not only by an individual test-taker’s sole performance.

Similar to May and Galaczi’s studies, Ducasse (2010) investigated raters’
judgments of peer-peer paired interactions in a Spanish speaking test (see also Ducasse, 2009; Ducasse & Brown, 2009). In her study, Spanish language teachers produced verbal protocols on their orientation to paired interactions performed by beginner-level Spanish students. They were not instructed on what to focus on while viewing the oral performances of test-takers; they were simply asked to comment on what made interaction successful or unsuccessful. The results showed that three salient categories were identified: (a) non-verbal interpersonal communication, (b) interactive listening, and (c) interactional management. The largest proportion of comments (48.7% of all comments) concerned interactional management, which includes topic change, turn organization and length, and topic development. It indicated that raters attended to how test-takers had used interactional resources in the interactions. In addition to these resources, Ducasse found that test-takers’ interactive listening—backchannels, signs of comprehension, and supportive listening—drew raters’ attention. It was suggested that the raters’ judgments of oral performances included comprehension of what the interlocutor said and attempts to co-construct interactions supportively, confirming what Galaczi (2008) had supposed. Ducasse’s study thus indicated that raters’ intuitive judgments of successful interactions were related to the co-constructed nature of interaction.

How test-takers interact with trained examiners (interlocutors) was also found to impact ratings in Brown’s (2005) study on interactions in IELTS. She asked eight accredited raters to assess performances on oral interviews as they normally do by applying the IELTS bandscale. Afterwards, they provided the justifications for their ratings. Brown reported that the examiners’ interactional patterns (structuring of topical sequences, questioning technique, and feedback and rapport) were different, and the test score was also different according to the patterns. In particular, it was shown that a
test-taker’s score on *Communicative effectiveness* varied markedly according to the examiner. The analysis of the justifications given by the raters showed that test-takers’ interactional behaviors—initiations, the quickness of response, and strategic behaviors—greatly influenced their judgments. Furthermore, all the raters commented on the ability to produce elaborated responses. Brown took this ability as interactional competence, as it can be regarded as the ability to sustain conversations. Although her study did not report on the frequency of comments, interactional behaviors appeared to be an influential area in the assessment of oral performances, as found in other studies (Ducasse, 2010; Galaczi, 2008; May, 2011a).

2.4.2.3 Non-verbal behavior

Non-verbal behavior is accounted for in some L2 communicative models as part of strategic competence or sociolinguistic competence. Young (2002, 2008) also stressed the importance of non-verbal resources (e.g., gesture, facial expressions) in the discursive practices. Some empirical studies also have explored the role of test-takers’ non-verbal behavior in raters’ holistic judgments.

It has been found that rater assessments of oral performances were influenced not only by test-takers’ verbal performances but also by non-verbal paralinguistic features. Gullberg (1998) examined whether visual information on test-takers affects raters’ evaluation of overall oral proficiency. The evaluated performances were video-recorded and audio-recorded French and Swedish data. Each of 20 linguistic laypersons was provided with either videos or audios and asked to assess linguistic management and narrative skills. All the performances were rated both with and without visual information, and the scores under the two conditions were statistically analyzed. The results showed that the majority of the test-takers received higher scores for overall
linguistic proficiency and narrative skills under the video condition, suggesting that the test-takers’ non-verbal behavior positively influenced the raters’ judgments.

Similar results were obtained by Nambiar and Goon (1993), who compared test scores given through the face-to-face format and those given through audio-recordings. In their study, two trained examiners rated fluency, accuracy, range, and effectiveness of L2 speakers’ oral performances. They rated the performances on the spot first and rescored the audio-recordings of the performances in a two-month interval. The results indicated that the mean score obtained through the face-to-face format were significantly higher. One of the possible reasons was the raters’ active participation in the interaction in the face-to-face format, which might distract their attention from linguistic limitations. It was also considered that non-vocal visual paralinguistic and extra-linguistic features positively affected rater evaluation or the lack of these features magnified the seriousness of linguistic limitations. The examiners’ notes suggested that dysfluency and errors in grammar and pronunciation were less noticeable when they rated performances in the face-to-face situation.

Some studies have also revealed that raters consciously attended to test-takers’ non-verbal behavior when they assessed oral performances. Orr (2002) analyzed verbal protocols provided by raters who assessed video-recorded performances on the Cambridge First Certificate in English (FCE) speaking test. Thirty-two official trained raters were asked to rate performances using the FCE rating scales, which are composed of (a) Grammar and vocabulary, (b) Discourse management, (c) Pronunciation, and (d) Interactive communication. The verbal protocols showed that the raters mentioned not only the four rating criteria but also 12 non-criterion features. Among the non-criterion features, the most frequently mentioned feature was the test-takers’ presentation of themselves, which entailed effort, body language, and preparedness for the test. Some
raters also mentioned eye contact while commenting on *Discourse management*. This indicated that the raters’ judgments of some criteria were influenced by test-takers’ non-verbal behaviors. Since the study did not provide the frequency of mentions of body language, it is unknown how strongly the raters attended to the feature. However, as Gullberg (1998) discovered, Orr’s study suggests that raters focus on a variety of features that are not stated in rating criteria and that body language may be one of them.

In a similar vein, Ducasse’s (2010) research using verbal protocol analysis revealed that raters attended to test-takers’ non-verbal behavior (17.4% of all comments) while rating peer-peer Spanish oral performances (see Section 2.4.2.2). The behaviors included hand gesture, gaze, laughter, body position, and facial expression. Since the raters were asked to watch video-recorded performances and comment on what made the interactions successful, this finding disclosed that raters’ judgments of interactional success were relatively strongly affected by test-takers’ non-verbal behavior exhibited during interaction. It suggests that raters might attend not only to linguistic forms but also to test-takers’ expressiveness, as stated in various L1 communicative competence models (Spitzberg & Hurt, 1987; Wiemann, 1977).

These studies, however, have not necessarily discussed the specific types of non-verbal behavior that were positively or negatively judged by raters. In fact, presence of some non-verbal behaviors does not guarantee the positive evaluation of raters. Neu (1990) examined gestures exhibited two test-takers (Yama and Ahmed) who received the remarkably similar scores for oral linguistic skills. Although Yama used more accurate and appropriate grammar and vocabulary than Ahmed did, his use of unintentional and complex gestures gave raters the impression that Yama had difficulties in speaking English. In contrast, Ahmed’s relaxed posture, head position, and simple gestures were considered to positively affect his test score though his accuracy was
poorer than Yama. This study suggests that more study should be done to explore the types of gestures that positively or negatively affect raters’ judgments. In fact, Gullberg (1998) examined what type of gesture was positively correlated with the raters’ judgments of overall linguistic proficiency and concluded that iconic gestures—gestures representing concrete objects and/or actions (McNeill, 2005)—possibly lead to positive assessments of proficiency. Although the visual information seemed to positively affect the raters’ judgments, the raters perceived that some behaviors did not enhance their comprehension of what test-takers had said.

2.4.3 Raters’ Judgments and Test-takers’ Level of Proficiency

Some empirical studies have found that the features contributing to raters’ judgments differ depending on the level of test-takers’ language proficiency. In other words, features that raters heed are not consistent across all language proficiency levels. Pollitt and Murray (1996) examined six TESOL specialists’ judgments by asking them to compare 10 pairs of video-taped oral performances on the Cambridge Certificate of Proficiency in English (CPE), concentrating on evaluating which performance of the pair is better. The raters then explained the perceived similarities and differences between each pair. In so doing, the researchers elicited what raters had paid attention to while making evaluative judgments. Their main finding was that the raters focused on different features at different proficiency levels (Figure 2.5). Pollitt and Murray found that the raters applied criteria of native-like behaviors, such as the content or elaboration of ideas, when comparing higher proficiency test-takers. In contrast, linguistic resources, comprehension, and pronunciation were only attended to when the raters compared pairs of lower proficiency test-takers. They argue that linguistic problems might become less obvious as proficiency grows; the rise made the raters begin to heed sophisticated
speech or idiomatic expressions as criteria. Their findings have important implications to what raters heed in relation to language proficiency.

Figure 2.5. Features heeded by raters according to test-takers’ proficiency (from Pollitt & Murray, 1996, p. 85).

Considering the major contribution of speech content to global scores found by some studies to be attributable to test-takers’ high English proficiency may be reasonable. For example, Hinofotis et al. (1981) investigated 10 TA applicants in an advanced oral communication course offered in the University of California, Los Angeles. Furthermore, Sato (2012) examined oral performances of relatively high-proficiency L2 learners: university students taking intermediate- and advanced-level general English courses. These studies revealed that the content (e.g., communication of information, elaboration, sophistication) affected raters’ judgments more than linguistic resources. In fact, research into raters’ judgments of writing performance also showed that raters tend to focus on the rhetoric and content of high-proficiency test-takers’ compositions (Cumming et al., 2001, 2002; Elder, Bright, & Bennett, 2007; O’Hagan, 2010; Weigle et al., 2003). In conclusion, raters’ focus shift
from language to content or ideas once test-takers exceed a certain threshold level of language proficiency.

The degree of rater attention to linguistic features also changes depending on speakers’ language proficiency. Examining test scores on a semi-direct speaking test, De Jong and Van Ginkel (1992) found that the relative contributions of linguistic-related criteria to the total scores differed depending on the raters’ global impression of performance. In their study, seven teachers of French rated test-takers’ oral performances on five tasks focusing on Pronunciation, Accuracy, Fluency, and Comprehensibility. Statistical analysis of the test scores revealed that Pronunciation most strongly contributed to the total score at lower levels of global impression, whereas the level of contribution of the four criteria became equal as impression scores rose. De Jong and Van Ginkel concluded that pronunciation was an elementary feature, and grammar did not contribute to the score in the absence of a certain extent of mastery in pronunciation.

The crucial role of pronunciation in low-proficiency levels was also pointed out by Iwashita et al. (2008), who investigated what features distinguished five proficiency levels of the TOEFL iBT speaking test (see Section 2.4.1). They found that some test-takers who exhibited poor pronunciation but wide-ranging vocabulary were judged as low-proficiency speakers. Iwashita et al. assumed that pronunciation was a decisive factor in raters’ judgment at low proficiency levels. This explanation seems reasonable since poor pronunciation is likely to hinder the judgment of syntactic or lexical knowledge; in this sense, pronunciation is the “first level hurdle” (Iwashita et al., 2008, p. 44) before being judged as one step better than the initial proficiency level. Once test-takers exceed the first level hurdle and make intelligible utterances, raters might not strongly heed pronunciation but pay attention to other features. This finding indirectly
supports Pollitt and Murray’s (1996) argument that “it is pointless and even counter-productive to write, for example, descriptors for pronunciation at high levels of proficiency” (p. 89).

Similar to De Jong and Van Ginkel’s (1992) study, Higgs and Clifford (1982) investigated the relative contributions of five linguistic features—Vocabulary, Grammar, Pronunciation, Fluency, and Sociolinguistic knowledge—to different levels of global language proficiency (Levels 1 to 5). Initially, the researchers hypothesized that Vocabulary and Grammar would most strongly contribute to the global proficiency of Level-1 learners (the lowest level) and Level-3 learners (the middle level), respectively. In order to test these hypotheses, 50 language teachers were asked to indicate their perceived importance of each feature at five proficiency levels. It was found that Pronunciation made a larger contribution at Level 1 than expected (it was the second largest contribution next to Vocabulary), although its contribution became the lowest of all the features at Level 3 and above. This partially supports the findings of the studies reviewed previously (De Jong & Van Ginkel, 1992; Iwashita et al., 2008; Pollitt & Murray, 1996). In addition, the contributions of Fluency and Sociolinguistic knowledge at Level 3 were larger than the predicted model, suggesting that Vocabulary and Grammar are not the only necessary characteristics for intermediate learners. This is also in accord with Pollitt and Murray (1996), in which raters would attend to sociolinguistic competence of test-takers with a middle level of proficiency.

In contrast to the studies above, De Jong et al. (2012) found that the relative contributions of (a) linguistic knowledge, (b) processing skills, and (c) pronunciation to oral proficiency were consistently larger for high-proficiency students than low-proficiency students (see Section 2.2.3). In their study, L2 Dutch students were divided into the 40% highest and the 40% lowest performing groups based on raters’
judgments. The regression weight of each predictor variable (see Table 2.3) was compared between the two proficiency groups. The result indicated that the contributions of all the predictor variables were larger for high-proficiency test-takers. Expressed differently, there was no unique factor contributing to low-proficiency test-takers more strongly than to their counterparts. Their research thus did not confirm the results of previous studies (De Jong & Van Ginkel, 1992; Higgs & Clifford, 1982) that demonstrated that the contributions of some features (vocabulary and pronunciation) were uniquely larger in the performances of low-proficiency test-takers. De Jong et al. (2012) argue that the difference might be due to the different criterion for the dependent variable. In their research, the dependent variable was *Functional adequacy*, consisting of linguistic proficiency, whereas the previous studies (De Jong & Van Ginkel, 1992; Higgs & Clifford, 1982) investigated raters’ impression of speaker from a listener perspective. The issue of research design will be discussed in a subsequent section.

In summary, the research reviewed in this section (De Jong & Van Ginkel, 1992; Higgs & Clifford, 1982; Pollitt & Murray, 1996) demonstrated the following. First, the relative contributions of factors underlying oral proficiency are dependent on the level of proficiency. The same features may not necessarily affect overall impressions of performance to the same degree across different levels. Second, the contribution of pronunciation is large at lower proficiency levels but becomes minor at higher proficiency levels (Iwashita et al., 2008; Wilds, 1975). Intelligible oral performances may make raters attend to features other than pronunciation or accent; it seems that native-like pronunciation is not necessarily required for high global evaluations. Third, raters tend to heed test-takers’ problematic characteristics. That is why raters pay attention to linguistic features at low-proficiency levels and gradually
apply non-linguistic criteria as the test-takers’ proficiency level rises. Content appears to be one of the salient features while rating high-proficiency test-takers (Hinofotis et al., 1981; Pollitt & Murray, 1996; Sato, 2012).

2.4.4 Research Approaches to Rater Judgment

The previous sections have reviewed empirical studies investigating the features that contribute to raters’ holistic judgments of L2 speakers’ oral performance. As can be seen, raters’ holistic judgments have been investigated using different research approaches (see Table 2.12), and therefore, different aspects of raters’ judgment were revealed and discussed. The following sections review research approaches used to investigate raters’ judgments and consider what aspects of judgments can be analyzed through each approach.

Table 2.12
Research Approaches Employed by Previous Studies

<table>
<thead>
<tr>
<th>Research approach</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reports</td>
<td>Eckes (2009); Higgs &amp; Clifford (1982); Plough et al. (2010);</td>
</tr>
<tr>
<td>Holistic ratings</td>
<td>De Jong &amp; Van Ginkel (1992); Hinofotis et al. (1981); Iwashita &amp; Grove (2003); McNamara (1990); Sato (2012);</td>
</tr>
<tr>
<td>Verbal reports</td>
<td>Brown (2007); Brown et al. (2005); Ducasse (2009, 2010); Ducasse &amp; Brown (2009); Kim (2009a, 2009b); Orr (2002); Pollitt &amp; Murray (1996); May (2009, 2011a, 2011b); Sato (2012); Zhang (2010); Zhang &amp; Elder (2011);</td>
</tr>
<tr>
<td>Objective measures</td>
<td>De Jong et al. (2012); Galaczi (2008); Gan (2010); Iwashita et al. (2008); Plough et al. (2010);</td>
</tr>
</tbody>
</table>

2.4.4.1 Self-reports

Self-reports are for eliciting people’s thoughts or cognitive processes during task performance (Cohen, 1994). Research participants are asked to describe their
general approaches to something through interviews or questionnaire surveys (e.g., “I am a systematic learner when it comes to learning a second language” (Gass & Mackey, 2000, p. 12)). Three studies employed self-reporting to elicit raters’ personal beliefs about the importance of criteria when assessing test-takers’ oral performance. The participants were asked to indicate the importance of criteria or describe the features perceived to be important without viewing or listening to any performance. Thus, their reports are not dependent on any particular performance but simply reflect their general beliefs. Although this method is practical and easily applied to collect a large number of sample’s thoughts, reports should be interpreted with caution. Cohen (1994) argues that self-reports do not necessarily reflect what participants would actually do in real-life settings; “Questionnaire or interview items are more likely to elicit learners’ beliefs about what they do rather than what they actually do” (p. 680).

2.4.4.2 Relationship between holistic and analytic ratings

The criteria strongly influencing raters’ judgment have been investigated by examining the relationship between holistic and analytic ratings. Since both ratings are elicited subjectively (i.e., based on raters’ judgments), De Jong et al. (2012) named this approach the subjective-subjective approach. Holistic ratings are derived from either raters’ overall assessments based on holistic scales or raters’ intuitive impression of oral performance. Analytic ratings are awarded based on given analytic rating scales composed of multiple features such as grammatical accuracy, vocabulary, pronunciation, and so forth. The relationship has been examined by obtaining correlation or regression coefficients between the two types of rating. That is, holistic ratings and analytic ratings are entered as the dependent variable and independent variables, respectively. It has been thought that the analytic criterion most strongly predicting holistic ratings is
considered to be influencing raters’ overall evaluation of oral performance.

An advantage of this research method is that it is possible to reveal raters’ unconscious attention to criteria when judging oral performances. McNamara (1990) found that trained raters for the OET associated Resources of grammar and expression with Overall communicative effectiveness even though they were trained not to emphasize the accuracy in the test. He argues that:

At the conscious level, then, and in terms of the intention of the design of the test, grammatical accuracy was downplayed. Yet the IRT [Item Response Theory] analysis shows that in both sessions of the test perceptions of grammatical and lexical accuracy in fact played a crucial role in determining the candidate’s total score. (p. 63)

This unconscious attention could not have been elicited if other methods had been used.

In contrast, a major disadvantage of this method is that studies have mainly showed the relative contributions of pre-determined criteria rather than what raters truly paid attention to. For example, Sato (2012) explored that only 67% of raters’ intuitive judgment was explained by five analytic criteria used in the test. In other words, approximately one third of raters’ judgments were influenced by non-criterion features (see also Section 2.4.2.1). Even though studies may find that raters emphasize a certain criterion, they simply show that the criterion is weighted compared to other criteria; it does not necessarily mean that raters did not consider non-criterion features.

Furthermore, if holistic ratings are given based on a pre-determined holistic rating scale, raters’ judgment is likely to be affected by features stated in the rating scale rather than raters’ pure impressions of the performance. Thus, this line of study is particularly useful to confirm whether raters bring any construct-irrelevant factors to their ratings.
2.4.4.3 Verbal reports

Verbal report analysis is also known as self-revelation, self-observation (Cohen, 1994), or introspective methods (McKay, 2009), in which participants are required to provide oral or written reports on their thought processes while performing certain tasks. Many studies have collected verbal protocols using stimulated recall (Gass & Mackey, 2000). In the studies, raters were first asked to rate test-takers’ oral performance with or without rating criteria. Then, they reviewed the video recordings of the same performance, and paused the video whenever they found any features they wished to comment on. Rater participants were asked to comment on the feature that influenced their ratings, judgment of oral proficiency, or perceptions of the test-takers. Unlike the studies using stimulated recall, Pollitt and Murray (1996) employed a different method to explore what raters paid attention to (see also Section 2.4.3). They asked the raters to compare pairs of performances and subsequently describe the similarities and differences between the pair. Sato (2012) collected raters’ general written comments after they had finished rating all performances without asking them to review the performances. Although there are some variations in the methods, basically the studies have aimed to explore the features attended to by raters in the process of judging test-takers’ oral performance.

An advantage of verbal protocols is that they are likely to reflect what participants would actually do in real-life settings more accurately than self-reports since raters are asked to describe their thought processes immediately after the task (Cohen, 1994; McKay, 2009). Furthermore, this method is exploratory in nature; rater comments are not necessarily restricted to features stated in predetermined rating
criteria. In fact, studies have demonstrated that it is not uncommon for even trained and accredited raters to attend to non-criterion features while rating (Brown, 2007; May, 2011a; Orr, 2002).

However, this method is limited by the fact that the degree of contribution of features cannot be precisely explored. The frequency of rater comments sorted in each category does not necessarily reflect the strength of raters’ attention. Moreover, it is often argued that verbal reports cannot explore raters’ unconscious and automatic thought processes (Wigglesworth, 2005). Detailed methodological issues in relation to verbal protocols are further discussed in Chapter 4.

2.4.4.4 Relationships between raters’ judgments and objective measures of performances

Some studies have investigated the relationship between raters’ judgments and various features gauged by objective measures or discourse analyses. This approach is called the subject-object approach because test-takers’ performance is subjectively assessed by raters and analyzed by objective measures (De Jong et al., 2012). In some of these studies, test scores were derived from raters’ evaluations based on given rating criteria. Further, the same performances were analyzed using Conversation Analysis or discourse analysis to examine interactional patterns (Galaczi, 2008; Gan, 2010) or linguistic-related features (Iwashita et al., 2008; Plough et al., 2010). Objectively analyzed features were usually quantified and statistically analyzed to reveal the relationship between raters’ judgments and objective measures of performances. In contrast, De Jong et al. (2012) elicited linguistic knowledge from independent measures (e.g., vocabulary tests, grammar tests) to avoid “the danger of circularity for a study that aims to define and decompose the construct of speaking proficiency” (p. 9).
An advantage of this approach is that it is possible to reveal raters’ unconscious attention to certain features. Furthermore, it is possible to analyze the role of specific linguistic features that cannot really be analyzed through raters’ judgment, such as grammatical complexity, vocabulary tokens, and number of filled pauses (see Iwashita et al., 2008). In this way, the analyzed features are not derived from raters’ judgments and are often restricted to linguistic-related or interactional features. Additionally, this line of study does not provide direct evidence regarding rater perception of observed features of performances. This approach thus cannot explore the relative contribution of non-criterion features and may not accurately reflect what raters actually attend to while making an evaluative judgment.

In conclusion, each approach can investigate limited aspects of evaluative judgment: (a) raters’ beliefs about rating behaviors, (b) the relative degree of contributions of criteria to overall judgment, (c) the unconscious orientations to pre-determined criteria, (d) non-criterion features raters consciously attend to while rating, and (e) specific linguistic-related features that affect their judgments but cannot be assessed subjectively. It is necessary to select or combine approaches in order to explore the aspect of raters’ judgments the researchers intend to investigate.

2.4.5 Summary

This section has reviewed empirical studies on how raters of speaking tests assess test-takers’ performance. Assessment criteria affecting their evaluative judgment were inconclusive; they include linguistic resources, fluency, pronunciation, speech content, interactional resources, mutual achievement, and non-verbal behavior. Plausible reasons behind the different findings are that the studies were conducted using (a) test-takers with different proficiency levels, (b) tests for different purposes, (c) different
elicitation tasks, and (d) different research approaches. As reviewed above, these facets are likely to affect the results reflecting the nature of raters’ judgments. Most studies commonly elicited evaluative judgments from language teachers and/or trained raters because the studies were conducted for the purpose of validation of particular speaking tests. As such, similar to L2 communicative theory development, only specialized views on communication are reflected in the research findings.

2.5 Research Question

Given the importance of perspectives of linguistic laypersons and a dearth of investigation, this study addresses the following research question:

What features of oral performance or behaviors of speakers affect linguistic laypersons’ impressions of communication ability?

It should be noted that exploring the components of English proficiency is NOT the aim of this study. Nor does the study aim to investigate the validity of assessment criteria used by language tests by comparing professional raters’ judgments and laypersons’ judgments. The aim of this study is to explore the constituents of linguistic laypersons’ perspectives on communication ability and to define communication ability from the perspective of linguistic laypersons. Revealing the strengths of influence of each feature or behavior is not considered as the primary aim. However, it is intended to explore the interrelationship among the features or behaviors derived from laypersons.

2.6 Summary

This chapter has provided justifications for investigating linguistic laypersons’
perspectives on L2 communication ability. Limitations of existing theories have been discussed by reviewing the details of their construction. It has been argued that the processes of L2 theory construction have been based mostly on either (a) reactions to or elaborations of previous theories or (b) empirical data from L2 speakers’ performance on language tests. Even though linguistic laypersons are most likely to be the ultimate arbiters in real-life contexts, theory construction has not included their perspectives. Furthermore, it has been reported that the perspectives of linguistic laypersons are different from those of language specialists whose main concern is with language proficiency. In the globalized world, English is used as a lingua franca, which makes investigation of NNESs’ perspectives important. Although the importance of linguistic laypersons’ perspectives on communication is acknowledged as such, previous studies have rarely paid attention to them, and studies on factors affecting linguistic laypersons’ judgments of communication ability have been scarce. Therefore, the research question, “What features of oral performance or behaviors of speakers affect linguistic laypersons’ impressions of communication ability?” guided this investigation.
Chapter 3 Methodology

3.1 Introduction

This chapter introduces the methodological approach underpinning the present study. The first section discusses an overview of the methodology and justifies the methodology adopted in the study. It also considers advantages and caveats of a data collection method, verbal protocol analysis. The subsequent section introduces the two oral proficiency tests used to elicit the linguistic laypersons’ judgment of communication ability. Following this, the specific data collection process and data analysis procedures are described.

3.2 Methodological Considerations

This study aimed to investigate linguistic laypersons’ subjective perspectives on L2 communication ability; more specifically, it was intended to explore the features of oral communication influencing their judgments. To answer the research question addressed in the study, a qualitative research approach was adopted.

One of the fundamental purposes of the qualitative approach is exploring unknown phenomena and generating new theories or hypotheses (Miles & Huberman, 1994). Corbin and Strauss (2008) state that one of the most important reasons for doing qualitative research is “to step beyond the known and enter into the world of participants, to see the world from their perspective and in doing so make discoveries that will contribute to the development of empirical knowledge” (p. 16). Therefore, researchers are required to interpret participants’ actions and words collecting textual data, acknowledging that knower and known are inseparable (Teddlie & Tashakkori, 2009). These characteristics of the qualitative research approach are suitable for this
study.

The worldview or paradigm that qualitative researchers adopt is constructivism, which acknowledges that the reality is multiple and socially constructed rather than single and tangible (Teddlie & Tashakkori, 2009). With regard to this, Corbin and Strauss (2008) explain further as follows:

The world is very complex. There are no simple explanations for things. Rather, events are the result of multiple factors coming together and interacting in complex and often unanticipated ways. … We believe that it is important to capture as much of this complexity in our research as possible, at the same time knowing that capturing it all is virtually impossible. (p. 8)

Accordingly, researchers are concerned with transferring the inference obtained from the research to other similar settings (i.e., transferability) rather than making time- and context-free generalizations. In this study, the data were collected through verbal protocol analysis.

3.2.1 An Overview of Verbal Protocol Analysis

Verbal protocol analysis (VPA) involves an analysis of participants’ verbal reports (or verbal protocols) on cognitive processes of their task performance. Green (1998) defined VPA as “a methodology which is based on the assertion that an individual’s verbalisations may be seen to be an accurate record of information that is (or has been) attended to as a particular task is (or has been) carried out” (pp. 1-2). It has been widely applied in the field of cognitive psychology underpinning the Information Processing (IP) theory of cognition (Ericsson & Simon, 1993). An important assumption
of the theory is that

information recently acquired (attended to or heeded) by the central processor is kept in STM [short-term memory], and is directly accessible for further processing (e.g., for producing verbal reports), whereas information from LTM [long-term memory] must first be retrieved (transferred to STM) before it can be reported. (Ericsson & Simon, 1993, p. 11)

This assumption suggests that it is possible to extract what has been stored in short-term memory through verbal reporting. The cognitive process is also defined as heeded information or thoughts. Therefore, according to IP theory, verbal protocols show a subset of the information that informants recently minded while completing tasks.

In order to ensure that verbal reports reflect cognitive processes of a task performance, Ericsson and Simon (1993) underscored that informants should not be asked to explain or justify their thoughts. They label the type of verbal protocols requiring explanation as Type 3 Verbalization and argue that such protocols alter the processes of task performance. Therefore, informants are instructed to simply verbalize what they are or were thinking while engaging in the task rather than explaining why they are or were showing certain behaviors. This point is quite important in order to directly access cognitive processes stored in short-term memory. In order not to influence informants’ cognitive processes, it is advised that researchers should minimize interactions with participants during verbal reporting (Green, 1998; McKay, 2009).

VPA is divided into different types according to how verbal protocols are elicited (Færch & Kasper, 1987; Gass & Mackey, 2000). Participants are instructed to express what they are thinking during the task (concurrent verbalization) or what they
were thinking immediately after the task (retrospective verbalization). Concurrent verbalization (or think-aloud) is considered to most accurately reflect participants’ cognitive processes and recommended whenever possible (Green, 1998). Retrospective verbalization can also access heeded information because it still remains in short-term memory after the task is completed (Ericsson & Simon, 1993). In some cases of retrospective verbalization, participants are asked to watch videos of their own performance to recall their cognitive processes. This version of VPA is referred to as stimulated recall (Gass & Mackey, 2000).

Second language research has widely applied VPA for investigating learners’ cognitive processes during their language use (see Gass and Mackey (2000, pp. 29-35)). Language testing researchers also have incorporated VPA into validation studies. For example, empirical studies have attempted to reveal raters’ rating process of compositions (Barkaoui, 2010; Cumming et al., 2001, 2002; Lumley, 2002, 2005; O’Hagan, 2010; Sakyi, 2000; Weigle, 1994). VPA has been also incorporated into studies on raters’ judgments of oral performances (Brown, 2000, 2007; Brown et al., 2005; Ducasse, 2009; Ducasse & Brown, 2009; Hubbard, Gilbert, & Pidcock, 2006; Kim, 2009a, 2009b; May, 2009, 2011a, 2011b; Orr, 2002; Zhang, 2010; Zhang & Elder, 2011). In these studies, think-aloud protocols were not employed for practical reasons (except for the study conducted by Hubbard et al. (2006)). Instead, raters were asked to engage in retrospective verbalization and/or stimulated recall.

Although VPA has been widely applied in various fields as a useful data collection method, the existing literature has discussed its limitations and disadvantages. Cohen (1994, pp. 680-681) summarized nine limitations of VPA argued in the previous research. One of the serious problems is that verbal protocols represent only a subset of cognitive processes or heeded information (Cumming et al., 2002; Lumley & Brown,
2005; McKay, 2009). In particular, automated processes are inaccessible by VPA since participants are unaware of the processes (Wigglesworth, 2005). Cohen (1994, 2011) also contends that a large part of cognitive processes is unconscious or too complex to capture in protocols.

More critically, researchers have challenged some fundamental assumptions of IP theory from the perspective of sociocultural theory. IP theory assumes that the information stored in short-term memory can be accessed through verbal reporting and that verbalization itself does not alter cognitive processes as long as researchers follow proper procedures for collecting verbal reports. In contrast, sociocultural theory posits that communicative activities mediate the development of higher mental processes and play a central role in mental development (Lantolf & Thorne, 2006). This perspective suggests that verbalization affects and alters participants’ cognitive processes (Smagorinsky, 1994; Swain, 2006). This seriously undermines the validity of verbal protocols because generated reports do not necessarily represent mental processes under the non-experimental condition. This problem is often observed in concurrent verbal protocols and referred to as reactivity.

Another crucial issue that IP theory does not take into account is the social role of speech. IP theory suggests that verbalization should be a monologue with no explanation or justification, and it is recommended that researchers should avoid interactions with participants during verbal reporting. Researchers usually sit behind participants, or participants are asked to verbalize without researchers present. However, sociocultural researchers believe that participants anticipate the audience and address them even though no one is present (Smagorinsky, 2001). In fact, some empirical studies have confirmed this social nature of verbal protocols (Barkaoui, 2011; Sasaki, 2003; Smagorinsky, 1997, 2001). The social and interactive nature of verbalization
leads to a threat of validity of verbal protocols referred to as *nonveridicality*. It refers to “the lack of correspondence between a protocol and the underlying primary process” (Gass & Mackey, 2000, pp. 107-108) and includes errors of omission and commission. The latter (reporting on what was not heeded) seriously undermine the validity of retrospective verbalization and stimulated recall. It is because, as sociocultural researchers posit, participants are likely to rationalize their protocols in an attempt to provide acceptable data to researchers (Cohen, 1994; Green, 1998; Greene & Higgins, 1994).

From the perspective of sociocultural theory, Smagorinsky (2001) and Swain (2006) negatively proposed the use of VPA as a method of eliciting participants’ cognitive processes. They suggested that it should be incorporated into research on how verbalization plays a role in mediating cognitive processes or language learning. Smagorinsky (2001) states that “I think … that they [protocols] would be used in service of different questions. In particular, the mediating role of speech itself could become an object of study” (p. 241). In addition, Swain (2006) warns that verbal protocols are not “brain dumps” (p. 110) that simply represent the information participants heed.

On the other hand, some researchers acknowledge the usefulness of the method in spite of its limitations (Lumley, 2005; Sasaki, 2003; Wigglesworth, 2005). VPA is the most viable method for accessing information that cannot be obtained by other methods (Cohen, 2011; McKay, 2009). As discussed above, it reflects what participants would actually do more accurately than other exploratory methods such as interviews and questionnaire surveys (Cohen, 1994). With regard to investigating rater’s cognitive processes, Lumley (2005) claims that “This methodology [VPA] is probably the only one presently available that can give us any insight, however hazy or indirect, into the
3.2.2 Verbal Protocol Analysis Procedures in the Present Study

The present study employed VPA as a method for investigating what contributes to linguistic laypersons’ impressions of communication ability. The main justification for the use of VPA is that it is the most viable option; as Greene and Higgins (1994) state, protocols “can at least offer a plausible explanation, providing more detail than we might obtain by simple speculation or by other methods alone” (p. 120). Part of linguistic laypersons’ evaluative judgments may be tapped by self-report, in which participants “provide descriptions of what they do, characterized by generalized statements” (Cohen, 2011, p. 80). For example, interviews or questionnaires are able to elicit participants’ general beliefs about what influences their judgments. However, these methods cannot investigate what actually influences participants’ evaluative judgments taking account of contextual factors that affect raters’ judgments, such as speakers’ English proficiency and the elicitation tasks used. Accordingly, VPA is suitable to elicit actual raters’ judgments of performances of various levels of test-takers engaging in different tasks.

Nevertheless, while conducting the study, limitations and threats to validity will be acknowledged. With regard to this point, Greene and Higgins (1994) suggest that researchers should consider how to collect and analyze verbal protocols “in a responsible way” and “reduce the possibility of misinterpretation” (p. 122). The procedures for data collection were meticulously designed following recommendations in the existing literature (Bowles, 2010; Cohen, 2011; Ericsson & Simon, 1993; Gass & Mackey, 2000; Green, 1998; Greene & Higgins, 1994; McKay, 2009).

In order to answer the research question, specific procedures were determined
referring to the classification scheme for VPA developed by Færch and Kasper (1987): namely, (a) relationship to specific action, (b) temporal relationship to action, (c) participant training, (d) procedural structure, (e) stimulus for recall, and (f) initiation of questions/recall interactions. Each procedure was determined based on previous studies and a pilot study that the researcher conducted.

First, participants were asked to verbalize their concrete and specific actions rather than non-specific and abstract actions. The present study was concerned with how linguistic laypersons judge communication ability of L2 speakers with different proficiency levels performing on different elicitation tasks. General reports on their judgments were not derived from VPA employed in the study.

The second consideration is the temporal distance between action and verbal reporting; that is, this is about whether verbalization should be concurrent, immediately retrospective, or delayed retrospective. This study incorporated retrospective verbalization because concurrent verbalization is not suitable for speaking or listening tasks (Wigglesworth, 2005). In fact, Hubbard et al. (2006) showed that experienced Cambridge CAE raters successfully provided concurrent verbal reports on their rating processes of speaking performances. However, given that rater participants in the present study are not accustomed to judging oral performance, the task of producing verbal protocols while judging communicative ability was deemed beyond the participants’ capacity.

The third point is with regard to participant training. It is claimed that training affects participants’ verbalization because researchers’ expectations may be conveyed through training (Brown, 2005; Gass & Mackey, 2000; Smagorinsky, 2001). This study conducted a brief training prior to verbalization to familiarize participants with the procedures. This was necessary for successful data collection because participants in the
present study were linguistic laypersons who had never formally evaluated L2 oral performances and provided verbal reports on their judgments. Accordingly, participants were asked to watch some sample performances, indicate their impressions, and generate verbal protocols. Feedback on the procedures was given if necessary during the practice session.

Fourth, with regard to the degree of procedural structure, this study employed a low-structure procedure that did not constrain the content of verbalization. Participants were thus allowed to verbalize anything that influenced their impressions of the speakers.

Fifth, the decision about whether or not participants would be given recall supports or stimuli was made based on the pilot study and previous studies. In the pilot study, four linguistic laypersons were asked to rate oral performances and verbalize what influenced their judgments either with or without recall supports. When the two versions of verbalization were compared, stimulated recall elicited a far larger amount of comments from the participants than verbal protocol without recall supports did. Moreover, the participants mentioned specific features of the performances in stimulated recall. These results suggest that stimulated recall is suitable to elicit the detailed features that influenced linguistic laypersons’ judgments. At the same time, the issue of nonveridicality was evident in the stimulated recall; some participants used the present tense for expressing their thoughts and changed their ratings during the recall session. These phenomena could be minimized by giving participants an opportunity to orally summarize the reasons for their rating immediately after indicating their impressions of the performance. In so doing, it was expected that in recall sessions participants would comment on specific features relevant to the summary statements. This elicitation technique has been used by several studies on raters’ judgments of L2
oral performances (Brown, 2007; Brown et al., 2005; Ducasse, 2009; Ducasse & Brown, 2009; May, 2009, 2011a, 2011b; Zhang, 2010; Zhang & Elder, 2011). The present study thus employed both a retrospective verbal protocol (without stimuli) and stimulated recall (Figure 3.1).

![Figure 3.1. Verbal protocol procedure in this study.](image)

The sixth consideration is related to who initiates verbalization. Since this study is concerned with the features affecting participants’ judgments, these participants took the initiative to verbalize rather than the researcher soliciting specific information from them.

Additionally, an important point to be considered is the language of verbal protocols. It is advised that participants should be allowed to provide verbal protocols in their L1 (Bowles, 2010; Gass & Mackey, 2000). To examine if the use of L2 impedes NNESs’ verbalization, three high-proficiency NNES participants (6.5-7.5 in IELTS; Chinese, Japanese, and Persian) were asked to provide verbal protocols in English in the pilot study. All of them responded that they had not faced any serious problems in
commenting in English, which indicates that verbalizing in English may not be a serious problem when research participants are high-proficiency English speakers. Given this result, the present study asked participants to provide verbal protocols in English. This decision was also related to a logistic reason. It was considered extremely difficult to translate various L1s for the data analysis and check the content of participants’ protocols during the data collection.

3.3 The CET-SET and the Cambridge English Examinations

This section introduces the two tests used to elicit linguistic laypersons’ evaluative judgments in this study: the College English Test-Spoken English Test (CET-SET) and the suite of Cambridge English Examinations. It also describes the parts of the tests used for data collection and their characteristics. The data were provided by the National College English Committee and Cambridge English Language Assessment.

3.3.1 CET-SET

The CET-SET is a spoken component of the College English Test, which aims “to examine the English proficiency of undergraduate students in China and ensure that Chinese undergraduates reach the required English levels specified in the National College English Teaching Syllabuses” (Zheng & Cheng, 2008, p. 408). The test is a general-purpose English proficiency test whose standard is set based on the national syllabuses. The test score is exploited as a prerequisite for bachelor’s degree in many Chinese universities (He & Dai, 2006). Therefore, the CET serves as an achievement test for university English education in China. At the same time, its scores can be used for employment in government organizations, foreign companies, or highly competitive positions (He & Dai, 2006; Zhang & Elder, 2009). Test-takers are Chinese
undergraduate students majoring in non-English disciplines. The CET is composed of four subsections: (a) listening comprehension, (b) reading comprehension, (c) cloze or error correction, and (d) writing and translation. The CET-SET is an independent subtest, and only test-takers with a sufficient score on the four subsections of the CET can sit the test. In other words, achieving a high score on the CET is a prerequisite for the speaking test; thus, CET-SET test-takers are considered intermediate to advanced L2 English learners in general.

The construct of the CET-SET is “the oral English communication ability of college and university students in China based on the National College English Teaching Syllabus” (Zhang & Elder, 2009, p. 298). This construct is specified in the analytic rating criteria, which consist of six components (Table 3.1). The criteria are weighted according to the importance determined by the test developer. The total score is obtained using the following formula: total score = raw scores for **Accuracy** and **Range** × 1.2 + raw scores for **Size** and **Discourse management** × 1.0 + raw scores for **Flexibility** and **Appropriacy** × 0.8. In other words, the CET-SET emphasizes lexico-grammatical accuracy and range. The total score is converted into one of seven levels (A⁺, A, B⁺, B, C⁺, C, and D) and reported to the test-takers. Descriptors of the levels (A, B, and C) are shown in Table 3.2.
Table 3.1
*The CET-SET Assessment Criteria (Zhang & Elder, 2009, p. 303)*

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Accuracy in pronunciation, stress/intonation and use of grammar and vocabulary</td>
</tr>
<tr>
<td>Range</td>
<td>Range of vocabulary and grammatical structures</td>
</tr>
<tr>
<td>Size</td>
<td>Size of contribution made by the candidate</td>
</tr>
<tr>
<td>Discourse management</td>
<td>Discourse management, the ability to produce extended and coherent discourse</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Flexibility in dealing with different situations and topics</td>
</tr>
<tr>
<td>Appropriacy</td>
<td>Appropriacy in the use of linguistic resources according to context</td>
</tr>
</tbody>
</table>

Table 3.2
*Band Grade and Descriptors (Zhang & Elder, 2009, p. 304)*

<table>
<thead>
<tr>
<th>Grade</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Able to make oral communication in English on familiar topics without difficulties</td>
</tr>
<tr>
<td>B</td>
<td>Able to make oral communication in English on familiar topics with some difficulties that do not affect communication</td>
</tr>
<tr>
<td>C</td>
<td>Able to make simple oral communication in English on familiar topics</td>
</tr>
</tbody>
</table>

The CET-SET employs a face-to-face group format, in which three test-takers sit in a group. After a brief warm-up, they are required to give an individual presentation on a given topic (1.5 minutes), discuss the topic in a group (4.5 minutes), and individually answer the questions asked by an examiner (Table 3.3). In Part 1, test-takers introduce themselves and answer some warm-up questions on a topic throughout the entire test. In Part 2, each test-taker is asked to give a 1.5-minute individual presentation on a given prompt (e.g., city traffic in China) after one-minute preparation. Other test-takers are given different prompts of the same line of topics and asked to listen to the presentation. Following all test-takers’ presentations, they are asked to discuss the topic for individual presentations for 4.5 minutes. An examiner provides a specific discussion topic and instructions to them. In Part 3, each test-taker is
asked further questions regarding the discussion topic.

Table 3.3

<table>
<thead>
<tr>
<th>Part</th>
<th>Time</th>
<th>Participants</th>
<th>Test Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.0 min</td>
<td>Examiner-Candidate</td>
<td>Verbal questions to each candidate</td>
</tr>
<tr>
<td>2</td>
<td>5.5 min</td>
<td>Examiner-Candidate</td>
<td>Visual stimulus</td>
</tr>
<tr>
<td>4.5 min</td>
<td>Candidate-Candidate</td>
<td>Group discussion</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5.0 min</td>
<td>Examiner-Candidate</td>
<td>Verbal questions to each candidate</td>
</tr>
</tbody>
</table>

Two accredited examiners rate the test-takers on the spot; one of them asks questions, provides instructions, and rates test-takers simultaneously, whereas the other examiner concentrates on rating without interacting with test-takers. The raters of the test are officially trained Chinese EFL university teachers.

The present study used individual presentations elicited from Part 2. The following shows the task characteristics of Part 2 using Fulcher’s (2003) framework (Table 3.4). The task orientation is open, as test-takers are allowed to develop the theme on their own although the topic is provided. Since the type of discourse produced is an extended monologue, this task is considered non-interactional. The presentations are made to examiners and peer test-takers of higher and equal statuses to the test-taker and who are strangers to the test-taker. Topics are related to everyday concerns and familiar to university students in China, and thus specialized topical knowledge is not required for the presentation. Nevertheless, in addition to language proficiency, general cognitive abilities such as logical thinking ability or organizing skills are necessary in the task. The setting of the task is a formal interview in which test-takers are seated facing two examiners with a table between them.
Table 3.4  
The Characteristics of Individual Presentations on the CET-SET

<table>
<thead>
<tr>
<th>Task features</th>
<th>Individual presentations on the CET-SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task orientation</td>
<td>Open</td>
</tr>
<tr>
<td>Interactional relationship</td>
<td>Non-interactional</td>
</tr>
<tr>
<td>Goal orientation</td>
<td>None</td>
</tr>
<tr>
<td>Interlocutor status &amp; familiarity</td>
<td>No interlocutor; audience are higher or equal status and unfamiliar</td>
</tr>
<tr>
<td>Topic(s)</td>
<td>General</td>
</tr>
<tr>
<td>Situations</td>
<td>Formal interview</td>
</tr>
</tbody>
</table>

Although the CET-SET contains interactions between an examiner and test-takers (Part 1) as well as a group discussion among three test-takers (Part 3), it was decided not to use them for the present study due to a lack of natural interactions.

Interactions between an examiner and test-takers (Parts 1 and 3) are formal interviews in which test-takers are only required to answer questions asked by the examiner. Moreover, as far as the researcher examined the pool of group discussions (the latter half of Part 2), most of the discussions appeared to include parallel interaction patterns or solo versus solo interaction (Galaczi, 2008). Test-takers had exhibited their own speaking performances with minimal acknowledgement of what other test-takers had said. In fact, the lack of interactional language functions in the CET-SET has been pointed out by He and Dai (2006) previously.

3.3.2 Cambridge English Examinations

The suite of Cambridge English Examinations (hereafter, Cambridge Exams) includes five English proficiency tests based on different levels of English proficiency (Table 3.5). Each test is set at one proficiency level defined by the CEFR (Council of Europe, 2001): from C2 to A2 levels. Test-takers are supposed to choose a test they wish
to undertake according to their own English proficiency. The examinations cover all the four language skill areas: reading, writing, listening, and speaking.

Table 3.5

Cambridge English Examinations

<table>
<thead>
<tr>
<th>Test</th>
<th>CEFR Level</th>
<th>Purpose/Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certificate of Proficiency in English (CPE)</td>
<td>C2</td>
<td>Academic and professional</td>
</tr>
<tr>
<td>Certificate in Advanced English (CAE)</td>
<td>C1</td>
<td>Academic and professional</td>
</tr>
<tr>
<td>First Certificate in English (FCE)</td>
<td>B2</td>
<td>Academic and professional</td>
</tr>
<tr>
<td>Preliminary English Test (PET)</td>
<td>B1</td>
<td>General and for schools</td>
</tr>
<tr>
<td>Key English Test (KET)</td>
<td>A2</td>
<td>General and for schools</td>
</tr>
</tbody>
</table>

The test constructs are claimed to be “everyday English at an/a intermediate/basic level” for PET and KET (UCLES, 2009a, 2009b) and “language proficiency in terms of a language user’s overall communicative ability” for CPE, CAE, and FCE (UCLES, 2007, 2008, n.d.). Test results are provided to test-takers in the form of a certificate. If a test-taker passes the test, he or she receives a certificate of a specific CEFR level. A large number of organizations accept the test certificate, such as universities, colleges, business organizations, and government departments. While certificates for PET and KET are mainly used for general purposes, those for CPE, CAE, and FCE are widely accepted by many employers, ministries, higher educational institutions, or government bodies as evidence for English proficiency or communication ability. Since test scores are used for a variety of purposes, these tests can be considered as general-purpose English proficiency tests. The tests are administered all over the world; English learners of a variety of nationalities take the tests.

The present study used test-takers’ performances on the speaking sections in three intermediate tests—the CAE, FCE, and PET. These tests adopt similar assessment
criteria. FCE and PET employ the same five criteria: *Grammar and vocabulary*, *Discourse management*, *Pronunciation*, *Interactive communication*, and *Global assessment*. The CAE assessment criteria are the same except that grammar and vocabulary are divided into *Grammatical resource* and *Lexical resource* and are assessed independently (Table 3.6). Two examiners, an interlocutor and an assessor, assess test-takers’ oral performances on the spot. The interlocutor conducts the test giving instructions to test-takers and awards a mark for *Global assessment*. The assessor does not interact with test-takers and focuses solely on rating the other criteria.

<table>
<thead>
<tr>
<th>Table 3.6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CAE Criteria (UCLES, n.d.)</strong></td>
</tr>
<tr>
<td>Criterion</td>
</tr>
<tr>
<td>Grammatical resource</td>
</tr>
<tr>
<td>Lexical resource</td>
</tr>
<tr>
<td>Discourse management</td>
</tr>
<tr>
<td>Pronunciation</td>
</tr>
<tr>
<td>Interactive communication</td>
</tr>
<tr>
<td>Global achievement</td>
</tr>
</tbody>
</table>

Since the speaking section of the tests employs a peer-peer paired format, a pair of test-takers sits the test simultaneously. All the tests are comprised of four parts as shown in Table 3.7. These three tests commonly include (a) a conversation between test-takers and an interlocutor, (b) an individual test-taker’s picture description, and (c) a discussion about issues related to picture description. Elicitation tasks used in the tests
are similar. Thus, the tests can elicit both monologic and dialogic discourse types from each test-taker. The task types used in the CAE and FCE are identical.

Table 3.7
_Elicitation Tasks in Three Cambridge Exams_

<table>
<thead>
<tr>
<th>Part</th>
<th>CAE</th>
<th>FCE</th>
<th>PET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Candidate-examiner interactions (3 min.)</td>
<td>Candidate-examiner interactions (3 min.)</td>
<td>Candidate-examiner interactions (2-3 min.)</td>
</tr>
<tr>
<td>2</td>
<td>Individual picture description (3 min.)</td>
<td>Individual picture description (3 min.)</td>
<td>Candidate-candidate interaction in a simulated situation (2-3 min.)</td>
</tr>
<tr>
<td>3</td>
<td>Candidate-candidate collaborative conversation (4 min.)</td>
<td>Candidate-candidate collaborative conversation (3 min.)</td>
<td>Individual picture description (3 min.)</td>
</tr>
<tr>
<td>4</td>
<td>Discussion about the topic raised in Part 3 (4 min.)</td>
<td>Discussion about the topic raised in Part 3 (4 min.)</td>
<td>Candidate-candidate conversation about the topic raised in Part 3 (3 min.)</td>
</tr>
</tbody>
</table>

The present study used test-takers’ dialogic performances on Part 3 of the CAE and FCE as well as Part 4 of the PET. In the CAE and FCE, the paired task is called a _collaboration task_, in which test-takers are given a visual stimulus consisting of several pictures and instructed to discuss each picture (e.g., the particular role of computers that each picture illustrates). They are required to negotiate to make a decision related to the visual stimulus (e.g., the role of computers that affects our lives the most). The following instructions are given: “First, talk to each other about X. Then decide Y.” In this task, test-takers are expected to sustain an interaction, exchange ideas, justify their opinions, agree or disagree, suggest, speculate, evaluate, and reach a decision through negotiation (UCLES, 2007, n.d.). In the PET, two test-takers are asked to engage in a casual conversation on a theme introduced in a previous part (e.g., different kinds of reading or writing you do). The theme is used for a discussion about their preferences,
experiences, habits, and so forth. Test-takers are expected to discuss their interests and give reasons for their views.

The following are the characteristics of the elicitation task used for this study (Table 3.8). Since the task orientation is open, test-takers are allowed to include their own thoughts or opinions although a topic for conversation or discussion is provided. It is a two-way dialogic interaction between two test-takers. They have an equal status in the sense that they are both candidates for the same test. The goal orientation is convergent, as the test-takers are required to engage in the same theme and negotiate with each other in order to reach a decision. Nevertheless, they do not have to agree with their partner, in particular during the conversation in the PET. Given topics are general and do not require specialized topical knowledge. Finally, although the interaction occurs in a testing room in front of two examiners, the situation is relatively casual since test-takers face each other without the examiner’s intervention.

Table 3.8
The Characteristics of Paired Interaction on CAE, FCE, and PET

<table>
<thead>
<tr>
<th>Task features</th>
<th>Conversations on the Cambridge ESOL exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task orientation</td>
<td>Open</td>
</tr>
<tr>
<td>Interactional relationship</td>
<td>Two-way interaction</td>
</tr>
<tr>
<td>Goal orientation</td>
<td>Convergent, divergent</td>
</tr>
<tr>
<td>Interlocutor status &amp; familiarity</td>
<td>Equal status, unfamiliar</td>
</tr>
<tr>
<td>Topic(s)</td>
<td>General</td>
</tr>
<tr>
<td>Situations</td>
<td>Casual conversation and discussion</td>
</tr>
</tbody>
</table>

3.4 Methods

This section provides information about the research participants, the instruments for data collection, the data collection procedure, and the data analysis procedure.
3.4.1 Participants

Twenty-six research participants were involved in the study as raters. The following describes the sampling methods used and the characteristics of participants. Research participants were recruited using four sampling criteria. Table 3.9 shows the criteria for sampling. It should be noted that being a native speaker of English was not a criterion.

Table 3.9

<table>
<thead>
<tr>
<th>Sampling Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria</td>
</tr>
<tr>
<td>1. Post-graduate students studying/researching at the University of Melbourne</td>
</tr>
<tr>
<td>2. Students specializing in disciplines other than applied linguistics and TESOL</td>
</tr>
<tr>
<td>3. Those who have no language teacher and rater training experiences</td>
</tr>
<tr>
<td>4. Those who have no formal language teaching and rating experiences</td>
</tr>
</tbody>
</table>

First, the researcher approached post-graduate students at the University of Melbourne. One of the justifications was to ensure that NNES participants had high enough English proficiency to comprehend test-takers’ spoken English and verbalize their judgments of their communication ability; at the University of Melbourne, 6.5 or 7.0 on IELTS or 79 on the TOEFL iBT (equivalent to B2 in the CEFR) is the minimum requirement for NNES applicants for graduate courses. Another rationale for this criterion was that post-graduate students had sufficient cognitive ability and vocabulary to generate metacognitive verbal protocols (the procedures will be described below). In general, informants vary in terms of verbal skills, and some of them may not be able to provide useful verbal data (Cohen, 2011). Therefore, Lumley (2005) warned that participants should be selected carefully since “one requirement is that they can articulate their thoughts at a metacognitive level” (p. 74) [emphasis in original].
The other three criteria were those for recruiting linguistic laypersons. As mentioned in Chapter 1, for the purpose of this study, *linguistic laypersons* were defined as those who do not have (a) specialized knowledge of applied linguistics, (b) experience of any training in language assessment and teaching, and (c) experience of rating and teaching L2 learners formally. As such, this study did not include applied linguistics researchers, students studying applied linguistics or TESOL, or professional language teachers or raters with formal training. To ensure that participants had none of the characteristics above, the researcher approached students specializing in disciplines other than applied linguistics and TESOL. In addition, this study included only those who had no language teacher training and rater training experience. Finally, those who had formal and extensive language teaching and rating experiences were not selected as research participants. However, it was difficult to find English-speaking participants who had never engaged in any kind of language teaching (e.g., as a private tutor). Therefore, it was decided not to exclude those who had language teaching experience for a short period of time (less than one year) as private tutors or instructors without a teacher license and training.

The recruits who met all the criteria above represented all of Kachru’s (1988) concentric circles: the Inner Circle, the Outer Circle, and the Expanding Circle. This is because the study intended to investigate the evaluative judgments made by people with various language backgrounds and English-learning experiences, including both NNESs and NESs.

Following the sampling procedures described above, 26 participants (raters hereafter) were recruited. However, among them, three raters were excluded from the study due to problems in the data collection procedures; thus, 23 raters were included in the study in total.
The background information of the 23 raters is given in Table 3.10. There were 15 female and 8 male raters. All of them were undertaking Master’s or PhD level courses in disciplines other than TESOL or applied linguistics. Rater S had engaged with English language teaching in France as a teaching assistant for approximately seven months. Nevertheless, she was included in the study because she had no previous language teacher training and her teaching experience was short and casual. No raters had undertaken any subjects on language teaching and applied linguistics, which confirmed their unfamiliarity with these fields. There were 10 raters from the Expanding Circle (Raters A to J: *EFL raters* hereafter), six raters from the Outer Circle (Raters K to P: *ESL raters* hereafter), and seven raters from the Inner Circle (Raters Q to W: *ENL raters* hereafter). Among those from the Expanding Circle, there were three raters (Raters D, F, and G) who had learned English from the early childhood through immersion. Thus, these raters are considered and labeled as ESL raters.
Table 3.10
*Rater Background Information (N=23)*

<table>
<thead>
<tr>
<th>Rater</th>
<th>F/M</th>
<th>Age</th>
<th>Nationality</th>
<th>L1</th>
<th>Course</th>
<th>Discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>M</td>
<td>26-30</td>
<td>Iranian</td>
<td>Farsi</td>
<td>PhD</td>
<td>Architecture</td>
</tr>
<tr>
<td>B</td>
<td>F</td>
<td>21-25</td>
<td>Vietnamese</td>
<td>Vietnamese</td>
<td>MA</td>
<td>Commercial law</td>
</tr>
<tr>
<td>C</td>
<td>F</td>
<td>21-25</td>
<td>Italian/German</td>
<td>Italian/German</td>
<td>MA</td>
<td>Engineering</td>
</tr>
<tr>
<td>D</td>
<td>M</td>
<td>31-35</td>
<td>Nepalese</td>
<td>Nepalese</td>
<td>PhD</td>
<td>Architecture</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>21-25</td>
<td>Japanese</td>
<td>Japanese</td>
<td>MA</td>
<td>International law</td>
</tr>
<tr>
<td>F</td>
<td>M</td>
<td>21-25</td>
<td>French</td>
<td>French</td>
<td>MA</td>
<td>Marketing</td>
</tr>
<tr>
<td>G</td>
<td>M</td>
<td>26-30</td>
<td>Spanish</td>
<td>Spanish</td>
<td>MA</td>
<td>Business administration</td>
</tr>
<tr>
<td>H</td>
<td>F</td>
<td>21-25</td>
<td>Chinese</td>
<td>Mandarin</td>
<td>MA</td>
<td>Marketing</td>
</tr>
<tr>
<td>I</td>
<td>F</td>
<td>36-40</td>
<td>Indonesian</td>
<td>Bahasa</td>
<td>MA</td>
<td>Social sciences</td>
</tr>
<tr>
<td>J</td>
<td>M</td>
<td>31-35</td>
<td>Vietnamese</td>
<td>Vietnamese</td>
<td>MA</td>
<td>International relations</td>
</tr>
<tr>
<td>K</td>
<td>F</td>
<td>21-25</td>
<td>Filipino</td>
<td>Filipino</td>
<td>MA</td>
<td>Social sciences</td>
</tr>
<tr>
<td>L</td>
<td>F</td>
<td>21-25</td>
<td>Kenyan</td>
<td>Kikuyu/Kiswahili</td>
<td>MA</td>
<td>Social sciences</td>
</tr>
<tr>
<td>M</td>
<td>F</td>
<td>36-40</td>
<td>Singaporean</td>
<td>Mandarin</td>
<td>MA</td>
<td>Education</td>
</tr>
<tr>
<td>N</td>
<td>F</td>
<td>26-30</td>
<td>Indian</td>
<td>English</td>
<td>MA</td>
<td>Social sciences</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>21-25</td>
<td>Malaysian</td>
<td>Mandarin</td>
<td>MA</td>
<td>Business administration</td>
</tr>
<tr>
<td>P</td>
<td>F</td>
<td>21-25</td>
<td>Bangladeshi</td>
<td>Bengali</td>
<td>MA</td>
<td>Economics</td>
</tr>
<tr>
<td>Q</td>
<td>F</td>
<td>21-25</td>
<td>Australian</td>
<td>English</td>
<td>MA</td>
<td>Social work</td>
</tr>
<tr>
<td>R</td>
<td>F</td>
<td>21-25</td>
<td>Australian</td>
<td>English</td>
<td>MA</td>
<td>Arts management</td>
</tr>
<tr>
<td>S</td>
<td>F</td>
<td>21-25</td>
<td>British</td>
<td>English</td>
<td>PhD</td>
<td>French literature</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>21-25</td>
<td>Australian</td>
<td>English</td>
<td>MA</td>
<td>Medicine</td>
</tr>
<tr>
<td>U</td>
<td>F</td>
<td>41-</td>
<td>New Zealander</td>
<td>English</td>
<td>MA</td>
<td>Physiotherapy</td>
</tr>
<tr>
<td>V</td>
<td>M</td>
<td>21-25</td>
<td>Canadian</td>
<td>English</td>
<td>MA</td>
<td>Dentistry</td>
</tr>
<tr>
<td>W</td>
<td>M</td>
<td>26-30</td>
<td>American</td>
<td>English</td>
<td>MA</td>
<td>Finance</td>
</tr>
</tbody>
</table>

Table 3.11 shows raters’ foreign language learning experience and perceived familiarity with NNEs’ spoken English. Eighteen raters had learned one or two foreign language(s) (excluding English). It should be noted that EFL and ESL raters (Raters A to P) learned English as a foreign or second language. Although many of them had learned foreign languages as a school subject and acknowledged that they were not necessarily fluent in those languages, some of them had more than 10-year learning experience and used the language for communication or their studies. Their foreign
language learning experience suggests that most raters had general meta-linguistic knowledge and were able to analyze language to some degree.

Table 3.11

<table>
<thead>
<tr>
<th>Rater</th>
<th>FL1</th>
<th>Length (FL1)</th>
<th>FL2</th>
<th>Length (FL2)</th>
<th>Familiarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Arabic</td>
<td>5 years</td>
<td>N/A</td>
<td>N/A</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>French</td>
<td>8 years</td>
<td>Russian</td>
<td>2 years</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>Mandarin</td>
<td>12 years</td>
<td>Hindi</td>
<td>2 years</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>Chinese</td>
<td>3 years</td>
<td>Spanish</td>
<td>1 years</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>German</td>
<td>10 years</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>French</td>
<td>4 years</td>
<td>N/A</td>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>H</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>Japanese</td>
<td>1 year</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>J</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>3</td>
</tr>
<tr>
<td>K</td>
<td>Spanish</td>
<td>2 weeks</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>L</td>
<td>French</td>
<td>5 years</td>
<td>N/A</td>
<td>N/A</td>
<td>3</td>
</tr>
<tr>
<td>M</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>French</td>
<td>4 years</td>
<td>Hindi</td>
<td>5 years</td>
<td>5</td>
</tr>
<tr>
<td>O</td>
<td>Cantonese</td>
<td>5 years</td>
<td>Malay</td>
<td>12 years</td>
<td>4</td>
</tr>
<tr>
<td>P</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>Q</td>
<td>German</td>
<td>2 years</td>
<td>Japanese</td>
<td>2 years</td>
<td>3</td>
</tr>
<tr>
<td>R</td>
<td>German</td>
<td>11 years</td>
<td>French</td>
<td>11 years</td>
<td>4</td>
</tr>
<tr>
<td>S</td>
<td>French</td>
<td>13 years</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>T</td>
<td>German</td>
<td>1 month</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>U</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>3</td>
</tr>
<tr>
<td>V</td>
<td>French</td>
<td>10 years</td>
<td>N/A</td>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>W</td>
<td>Spanish</td>
<td>4 years</td>
<td>N/A</td>
<td>N/A</td>
<td>4</td>
</tr>
</tbody>
</table>

Note. 1=Not familiar; 5=Very familiar.

Table 3.11 also indicates the degree to which the raters perceived their familiarity with English spoken by NNESs in general. Most raters marked 3 or above, indicating that they were at least moderately familiar with NNESs’ English. Although
Rater M responded 1 (= not familiar), she stated that she was particularly familiar with English spoken by Singaporean, Malay, and Filipino speakers. Their level of familiarity suggests that they had opportunities to communicate with NNESs in English and hear their accents.

The EFL and ESL raters had a wide range of English proficiency and distinct English learning experiences. Table 3.12 illustrates the length of time spent in Australia, their English proficiency test scores, the age when they started learning English, and the place they spent most of time learning English. The levels of English proficiency ranged from IELTS 6.5 to 8.5, which is approximately equivalent to the CEFR B2 to C2 levels. Raters M and O, who were from Singapore and Malaysia, were not required to submit a proficiency test score for their admission to graduate school. The raters’ scores on English proficiency tests suggest that they were independent to highly proficient English users in terms of the CEFR scale (Council of Europe, 2001). With regard to English learning experience, ESL raters had started to learn the language from the early age (from three to six) and used it as a means to communicate at school. In addition, English had been the primary language at home for Raters K, L, and N. English learning experience of the raters from the Expanding Circle was widely different. Raters D, F, and G had learned English through immersion at English-medium schools, which indicates that their learning environment was in that sense similar to that of the ESL raters. The rest of the raters (Raters A, B, C, E, H, I, and J) had studied English as one of the school subjects in their home countries. Their onset ages were relatively higher than those of the ESL raters. Nonetheless, Rater C had been in Britain for one year when she was four years old and started to learn the language there.

---

3 The TOEFL and TOEIC scores of Raters A, C, F, and G were approximately equivalent to IELTS 7.5, 8.0, 7.5, and 8.0, respectively.
### Table 3.12

**Raters’ English Proficiency and English Learning Experience**

<table>
<thead>
<tr>
<th>Rater</th>
<th>Yrs in AU</th>
<th>Test score</th>
<th>Age started learning English</th>
<th>Places learning English</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.5 yrs</td>
<td>TOEFL PBT 617</td>
<td>12</td>
<td>school, language institute</td>
</tr>
<tr>
<td>B</td>
<td>6 mths</td>
<td>IELTS 6.5</td>
<td>15</td>
<td>school, home</td>
</tr>
<tr>
<td>C</td>
<td>3 mths</td>
<td>TOEFL iBT 113</td>
<td>4</td>
<td>school, Britain (1 yr)</td>
</tr>
<tr>
<td>D</td>
<td>4 yrs</td>
<td>IELTS 8.0</td>
<td>9</td>
<td>school</td>
</tr>
<tr>
<td>E</td>
<td>6 mths</td>
<td>IELTS 8.0</td>
<td>13</td>
<td>school</td>
</tr>
<tr>
<td>F</td>
<td>3 mths</td>
<td>TOEFL iBT 106</td>
<td>0</td>
<td>school, home</td>
</tr>
<tr>
<td>G</td>
<td>3 mths</td>
<td>TOEIC 985</td>
<td>2</td>
<td>school, academy</td>
</tr>
<tr>
<td>H</td>
<td>7 mths</td>
<td>IELTS 6.5</td>
<td>9</td>
<td>school</td>
</tr>
<tr>
<td>I</td>
<td>1 mth</td>
<td>IELTS 7.0</td>
<td>18</td>
<td>school, work</td>
</tr>
<tr>
<td>J</td>
<td>1 mth</td>
<td>IELTS 7.0</td>
<td>15</td>
<td>school</td>
</tr>
<tr>
<td>K</td>
<td>1 yr</td>
<td>IELTS 8.0</td>
<td>4</td>
<td>school, home</td>
</tr>
<tr>
<td>L</td>
<td>3.5 mths</td>
<td>IELTS 7.5</td>
<td>6</td>
<td>school, home</td>
</tr>
<tr>
<td>M</td>
<td>1 yr</td>
<td>N/A</td>
<td>5</td>
<td>school</td>
</tr>
<tr>
<td>N</td>
<td>3 mths</td>
<td>IELTS 8.5</td>
<td>4</td>
<td>school, home</td>
</tr>
<tr>
<td>O</td>
<td>2.5 mths</td>
<td>N/A</td>
<td>5</td>
<td>school</td>
</tr>
<tr>
<td>P</td>
<td>8 mths</td>
<td>IELTS 8.5</td>
<td>3</td>
<td>school</td>
</tr>
</tbody>
</table>

#### 3.4.2 Instruments

The following instruments were developed and used in this study: (a) a plain language statement (b) a consent form, (c) a participant background questionnaire, (d) written instructions for verbal report procedures, (e) video-recorded oral performance data, (f) a rating form, and (g) a set of post-session semi-structured interview questions. The drafts were trialed with four linguistic laypersons and two PhD candidates in applied linguistics, and modified according to their feedback.

#### 3.4.2.1 Plain language statement

The plain language statement described the purpose of the study and procedures for data collection using non-technical terms (Appendix 1). It also stated that
participation is voluntary, that confidentiality is protected, and that participants may withdraw at any time without penalty. The contact information of the human ethics office at the university was provided.

3.4.2.2 Consent form

A consent form was prepared to meet research ethics requirements. The consent form contained the same information as the plain language statement. The research was approved by the Human Research Ethics Committee at the University of Melbourne.

3.4.2.3 Participant background questionnaire

A questionnaire was designed to collect biographical data from the research participants, including age, gender, nationality, first language, and area of academic expertise. NNES participants were asked to report the IELTS or TOEFL scores that they had achieved most recently. It also included questions regarding their foreign language learning experience and familiarity with various foreign accents. The participants were also asked whether they had teaching or rating experience. Separate questionnaires were developed for EFL and ESL raters and ENL raters (Appendices 2 and 3).

3.4.2.4 Written instructions for verbal report procedures

A five-page booklet containing instructions for verbal reporting was developed for the participants (Appendix 4). The instruction booklet contained the procedures for verbal reporting illustrated in Figure 3.1 (Section 3.2.2). It also explained how to indicate their impressions of test-takers, the number of performances raters would watch, and the types of oral performances (i.e., individual presentations and pair conversations).
The instructions given to the raters underscored that they were expected to consider the communication ability of the speakers:

First, you will be asked to watch speaking performances of non-native English speakers in some English tests. While watching each performance, please judge the speaker’s communication ability. Please follow your impression and focus on any aspects of the performances as you like. You can take notes of your impression.

As can be seen, the raters were not provided with any pre-determined criteria or specific aspects of oral performances that they were required to focus on. Instead, the interpretation of communication ability was completely dependent on each rater. This open-ended approach was considered suitable for this study since it makes the ratings and verbal reports represent raters’ pure interpretation of communication ability without any influences of external criteria.

A possible problem with the instructions is that the raters might misinterpret that they are being asked to evaluate test-takers’ English proficiency. In the pilot study, some participants assumed that they were required to comment on the test-takers’ levels of English proficiency. Some contextual factors in the research (e.g., the researcher’s status as a PhD candidate in applied linguistics) might have led them to assume that English proficiency was the central concern of the study. Since this study aims to explore the features—not restricted to language proficiency—that are salient in the ability to communicate, it was necessary to emphasize what they are asked to focus on. Thus, it was determined that a verbal explanation would also be given to avoid any misinterpretation on the part of the raters. The explanation verbally provided to the raters was as follows:
Imagine that you are talking with the persons and thinking about how good their communication ability is. You can look at any aspects of the performances as you like. It doesn’t have to be their English. It’s all up to you what you pay attention to.

The first verbal reports that the raters provided were overall reasons for their ratings. The instructions below were given:

**Please talk about the reasons for your rating.** You can begin to talk immediately after indicating your impression of the speaker’s communication ability on the form. Please talk to the IC recorder in front of you. Please try to say as much as possible.

This type of verbal report is referred to as *Type 3 Verbalization* (Ericsson & Simon, 1993), which is not recommended as an instrument for deriving participants’ cognitive process. Rather, Ericsson and Simon (1993) claim that it is more appropriate to give instructions that do not require justifications or explanations, such as “tell me everything you were thinking as you watched the performances.” However, it was felt that such instructions confuse the raters and make them wonder what they are expected to talk about. This study thus employed Type 3 Verbalization that solicits the reasons for the rating since it aimed to elicit specific information about raters’ judgments or the reasons for their judgments. In fact, Bowles (2010) advocates the use of Type 3 Verbalization when it is necessary to investigate participants’ justifications for their behaviors. Previous studies also used this type of verbal protocols for investigating raters’ judgments (Brown, 2007; Brown et al., 2005; Ducasse, 2010; Kim, 2009a, 2009b; May, 2009, 2011a, 2011b; Zhang, 2010; Zhang & Elder, 2011).

The second verbal reporting was stimulated recall. The instructions were as
follows:

After talking about the reasons, you will be asked to watch the performance again. **This time, I would like you to elaborate on the reasons you have mentioned. Please stop the video every time you find anything that affected your impression, and talk about it in detail.** Stop as frequently as you can and comment anything you like. Please comment in as much detail as you can.

Similar to the first verbalization, the second verbalization was also Type 3 Verbalization in which the raters were required to extract specific information in their short-term memory and explain it. For this round, they were expected to elaborate on the reasons for their ratings by providing detailed features of the performance that affected their impressions.

For both verbal reports, the raters were asked not to face the researcher but directly speak to an IC recorder placed in front of them. With regard to this, some pilot study participants pointed out that talking to the machine was unnatural and made them uncomfortable. Despite these reactions, it was necessary to make them directly talk to the IC recorder, as it is important to attempt to minimize social interaction with the researcher as much as possible to prevent the raters from saying what they believe they are expected to. If they were allowed to speak to the researcher, the raters might adjust what they say based on the researcher’s subtle reactions or facial expressions (Corbin & Strauss, 2008).

3.4.2.5 Video-recorded oral performance data

Video recordings of oral performances were prepared for data collection. This
study employed monologic performances on the CET-SET and paired interactions from the Cambridge Exams. Ten videos of oral performances (13 test-takers) were prepared for data collection.

Seven monologic performances were selected from the pool of CET-SET data (Table 3.13). Seven test-takers with different overall scores (from A+ to D) were selected to collect the raters’ judgments of L2 speakers with various English proficiency levels. Individual presentations given in Part 2 of the test were chosen from the test tasks (see Section 3.3.1). In this part, each test-taker had received a topic to be presented shortly before the performance and presented for 1.5 to 2 minutes. The topics to be presented were regarding air pollution such as the solutions, causes, and consequences of air pollution. The number of female and male test-takers was also considered in the selection process. Visual and sound quality for each video was confirmed as suitable. Performances with low visual and sound qualities were excluded from the performances used for data collection.

<table>
<thead>
<tr>
<th>Test-taker</th>
<th>Gender</th>
<th>Topic</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>CET A+</td>
<td>Female</td>
<td>How to cope with air pollution</td>
<td>1:27</td>
</tr>
<tr>
<td>CET A</td>
<td>Male</td>
<td>Causes of air pollution</td>
<td>1:41</td>
</tr>
<tr>
<td>CET B+</td>
<td>Female</td>
<td>Consequences of air pollution</td>
<td>1:26</td>
</tr>
<tr>
<td>CET B</td>
<td>Male</td>
<td>Causes of air pollution</td>
<td>2:01</td>
</tr>
<tr>
<td>CET C+</td>
<td>Female</td>
<td>How to cope with air pollution</td>
<td>1:29</td>
</tr>
<tr>
<td>CET C</td>
<td>Female</td>
<td>Consequences of air pollution</td>
<td>2:00</td>
</tr>
<tr>
<td>CET D</td>
<td>Male</td>
<td>Causes of air pollution</td>
<td>1:20</td>
</tr>
</tbody>
</table>

Second, a pair of peer-peer interactions was selected from each of the CAE, FCE, and PET (Table 3.14). Unlike the CET-SET, these tests included test-takers with
different nationalities and L1s. These performances were selected because (a) the
test-takers had a wide variety of language backgrounds and (b) their visual and sound
qualities were better than other performances in the pool. The prompts given to the
test-takers were: “Discuss how each weather condition in the picture affects people’s
lives and choose two conditions that could have the most harmful effect” (CAE);
“Discuss the advantages and disadvantages of various jobs in the Olympic Games in the
picture and decide a job that would be most interesting to do” (FCE); and “Talk about
the place you live and say what you like and dislike about it” (PET). Tables 3.15 and
3.16 show each test-taker’s scores for the analytic criterion.

Table 3.14

<table>
<thead>
<tr>
<th>Test-taker</th>
<th>Test</th>
<th>Gender</th>
<th>Nationality</th>
<th>Topic</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anita</td>
<td>CAE</td>
<td>Female</td>
<td>Mexican</td>
<td>Weather conditions in the world</td>
<td>3:24</td>
</tr>
<tr>
<td>Marie</td>
<td>CAE</td>
<td>Female</td>
<td>Swiss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aida</td>
<td>FCE</td>
<td>Female</td>
<td>Italian</td>
<td>Jobs in the Olympic Games</td>
<td>3:41</td>
</tr>
<tr>
<td>Chloe</td>
<td>FCE</td>
<td>Female</td>
<td>French</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asad</td>
<td>PET</td>
<td>Male</td>
<td>Saudi</td>
<td>Place they live and furniture</td>
<td>3:38</td>
</tr>
<tr>
<td>Chulsoo</td>
<td>PET</td>
<td>Male</td>
<td>Korean</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Pseudonyms are used.

Table 3.15

<table>
<thead>
<tr>
<th>Test-taker</th>
<th>GR</th>
<th>LR</th>
<th>DM</th>
<th>P</th>
<th>IC</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anita</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Marie</td>
<td>3.0</td>
<td>3.0</td>
<td>3.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Note. GR=grammatical resource; LR=lexical resource; DM=discourse management;
P=pronunciation; IC=interactive communication; GA=global achievement.
Table 3.16

FCE and PET Test-takers’ Scores for Each Analytic Rating Criterion

<table>
<thead>
<tr>
<th>Test-taker</th>
<th>GV</th>
<th>DM</th>
<th>P</th>
<th>IC</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aida</td>
<td>3.5</td>
<td>3.0</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Chloe</td>
<td>3.0</td>
<td>3.0</td>
<td>3.5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Asad</td>
<td>3.5</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Chulsoo</td>
<td>3.5</td>
<td>4.0</td>
<td>3.5</td>
<td>4.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Note. GV=grammar and vocabulary. Although performances on FCE and PET are assessed based on the same criteria, their scores are not directly comparable because they apply different standards.

The 10 video recordings of oral performances were randomly ordered for each rater to minimize any order effect. A laptop computer with high-quality external speakers was prepared for showing the videos to the raters.

Finally, the following two performances were selected from the CET-SET and PET for practice: (a) a presentation on the consequences of air pollution demonstrated by a female test-taker of C+ level and (b) a paired conversation about favorite types of meal performed by a Vietnamese male and a Syrian female.

3.4.2.6 Rating form

A rating form was developed and used to record participants’ ratings of test-takers’ communication ability (Appendix 5). Each page contained a photograph of the test-taker(s), the title of the presentation (for individual presentations), the names of the test-takers (for peer-peer conversations), space to take notes, and a rating scale.

The rating scale used in the study was a seven-level scale bounded by Poor and Excellent, as illustrated in Figure 3.2. This type of scale is referred to as a semantic differential scale, which “consists of an unspecified number of seven-point rating scales that are bipolar, with each extreme defined by an adjective” (Davies et al., 1999, p. 177). The rationale for using this scale rather than other types of rating scales (e.g., Likert
Scale) is that semantic differential scales have been widely adopted for language attitudes research that investigates linguistic laypersons’ attitudes to language. One of the advantages of the scale is that it is able to elicit participants’ snap judgments and minimize their mental processing (Garrett, 2010). Since levels are not labeled as in a Likert Scale (e.g., average or good), it is expected that each rater will interpret the levels differently and use the scale inconsistently. Moreover, this study was concerned with what contributed to their judgments of test-takers’ communication ability (i.e., their verbal reports) rather than with inter- and intra-rater reliability. Thus, a simple scale that linguistic laypersons can easily apply was considered more appropriate.


*Figure 3.2. Semantic differential scale for individual presentations.*

The number of levels was set at seven because seven-point scales are widely used in speaking tests to discriminate test-takers’ proficiency levels (Luoma, 2004). The pilot study showed that the participants were able to apply the scale without great difficulty. The participants in the pilot study also commented that the semantic differential scale is easy to use in order to indicate their impressions.

In the case of the paired conversations between the two test-takers, the raters were asked to indicate their impressions of each test-taker’s communication ability (Figure 3.3). The name of each test-taker was given so that the raters could judge each test-taker without confusion (the video footages of the paired interactions also showed each test-taker’s name). This was also expected to facilitate verbal reports.
3.4.2.7 Post-session semi-structured interview questions

Semi-structured interview questions were developed to gather supplementary information on the raters’ evaluative judgments, as recommended by Green (1998). The questions are listed in Appendix 6.

A first set of questions asked the rater perceptions of the activity of judging oral performances and making verbal protocols. The main concern of the questions was with the transferability of the protocols obtained in this study. The raters were asked if the features they commented would affect their impression of people in real life. They were also asked about perceived difficulties in judging the performances and making comments on their thoughts in English (only the EFL raters).

The other set of questions were those asking the degree of contribution of various features to the raters’ impressions. The features included are (a) grammatical accuracy, (b) pronunciation or accent, (c) fluency, (d) non-verbal behaviors, (e) the content of speech, (f) speech organization, and (g) confidence. These features were selected based on previous studies (see Section 2.4). In this study, the raters were asked to initiate verbalization and allowed to determine the content of verbal reports rather than providing information based on the researcher’s inquiries. Therefore, it was impossible to obtain any information on the features that were not mentioned in their reports. In fact, a lack of comments on a certain feature does not necessarily mean that
the raters did not consider it during the task performance (Barkaoui, 2011; Lumley, 2005). For example, even though participants may not have mentioned the grammatical accuracy of the performance, it does not automatically mean that they had not considered accuracy or that accuracy had not influenced their impressions at all.

At the same time, as Green (1998) advised, the raters’ responses to the interview questions cannot replace their verbal protocols but should be treated as supplementary data. This is because, as Cohen (1994, 2011) argued, verbal protocols reflect what participants would actually do more accurately than interview responses. The interview data can only reveal the extent to which various features affected the raters’ impressions in general without taking into account the differences in test-takers’ proficiency levels and the elicitation tasks.

3.4.3 Data Collection Process

The procedures for data collection were as follows. First, prior to data collection, each rater received an e-mail with the plain language statements, and his or her eligibility for the research was confirmed. Only those who fulfilled the criteria in Table 3.9 and who had agreed to participate in the study met the researcher individually.

Data collections were conducted in a small project room in a library at the University of Melbourne. The raters received an explanation for the purpose of the research and signed the consent form in accordance with the University of Melbourne Human Research Ethics guidelines. Afterwards, they were asked to fill out the background information questionnaire.

Following this, the raters received the instruction booklet and the rating form. They were asked to read the instructions while the researcher was preparing the equipment for data collection. Following this, the researcher gave the instructions orally.
to make sure that the participants fully understood what they were expected to do. In particular, the following points were emphasized so as to avoid misunderstandings: (a) they were expected to rate communication ability and thus allowed to focus on any features of oral performances and behaviors of the speakers, and (b) they were asked to talk to an IC recorder placed in front of them and speak as much as possible.

After receiving the instructions for verbal reporting, the raters practiced rating communication ability and providing verbal protocols. They first practiced using the individual presentation on the CET-SET demonstrated by a female C+ test-taker. When participants misunderstood the procedures or instructions, the researcher pointed this out and corrected their misunderstanding. If raters wished to practice more, another performance (a female-male paired interaction on the PET) was played, but not otherwise.

Following the practice(s), the raters moved on to the data collection stage. The researcher sat behind them and did not intervene in their activity unless necessary. Breaks were occasionally offered during the session to prevent fatigue from affecting their judgments and protocols. After the raters judged and commented on all the performances, a post-session semi-structured interview was conducted. All of their verbal protocols and responses to the interview questions were audio recorded. Each data collection session lasted 1.5 to 3 hours. At the end of the session, they received an honorarium ($70) for their research participation. Figure 3.4 illustrates the activities the raters engaged in and the data collected from this study.
English was used throughout the data collection. The researcher used English in order to give the instructions and ask interview questions. The raters also received the written instructions and provided verbal protocols in English. However, Rater E (the Japanese participant) was asked to comment and answer questions in Japanese (the researcher’s L1).

### 3.4.4 Data Analyses

As Figure 3.4 shows, this study collected both quantitative data (ratings given by the raters) and qualitative data (protocol and interview data). Analyses were conducted as follows.
3.4.4.1 Analysis of quantitative data

The 23 raters’ ratings of 13 test-takers (in total 299 ratings) were statistically analyzed. Since this study was concerned with laypersons’ first impressions of the test-takers, the original scores were analyzed in case the raters changed their ratings while providing verbal protocols.

Many-facet Rasch analysis was conducted using FACETS 3.68.1 (Linacre, 2011). First, the rater measurement, including severity and inter-rater reliability, was estimated and examined to confirm the degree of agreement among the raters. Second, the test-taker measurement was estimated. The reliability of separation index was obtained to examine how widely the raters had separated the 13 test-takers’ ability. In addition, each test-taker’s perceived communication ability was estimated, and the rank order of the test-takers’ ability was calculated to examine how much rater impressions aligned with their English proficiency as measured by the proficiency tests.

3.4.4.2 Analysis of qualitative data

The qualitative data analyzed in the present study were the audio recordings of (a) the test-takers’ oral performances, (b) the raters’ verbal protocols (summary statements and stimulated recall data), and (c) the raters’ responses to the post-session interview.

The test-takers’ oral performances, the verbal protocols, and the raters’ responses to the post-session interview were transcribed using different orthographic transcription systems. For the test-takers’ performances, the researcher adopted the Conversation Analysis (CA) transcription system (Atkinson & Heritage, 1984; Lazaraton, 2002), which requires the transcriber to include as many details as possible in the transcript, such as periods of silence, lengthened sounds or syllables, fillers,
features of aspiration, overlapping, and so forth (see Appendix 7). The reason for using the CA transcription system was to identify and describe the specific features that the raters commented on. In so doing, it was possible to know in detail what features are positively or negatively perceived.

In contrast to the test-takers’ oral performances, the raters’ verbal protocols and the interview data were transcribed excluding detailed speech features because the main concern was with the content rather than the manner of speech. Table 3.17 shows the transcription conventions for verbal protocols. Since Rater E was asked to speak in Japanese throughout the data collection, her data was transcribed in Japanese and translated into English by the researcher.

<table>
<thead>
<tr>
<th>Text feature</th>
<th>Convention</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quotation from utterance</td>
<td><em>italics</em></td>
<td>Again, the usage of harm to the earth.</td>
</tr>
<tr>
<td>Paraphrase or correction</td>
<td>“</td>
<td>So she ended it with “Okay now I agree with you. Leave it alone.”</td>
</tr>
<tr>
<td>Emphasis</td>
<td>CAPS</td>
<td>I like when she said government SHOULD.</td>
</tr>
<tr>
<td>Inaudible talk</td>
<td>xxx</td>
<td>Don’t move them a lot xxx.</td>
</tr>
<tr>
<td>Unclear talk</td>
<td>( )</td>
<td>She asks her opinions of co … (co-conspirator).</td>
</tr>
<tr>
<td>Non-vocal action</td>
<td>(() )</td>
<td>I don’t know what is that. (laughs)</td>
</tr>
</tbody>
</table>

The transcribed verbal protocols—the summary statements and stimulated recall data—were segmented prior to coding. As an initial stage of analyzing protocol data, Green (1998) advises that “protocols are segmented, each segment corresponding to a chunk of behaviour, such as a statement or a phrase” (p. 19). Accordingly, protocol data need to be divided into chunks based on a unit of analysis determined by researchers.
Many studies using verbal protocol segmented protocol data according to the content or the idea emerged from the data rather than based on grammatical units (Brown et al., 2005; Ducasse, 2010; Lumley, 2005; May, 2011a). For example, Brown et al. (2005) employed the ideas unit as a unit of analysis, which was defined as “a single or several utterances, either continuous or separated by other talk but falling within the same turn, with a single aspect of the performance as the focus” (p. 14). Thus, a single segment could consist of a sentence or even a paragraph, depending on the idea presented by the informant, and researchers need to segment their data intuitively. One of the limitations of using a meaning-based unit of analysis is that segmentation is subjective (Lumley, 2005). This is why May (2011a) acknowledges the importance of checking inter-segmenter reliability. In addition, researchers are required to segment their data having a coding scheme in mind. In other words, segmentation and coding are inseparable, making the segmentation process a highly iterative process.

This study used a linguistically structure-based unit, the communication unit (C-unit), in segmenting the protocol data. In essence, a C-unit is an independent clause with its modifiers (Loban, 1976). In this sense, it is basically the same as the T-unit, which is defined as “one main clause plus whatever subordinate clauses happened to be attached or embedded within it” (Hunt, 1966, p. 735). The C-unit is more suitable for analyzing spoken discourse since it takes account of the unique characteristics of the spoken language (Foster, Tonkyn, & Wigglesworth, 2000), such as sentence fragments occurring due to hesitation and false starts. Moreover, a C-unit is also comprised of “answers to questions which lack only the repetition of the question elements to satisfy the criterion of independent predication” (Loban, 1976, p. 105). Loban claims that the segmentation is dependent on the grammatical structure of utterances but the speaker’s intonation patterns should be considered whenever necessary.
The following are examples from the data of this study. Each of the three examples below constitutes one C-unit, since they consist of one independent clause, respectively:

- So she had some ums and errs here, (U-A⁺-SR-2725)\(^4\)
- and the way she talks is more friendly and draws your attention more than the other girl. (A-Aida-SR-148)\(^5\)
- I guess she can’t express it because she doesn’t have the vocabulary. (D-Chloe-SR-468)

Sentence fragments are counted as a separate C-unit when the intonation pattern indicates a complete thought. For example, the following comments are counted as one C-unit, respectively, based on their final falling intonation:

- Again, the pronunciation of the words. (A-D-SR-102)
- Okay, much better. (U-A⁺-SS-5337)

In the case that raters summarize or quote the speaker’s utterance in the comments, the entire quotation is counted as a single C-unit as follows:

- So she ended it with “Okay now I agree with you. Leave it alone.” (N-Chloe-SR-2016)

\(^4\) Raters, test-takers, and type of verbal report are indicated in parenthesis. For example, U-A⁺-SR-2725 indicates that the segment was extracted from comments made by Rater U on CET A⁺ and the comment was provided in stimulated recall (stimulated recall=SR; summary statement=SS). The following number (2725) is the segment number.

\(^5\) This segment has one subject (the way she talks) and one compound predicate (is more friendly and draws your attention more than the other girl).
This segmentation process using C-units is highly consistent and independent of the researcher’s intuition because the segmentation is essentially structure-based. Similar to the meaning-based approach, the frequency of segments does not necessarily reflect the strength of rater attention. However, it can be considered that the number of C-units on a particular feature represents how much the rater elaborates on the feature, which appears to reflect the saliency of the feature to some degree.

As a result of segmentation, the protocol data were divided into 5,762 segments. Subsequently, each segment was scrutinized and grouped into similar themes. First, a microanalysis was conducted by closely observing the segments. A microanalysis is to “break into the data to make some sense out of the materials” (Corbin & Strauss, 2008, p. 59). The researcher went through all the segments, taking notes of the themes emerging from the data. These emerging themes were used to develop coding categories that group similar concepts. A single set of code categories was developed for the monologic and dialogic tasks since a large part of the emerging themes was the same.

Each of the segments was coded into the categories that were initially developed. A set of guidelines was consistently applied for coding the protocol data (Table 3.18). The researcher coded the segments interpreting their evaluative direction (positive, negative, or neither). In addition, the researcher noted the relationships among emerging themes whenever possible. For example, a cause-and-effect relationship was indicated using an arrow (→) when a single segment or multiple segments presented the relationship (e.g., *She looked confident because she spoke fluently* was noted as *fluency → confidence*). This process was considered important because exploring relationships among concepts is one of the goals of theory construction (Corbin & Strauss, 2008). After the initial coding, the coding categories were refined. Some categories were merged or divided in order to make the categories useful.
Table 3.18  
Guidelines for Coding Segments

Guideline
1. Interpret what behaviors/features the rater is commenting on.
2. Indicate the evaluative direction (positive, negative, or neither).
3. Code the main clause of segment.
4. Code the subordinate clause when the main clause is “it seems that …,” “we can see that …,” “I think/feel/notice that …,” or “it would be better if ….”
5. Code both main and subordinate clauses when the latter contains the reason(s) for the evaluation of overall communication ability.
6. Assign more than one code if necessary.
7. Indicate the relationships among emerging themes whenever possible.

The majority of the segments consisted of one theme were assigned a single code; however, multiple codes were assigned if a single segment contained more than one theme. This was inevitable due to the structurally oriented segmentation approach.

For example, the following segment was coded both as pronunciation and vocabulary:

- Too poor pronunciation and vocabulary. (J-D-SR-1460)

When the main clause of the segment showed the raters’ global judgment of communication ability, the subordinate clause stating its reasons was also coded at the same time (see Guideline No. 5 in Table 3.18). For example, the following segment was coded as a global ability and body movement:

- Okay, so this is one of the more lower rated communication abilities that I will rank because firstly, she kept on touching her hair, which was really distracting. (K-C-SS-4436)
All the segments were assigned one code or multiple codes in this way. As a result, the majority of the segments were single coded while some segments contained multiple themes (Table 3.19). At most, six codes were assigned to a single segment (as in the following example):

- But Aida was really relaxed, easy going, used good vocabulary, better sentence structure, better grammar, and was really more motivated about talking and was really interested about talking and putting different ideas and talking about different things and at the very end just had a good concluding remark as well. (A-Aida-SS-3401)

Table 3.19

<table>
<thead>
<tr>
<th>Number of Code Assigned</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of segments</td>
<td>5380</td>
<td>317</td>
<td>60</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5762</td>
</tr>
</tbody>
</table>

This structural-oriented segmentation approach was limited in that a large number of segments contained multiple themes. This violated a general principle proposed by Green (1998), who states that “the general principle of protocol analysis is that each segment of protocol represents a single process” (p. 75). Although structural-oriented segmentation avoided arbitrariness, approximately 6.6% of the segments contained more than one theme because of the approach.

To confirm inter-coder reliability, a second coder (a PhD candidate in applied linguistics) was asked to code part of the protocol data. Inter-coder reliability was checked through the following two phases. In the first phase, she was asked to code approximately 10% of the data into nine main categories (to be introduced in the next
chapter). The agreement rate was 86.9%. A discussion was made where she and the researcher disagreed, and the categorization was modified accordingly. In the second phase, the second coder was given 40-50 segments from each main category and asked to code them into its subcategories. The agreement rates ranged from 82.5% to 95.0%. The second coder and the researcher also discussed the coding of the segments whenever there was a disagreement. The categorization was finalized through discussion.

The transcripts of raters’ responses to the post-session interview were analyzed using NVivo 10 (QSR International, 2012). The interview data were analyzed mainly to examine the transferability of the protocol data into non-testing contexts and interpret the perceived importance of the performance features.

3.5 Summary

This chapter has described the methodology adopted in the present study and specific methods for collecting and analyzing data. This study primarily used a qualitative approach given that the purpose of the study was to explore the influential features on linguistic laypersons’ judgments. A verbal protocol analysis was used to elicit raters’ judgments of L2 speakers’ oral performances. Semi-structured interviews were conducted to collect data supplementing the verbal protocol data. These data were qualitatively analyzed to explore what factors had affected raters’ impressions of the speakers. In addition, ratings given by the raters were statistically analyzed to examine the relationship between the judgments of communication ability and the speakers’ language proficiency.
Chapter 4 Results

4.1 Introduction

This chapter presents the results of the analyses of both quantitative and qualitative data in this study. It firstly presents the results of quantitative analysis of the ratings provided by the raters. The subsequent part of the chapter shows the results of qualitative analysis of the verbal protocol and interview data, which reflect the test-takers’ behaviors and speech features that affected the raters’ impressions. The behaviors and features explored are classified into nine categories and discussed in detail with sample extracts.

4.2 Ratings of Rater Impressions

In total, 23 raters’ ratings of 13 test-takers (in total 299 ratings) were examined. They were asked to intuitively indicate their impressions of the test-takers’ communication ability on a scale of 1 (Poor) to 7 (Excellent). The original ratings were analyzed when raters changed their ratings while providing verbal protocols. The ratings thus reflected their first impressions of test-takers’ communication ability. The study conducted many-facet Rasch analysis using FACETS 3.68.1⁶ (Linacre, 2011).

4.2.1 Rater Measurement

Table 4.1 provides the rater measurement report, including the severity measures, fit statistics, and single rater-rest of the raters (SR/ROR) correlations. The most severe rater was Rater V, and the most lenient rater was Rater L. The rater

⁶ Each of the original ratings (ranging from 1 to 7) was multiplied by 2 for converting all of the ratings into integers (e.g., 4.5×2=9) since FACETS requires the data to be in the form of integers. Thus, the data ranged from 2 to 14.
separation reliability was .90, which indicates that the raters’ impressions of 13 test-takers were widely spread. The rater separation index (strata) was 4.24, indicating that there were approximately four statistically distinct groups of rater severity. This wide range of severity is not surprising given that the raters were not provided with any benchmark or standard but simply asked to rate using their intuitions. Moreover, the rating scale used in the study (semantic differential scale) did not contain any descriptors or labels for levels. The ratings they gave during the practice session seemed to affect their overall ratings. Prior to rating the 13 test-takers, the raters watched a performance demonstrated by a female CET-SET test-taker of C+ to familiarize them with the procedure. Her performance might become the benchmark of the subsequent performances. In fact, the six most severe raters (Raters V, J, G, N, F, and A) rated her performance 1 or 2, and the six most lenient raters (Raters L, T, H, B, R, and Q) rated her performance 5.

Table 4.1

Rater Measurement Report (Arranged by Severity)

<table>
<thead>
<tr>
<th>Total</th>
<th>Total Observed</th>
<th>Fair-M</th>
<th>Model</th>
<th>Inf Sup</th>
<th>Out Otr</th>
<th>Estim. Corr.</th>
<th>Exact Agree.</th>
<th>Exact Agree.</th>
<th>Obs %</th>
<th>Exp %</th>
<th>Nu Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>13</td>
<td>6.5</td>
<td>5.57</td>
<td>1.16</td>
<td>.20</td>
<td>.89</td>
<td>.92</td>
<td>.10</td>
<td>.20</td>
<td>.79</td>
<td>18.5</td>
</tr>
<tr>
<td>92</td>
<td>13</td>
<td>7.1</td>
<td>6.64</td>
<td>.89</td>
<td>.24</td>
<td>.95</td>
<td>.91</td>
<td>.08</td>
<td>.98</td>
<td>.92</td>
<td>22.4</td>
</tr>
<tr>
<td>95</td>
<td>13</td>
<td>7.3</td>
<td>6.76</td>
<td>.84</td>
<td>.20</td>
<td>.97</td>
<td>.92</td>
<td>.08</td>
<td>.98</td>
<td>.92</td>
<td>22.4</td>
</tr>
<tr>
<td>100</td>
<td>13</td>
<td>7.7</td>
<td>7.20</td>
<td>.69</td>
<td>.20</td>
<td>1.00</td>
<td>1.00</td>
<td>.08</td>
<td>.98</td>
<td>.92</td>
<td>20.6</td>
</tr>
<tr>
<td>108</td>
<td>13</td>
<td>7.7</td>
<td>7.20</td>
<td>.69</td>
<td>.20</td>
<td>1.00</td>
<td>1.00</td>
<td>.08</td>
<td>.98</td>
<td>.92</td>
<td>20.6</td>
</tr>
<tr>
<td>107</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>106</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>103</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>101</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>100</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>98</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>96</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>94</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>92</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>88</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>86</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>84</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>82</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>80</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>78</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>76</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
<tr>
<td>74</td>
<td>13</td>
<td>8.2</td>
<td>8.18</td>
<td>.93</td>
<td>.23</td>
<td>1.01</td>
<td>1.01</td>
<td>.71</td>
<td>.89</td>
<td>.89</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Model: Fixed (all same) chi-square: 205.5 d.f.: 32 significance (probability): .00
Model: Random (normal) chi-square: 188.4 d.f.: 21 significance (probability): .00
Inter-rater agreement: 93.9% Exact agreements: 858 > 99.9% Expected: 855.0 > 99.9%
Inter-rater reliability was also examined in Table 4.1. The mean SR/ROR correlation was .85 (see the column labeled “Corr. PtBis”). The SR/ROR correlation shows the degree to which raters agree about which performances are judged better and which are judged worse (Linacre, 2012), and a coefficient greater than .70 is considered high for a rating scale with several points (Myford & Wolfe, 2004). Therefore, the mean correlation of .85 is considered high and indicates that the raters generally agreed about the ranking of test-takers’ performances. Although infit and outfit indices of Raters H, I, and E appeared to indicate underfit, their SR/ROR correlations were greater than .70. Moreover, Observed % of exact agreement (25.5%) was approximately equal to Expected % of exact agreement (26.0%), which suggests that the raters acted as “independent experts” showing variation under identical rating conditions rather than “scoring machines” who give identical ratings (Eckes, 2011; Linacre, 2012). In summary, the rater measurement report shows that the raters agreed on whose performances were better and worse although they did not award identical ratings and the ratings were widely spread.

4.2.2 Test-taker Measurement

Table 4.2 shows the test-taker measurement report, including the ability measure of each test-taker and the reliability of separation index. The high degree of reliability index (.98), which is equivalent to Cronbach’s Alpha (Linacre, 2012), indicates that the ratings given by the 23 raters widely separated the test-takers. The separation index (9.39) shows that the test-takers could be separated into roughly nine statistically distinct strata according to their ability.
Table 4.2

Test-taker Measurement Report (Arranged by Ability)

<table>
<thead>
<tr>
<th>Total Score</th>
<th>Total Count</th>
<th>Observed Average</th>
<th>Fair-M</th>
<th>Model</th>
<th>Infit M-Sqa Std</th>
<th>Outfit M-Sqa Std</th>
<th>Estim. Phi</th>
<th>Corr. PTTbis</th>
<th>N test-taker</th>
</tr>
</thead>
<tbody>
<tr>
<td>282</td>
<td>23</td>
<td>12.3</td>
<td>12.20</td>
<td>2.03</td>
<td>.19</td>
<td>.94</td>
<td>-.1</td>
<td>.94</td>
<td>.0</td>
</tr>
<tr>
<td>265</td>
<td>23</td>
<td>11.3</td>
<td>11.55</td>
<td>1.62</td>
<td>.19</td>
<td>.77</td>
<td>-.6</td>
<td>.76</td>
<td>-.6</td>
</tr>
<tr>
<td>250</td>
<td>23</td>
<td>10.8</td>
<td>10.89</td>
<td>1.17</td>
<td>.17</td>
<td>1.42</td>
<td>1.1</td>
<td>1.42</td>
<td>1.1</td>
</tr>
<tr>
<td>246</td>
<td>23</td>
<td>10.7</td>
<td>10.81</td>
<td>1.12</td>
<td>.17</td>
<td>.46</td>
<td>-.9</td>
<td>.75</td>
<td>-.8</td>
</tr>
<tr>
<td>244</td>
<td>23</td>
<td>10.6</td>
<td>10.64</td>
<td>1.00</td>
<td>.17</td>
<td>.64</td>
<td>-1.1</td>
<td>.61</td>
<td>-1.1</td>
</tr>
<tr>
<td>238</td>
<td>23</td>
<td>10.3</td>
<td>10.59</td>
<td>.86</td>
<td>.16</td>
<td>1.06</td>
<td>-.2</td>
<td>1.27</td>
<td>-.8</td>
</tr>
<tr>
<td>219</td>
<td>22</td>
<td>8.5</td>
<td>8.70</td>
<td>.54</td>
<td>.15</td>
<td>.92</td>
<td>-.1</td>
<td>.97</td>
<td>-.3</td>
</tr>
<tr>
<td>200</td>
<td>23</td>
<td>8.7</td>
<td>8.62</td>
<td>.25</td>
<td>.14</td>
<td>1.02</td>
<td>-.1</td>
<td>.91</td>
<td>-.1</td>
</tr>
<tr>
<td>199</td>
<td>23</td>
<td>8.7</td>
<td>8.84</td>
<td>.19</td>
<td>.14</td>
<td>.96</td>
<td>.0</td>
<td>1.09</td>
<td>.3</td>
</tr>
<tr>
<td>168</td>
<td>23</td>
<td>7.3</td>
<td>7.36</td>
<td>-.27</td>
<td>.14</td>
<td>.91</td>
<td>-.5</td>
<td>.76</td>
<td>-.6</td>
</tr>
<tr>
<td>168</td>
<td>23</td>
<td>7.3</td>
<td>7.36</td>
<td>-.27</td>
<td>.14</td>
<td>.91</td>
<td>-.5</td>
<td>.76</td>
<td>-.6</td>
</tr>
<tr>
<td>122</td>
<td>23</td>
<td>5.3</td>
<td>5.10</td>
<td>-1.23</td>
<td>.16</td>
<td>.88</td>
<td>-.2</td>
<td>.97</td>
<td>0</td>
</tr>
<tr>
<td>77</td>
<td>23</td>
<td>3.3</td>
<td>3.27</td>
<td>-2.22</td>
<td>.19</td>
<td>.52</td>
<td>-1.8</td>
<td>.50</td>
<td>-1.6</td>
</tr>
</tbody>
</table>

Model, Pseudo: RMSE .18 Adj (True) S.D. 1.12 Separation 8.00 Strat. 8.98 Reliability .98
Model, Sample: RMSE .16 Adj (True) S.D. 1.06 Separation 7.06 Strat. 8.77 Reliability .98
Model, Random (normal) chi-square: 500.2 d.f.: 12 significance (probability): .00
Model, Random (t-distributed) chi-square: 500.2 d.f.: 12 significance (probability): .00

Table 4.2 is arranged by test-takers’ ability in the order of the most highly judged person (Anita) to the least highly judged person (CET D). Figure 4.1 illustrates the same information in a vertical map. As Table 4.2 and Figure 4.1 show, the order of perceived communication ability of the CET-SET test-takers aligned with that of their test scores (CET A+ was the highest score achiever, and CET D was the lowest score achiever); CET A+ was awarded the highest rating (1.63 logits), whereas the CET D was awarded the lowest rating (-2.22 logits). However, the communication abilities of some test-takers were not clearly distinguished by the raters. For example, the differences in the estimated measure between CET A and CET B+ and between CET B and CET C+ were 0.12 logits and 0.08 logits, respectively. This indicates that the raters judged the ability of these test-takers almost equally even though their English proficiency was different according to the test scores.
As opposed to the CET-SET, the rater impressions of the Cambridge Exams test-takers did not necessarily align with their English proficiency as gauged by the tests. Among the six test-takers, Anita received the highest ratings from raters (2.03 logits). She was a CAE test-taker (a more advanced test than the FCE and PET) and had obtained the higher logit score than her partner (Marie). However, another CAE test-taker, Marie, received ratings lower than FCE and PET test-takers. This result is worth noting because Marie, who passed the CAE, was predicted to have better English proficiency than those who passed the FCE and PET. Furthermore, the two PET test-takers (Asad and Chulsoo) were evaluated unexpectedly more highly than the FCE test-takers, who were supposed to be more advanced English speakers than the PET test-takers. In particular, the logit values of the PET test-takers indicate that their

Figure 4.1. Vertical map from the many-facet Rasch analysis.
communication abilities were distinguished from the three more advanced English
speakers (the differences were at least 0.44 logits). On the other hand, two speakers with
different levels of English proficiency—Marie (CAE) and Chloe (FCE)—were
perceived almost the same by the raters since the difference was only 0.06 logits.

In summary, the raters rarely awarded the identical rating to each test-taker, and
thus the level of raters’ severity was significantly different from each other. However,
they appeared to agree on whose performances were better and whose were worse. The
raters’ impressions of the CET-SET test-takers were aligned with their test scores or
English proficiency. However, high Cambridge Exams test score achievers did not
necessarily receive high impression ratings from the raters, suggesting that high English
language proficiency did not directly lead to the raters’ high evaluation of
communication ability.

4.3 The Emerging Features Contributing to Raters’ Judgments

Table 4.3 shows the main categories and their subcategories of themes that
emerged from the protocol data. Nine main categories and 29 subcategories were
explored. First, Demeanor represented the perceived confidence and anxiety of the
test-takers during their performances. Second, Non-verbal Behavior included the
speakers’ physical movements, eye contact, postures, and facial expressions. Third,
Pronunciation was the test-takers’ overall pronunciation, L1 accent, prosodic features,
and paralanguage or articulation. Fourth, Linguistic Resources contained the use and
choice of vocabulary, grammatical accuracy, grammatical complexity, and other
lexico-grammatical features. Fifth, Fluency represented the perceived smoothness and
temporal impressions of utterances, including the rate of speech, pause phenomena, and
repair phenomena. Sixth, Content included the quality of ideas, the organization of
Table 4.3
*The Nine Main Categories and Their Subcategories*

<table>
<thead>
<tr>
<th>Main Categories</th>
<th>Subcategories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Demeanor</td>
<td>1.1 Confidence</td>
</tr>
<tr>
<td></td>
<td>1.2 Anxiety</td>
</tr>
<tr>
<td></td>
<td>1.3 Other (e.g. attitudes, concentration)</td>
</tr>
<tr>
<td>2. Non-verbal Behavior</td>
<td>2.1 Body movement</td>
</tr>
<tr>
<td></td>
<td>2.2 Eye contact</td>
</tr>
<tr>
<td></td>
<td>2.3 Other (e.g. posture, facial expression)</td>
</tr>
<tr>
<td>3. Pronunciation</td>
<td>3.1 Pronunciation</td>
</tr>
<tr>
<td></td>
<td>3.2 Accent</td>
</tr>
<tr>
<td></td>
<td>3.3 Prosody &amp; Paralanguage</td>
</tr>
<tr>
<td>4. Linguistic Resources</td>
<td>4.1 Vocabulary &amp; Wording</td>
</tr>
<tr>
<td></td>
<td>4.2 Grammar</td>
</tr>
<tr>
<td></td>
<td>4.3 Other (e.g. textualization, variety of expressions)</td>
</tr>
<tr>
<td>5. Fluency</td>
<td>5.1 Overall fluency</td>
</tr>
<tr>
<td></td>
<td>5.2 Rate &amp; Amount of speech</td>
</tr>
<tr>
<td></td>
<td>5.3 Pause Phenomena</td>
</tr>
<tr>
<td></td>
<td>5.4 Repair Phenomena</td>
</tr>
<tr>
<td>6. Content</td>
<td>6.1 Ideas</td>
</tr>
<tr>
<td></td>
<td>6.2 Framing of ideas</td>
</tr>
<tr>
<td></td>
<td>6.3 Topical knowledge</td>
</tr>
<tr>
<td>7. Interaction</td>
<td>7.1 Interaction &amp; Engagement</td>
</tr>
<tr>
<td></td>
<td>7.2 Interactional pattern</td>
</tr>
<tr>
<td>8. Overall Impression</td>
<td>8.1 Overall performance &amp; Global ability</td>
</tr>
<tr>
<td></td>
<td>8.2 Overall message conveyance</td>
</tr>
<tr>
<td></td>
<td>8.3 Overall comprehensibility of message</td>
</tr>
<tr>
<td></td>
<td>8.4 Overall English proficiency</td>
</tr>
<tr>
<td>9. Other</td>
<td>9.1 Miscellaneous speech features and speaker behaviors</td>
</tr>
<tr>
<td></td>
<td>9.2 Comments unrelated to speaker behaviors</td>
</tr>
<tr>
<td></td>
<td>9.3 Rater’s general perception</td>
</tr>
<tr>
<td></td>
<td>9.4 Fillers, Procedures, &amp; Unclear Comments</td>
</tr>
</tbody>
</table>

Table 4.4 shows how many segments of the CET-SET and the Cambridge Exams were categorized into each main category, respectively. The commonalities between the two tests were the high frequency of segments on *Overall Impression* and

136
Content. The raters most frequently commented on Overall Impression including overall performance, message conveyance, comprehensibility, and English proficiency (21.1% and 19.4%, respectively). The frequency of the comments on Content—the speaker’s ideas, framing of ideas, and topical knowledge—was the second largest in both tests (15.1% and 13.7%, respectively). Table 4.4 also indicates that there were some differences in the frequency of comments between the two tests. First, the number of segments referring to Interaction was the smallest in the CET-SET (0.7%) but quite large in the Cambridge Exams (12.0%). Second, comments on Fluency were relatively large in the CET-SET (13.5%) whereas those in the Cambridge Exams were the second smallest (6.6%). Finally, rater comments on Linguistic Resources of the CET-SET test-takers were in sixth place (10.4%), but those of the Cambridge Exams test-takers were in third place (12.7%).

Table 4.4

Frequency of Segments on Each Main Category

<table>
<thead>
<tr>
<th>Main Category</th>
<th>CET-SET</th>
<th></th>
<th>Cambridge</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>1. Demeanor</td>
<td>213</td>
<td>5.7</td>
<td>157</td>
<td>6.3</td>
</tr>
<tr>
<td>2. Non-verbal Behavior</td>
<td>351</td>
<td>9.4</td>
<td>238</td>
<td>9.6</td>
</tr>
<tr>
<td>3. Pronunciation</td>
<td>432</td>
<td>11.5</td>
<td>253</td>
<td>10.2</td>
</tr>
<tr>
<td>4. Linguistic Resources</td>
<td>391</td>
<td>10.4</td>
<td>315</td>
<td>12.7</td>
</tr>
<tr>
<td>5. Fluency</td>
<td>507</td>
<td>13.5</td>
<td>163</td>
<td>6.6</td>
</tr>
<tr>
<td>6. Content</td>
<td>565</td>
<td>15.1</td>
<td>339</td>
<td>13.7</td>
</tr>
<tr>
<td>7. Interaction</td>
<td>28</td>
<td>0.7</td>
<td>303</td>
<td>12.2</td>
</tr>
<tr>
<td>8. Overall Impression</td>
<td>791</td>
<td>21.1</td>
<td>480</td>
<td>19.4</td>
</tr>
<tr>
<td>9. Other</td>
<td>465</td>
<td>12.4</td>
<td>228</td>
<td>9.2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3743</td>
<td>100.0</td>
<td>2476</td>
<td>100.0</td>
</tr>
</tbody>
</table>

### 4.4 Demeanor

Segments referring to the test-takers’ demeanor were 5.7% and 6.3% of all the
comments on the performances on the CET-SET and the Cambridge Exams respectively.
The emerging themes within this category were (a) confidence, (b) anxiety, and (c) other
demeanor factors. Table 4.5 shows the frequency of segments found in each subcategory.
The number of segments on the subcategories of the two tests was not significantly
different ($\chi^2(2) = 4.121, p > .05$).

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET</th>
<th></th>
<th>Cambridge</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>1.1 Confidence</td>
<td>72</td>
<td>1.9</td>
<td>56</td>
<td>2.3</td>
</tr>
<tr>
<td>1.2 Anxiety</td>
<td>88</td>
<td>2.4</td>
<td>50</td>
<td>2.0</td>
</tr>
<tr>
<td>1.3 Other</td>
<td>53</td>
<td>1.4</td>
<td>51</td>
<td>2.1</td>
</tr>
<tr>
<td>Demeanor Total</td>
<td>213</td>
<td>5.7</td>
<td>157</td>
<td>6.3</td>
</tr>
</tbody>
</table>

4.4.1 Confidence

The raters mentioned the degree of the test-takers’ confidence that they had perceived in both tests (1.9% and 2.3% of all the comments, respectively). Their comments included the level of confidence in general (Extract 1) and inferences about the test-takers’ confidence in the subject matter or English (Extract 2). In the paired interactions, some raters commented on the differences in confidence between two speakers (Extract 3).

1. She spoke quite confidently, (T-C$^+$-SS-5281)

2. She was obviously very confident in speaking English and very confident in the subject matter. (R-A$^+$-SS-5079)

3. I think she has slightly less self-confidence than Aida. (O-Chloe-SS-4847)
In the post-session interview, 12 raters mentioned the importance of this feature and acknowledged that it strongly affected their impression. Furthermore, four raters (Raters L, P, Q, and U) answered that confidence was one of the three features that had the greatest effect on their impressions of communication ability (Extract 4). At the same time, Raters I and K touched upon the differences in its importance between the monologic and dialogic tasks. Confidence was more important for the individual presentations than the paired interactions that occurred in a relaxing atmosphere (Extract 5).

4 Like I have met people whose English is not that really good, they are really struggling, but they are so confident. That eventually … because unless you are confident, it’s really hard to grab one’s attention, and if you cannot grab one’s attention, you cannot really communicate to that person. So if you are confident that’s like half of your job done. (P-Interview)

5 I think in the paired conversations I don’t think confidence was as important because it’s a more relaxed environment. I think the air pollution one, they looked more pressured to talk. But with the paired conversation, it’s fairly easy everyday topics and their environment also looked more relaxed like the room looked relaxing and the proctors looked very kind, even if they didn’t look friendly. (K-Interview)

Some protocol data indicated relationships with other categories. In many cases, the raters seemed to judge the test-takers’ confidence from visual cues such as body language, eye contact, and posture (Extract 6). Although the raters did not specifically explain what kind of behaviors led to the positive evaluation of confidence, avoiding
eye contact was at least thought of as an indicator of lack of confidence. Some raters’ responses in the interview revealed that confidence was also judged through facial expression; in particular, smiles led to positive rater impressions. Additionally, it was found that the test-takers’ articulation or voice dynamics affected raters’ judgments of confidence (Extract 7). Loud and clear articulation appeared to lead to positive evaluation of confidence.

6 Even though he corrected himself sometimes or would pause, he sounded quite confident while talking because his posture and he just kept looking straight at the audience and focusing on talking to them. (K-A-SR-1516)

7 Also the volume of her voice, she is not projecting outwards. So she doesn’t seem very confident. (W-C-SR-3125, 3126)

Some comments provided in the interview revealed that fluent utterances had an impact on raters’ judgments of confidence. Negative fluency-related features—hesitations, stuttering, filled pauses, and unfilled pauses—appeared to make the speakers look unconfident (Extract 8). In the paired interactions, some comments indicated that the test-takers’ interactional patterns affected how confident they looked (Extract 9). Interactional behaviors, such as dominating conversation, taking an active role, and initiating interaction, made the speakers appear highly confident.

8 If you have a shivering person “and – and – and” in talking like that, you seem like he is not very confident with what he is saying. (F-Interview)

9 So for Anita since I think she has been doing most of the talking, clearly she shows good confidence. (P-Anita-SS-4937)
Rater comments also showed that the perceived confidence directly influenced the raters’ impression of global communication ability and facilitated the comprehensibility of the speakers (Extracts 10-11).

10 I judge her communication ability very good because at the very beginning she seems confident and she looks totally aware of the whole things that she is going to talk about. (A-A\textsuperscript{+}-SS-3272)

11 I thought this speaker was generally quite confident, which sort of helped understanding him despite his hesitations (C-A-SS-3559)

The raters also commented on contrasts between confidence and linguistic resources, indicating that confidence might be able to compensate for linguistic problems in utterances (Extract 12). Some pronunciation problems were also compensated for by confidence (Extract 13). The raters’ overall impressions were not necessarily negative despite linguistic mistakes insofar as the speakers were confident.

12 Just in this part I can find so many places with wrong quite easy. / But he speaks it quite confidently as well as … / It’s easy to understand what he’s trying to say even if there are mistakes. (D-Asad-SR-474, 475, 476)

13 Even though English or pronunciation is a little bit bad, confident speakers speak very clearly, and I understand. But even though speakers have good English proficiency, I don’t understand very much if they speak quietly without confidence. (E-Interview)

Several segments coded as Rater’s General Perception included their beliefs
about confidence. Rater H mentioned that confidence helped NNESs avoid being unduly focused on linguistic mistakes, which in turn could lead to better communication (Extract 14).

14 For non-native speakers I think many of us, we have the problem of inconfident when we talk with others whether they are native speakers or non-native speakers. / So if we have the confidence to say what we want to say, we don’t care about whether we will make mistakes. / We can have very good communication with others. (H-A+SR-1009, 1010, 1011)

Figure 4.2 illustrates the relationships between the speaker’s confidence and other features. On the one hand, non-verbal behavior, articulation, fluency, and interactional pattern affected the raters’ judgments of confidence; on the other hand, the level of the speakers’ confidence affected the judgments of communication ability and comprehensibility, and compensated for problems with linguistic resources and pronunciation.
4.4.2 Anxiety

Comments on anxiety accounted for 2.4% and 2.0% of all the comments in the CET-SET and the Cambridge Exams, respectively. Comments within this subcategory reflected the raters’ inferences about the test-takers’ state of mind such as nervousness, comfort, security, calmness, and being relaxed (Extracts 15-18). A high degree of nervousness and anxiety was generally perceived negatively.

15 Obviously, he is a little bit nervous (A-A-SR-3293)
16 She is not comfortable at doing this exercise at all. (F-C-SS-3888)
17 Marie looks really insecure. (G-Marie-SR-928)
18 but he’s quite calm. (I-D-SS-4292)
The raters appeared to infer the level of anxiety based on non-verbal behavior, including hand movement, eye gaze, posture, and laughter (Extracts 19-21). More specifically, behaviors like playing with a ring, scratching nails and fingers, rubbing legs, maintaining tense shoulders, and avoiding eye contact made the test-takers look nervous and uncomfortable. Although smiles occasionally seemed to make the speakers look confident, it also made some raters feel that the speakers were uncomfortable with the situation. Moreover, the voice quality contributed to how comfortable the speakers sounded; low voice volume was generally considered a sign of anxiety.

19 and anxiety is demonstrated by the fact that she is scratching her nails and fingers and has a much lower voice than her friend Anita.

(F-Marie-VP-3905)

20 Obviously she is nervous, looking different directions, losing eye-contact with the persons she is talking with and … (A-B\textsuperscript{+}-SR-31)

21 So they are not uncomfortable with the situation, although you can tell Chulsoo is slightly more uncomfortable because, you know, there are obviously point he sort of laughs, which I find to indicate that he is … like he smiles … I mean those are indications that he is not as comfortable with the language. (L-Asad/Chulsoo-SS-4572)

Although anxiety did not directly affect the comprehensibility of the speakers’ message, it distracted the raters and negatively affected judgments of communication ability (Extracts 22-23). Furthermore, anxiety made the raters feel uncomfortable. In contrast, relaxation was considered to help communication in the interaction (Extract 24).
22 she came across as quite nervous in her body language, with tense shoulders / but it didn’t really impact how you could understand her. (S-A^+SS-5184, 5185)

23 She become nervous. / And yeah, it caused that her communication ability may seems to not really good. (I-C-SS-4287, 4288)

24 And he is very relaxed. / So that helps. (P-Chulsoo-SR-2291, 2292)

Figure 4.3 summarizes the relationships between the speaker’s anxiety and other features. On the one hand, non-verbal behavior and articulation affected the raters’ judgments of anxiety; on the other hand, the speakers’ anxiety affected the raters’ judgments of communication ability.

![Figure 4.3. Relationships between anxiety and other features.](image)

**4.4.3 Other Demeanor Factors**

Features coded as *Other* accounted for 1.4% and 2.1% of all the comments. Main themes emerging from the comments on the CET-SET test-takers were (a) the speakers’ concentration on the speech, (b) willingness to talk, (c) manner of speaking,
and (d) personal characteristics. Themes emerging from the comments on the Cambridge Exams were quite similar: (a) the speakers’ willingness to communicate, (b) manner of speaking, (c) personal characteristics, and (d) concentration on the interaction.

The largest number of comments on the individual presentations was related to the speakers’ concentration on the speech (Extract 25). The raters negatively evaluated those who lost concentration, looked confused, and lost track. Concentration was mentioned when the raters watched the paired performances although the frequency of such comments was relatively low (Extract 26).

25 I generally think she felt comfortable and was concentrated on what she wants to convey, was not that nervous … (A-A⁺-SS-3225)

26 That is good about his communication skills just trying to look at the other person’s face and at the same time concentrating on what he is talking about. (A-Asad-SR-161)

The most frequently stated feature in the paired interaction was willingness to communicate with the partner. More specifically, the raters focused on how enthusiastic and interested the speakers seemed (Extracts 27-28). Although one rater judged willingness from facial expression, other raters did not provide specific reasons for their judgment of willingness to communicate.

27 But I think Anita was full of passion in this conversation.
   (H-Anita-SS-4199)

28 They both came across as interested in each other’s conversation that they
were following the conversations (U-Asad/Chulsoo-SS-5438)

In both tests, the raters made a number of comments on the test-takers’ general manner of speaking (Extract 29). Although these comments were not specific enough to show what behavior was actually considered, some raters pointed out several features related to the manner of speaking. A natural way of talking was positively evaluated whereas contrived or forced speech was negatively perceived (Extract 30).

29 The way she speak even must look like the native speaker as far as I see.
   (B-B−-SR-217)

30 The only thing is that it felt a little contrived in that I felt the emotion of what was being said didn’t quite match if that made sense. (T-A+−SS-5246)

Some raters mentioned personal characteristics about the speakers such as talkativeness, self-esteem, and friendliness through inference from their demeanor (Extract 31).

31 Aida came across as friendly and more outgoing, and therefore probably a little bit better but not enough, little bit, (U-Aida-VP-5421)

4.4.4 Differences among Proficiency Levels

The number of comments on Demeanor given to each test-taker was examined. Table 4.6 shows the frequency of segments referring to demeanor—confidence, anxiety, and other aspects of demeanor—of the CET-SET test-takers. The number of positive and negative segments suggests that language proficiency gauged by the CET-SET was
moderately aligned with raters’ judgments of the speaker’s demeanor. The three most proficient test-takers (CETs $A^+$, $A$, and $B^+$) received more positive comments than negative ones, whereas negative comments given to less proficient test-takers (CETs $B$, $C^+$, $C$, and $D$) surpassed positive ones.

Table 4.6

<table>
<thead>
<tr>
<th></th>
<th>$A^+$</th>
<th>$A$</th>
<th>$B^+$</th>
<th>$B$</th>
<th>$C^+$</th>
<th>$C$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>24</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Negative</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>15</td>
<td>19</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>Neither</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>20</td>
<td>23</td>
<td>26</td>
<td>27</td>
<td>52</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 4.7 displays the frequency of segments referring to the demeanor of the Cambridge Exams test-takers. The evaluation of demeanor was not necessarily linked to language proficiency as measured by the test. The raters gave Asad the largest number of positive comments on demeanor followed by Anita, whose language proficiency was evaluated most highly by the test. In fact, the two PET test-takers obtained many positive comments on their demeanor. In contrast, Marie’s demeanor was perceived most negatively. Although the evaluation of demeanor was not aligned with English proficiency, it seemed to be more closely related to rater impression of communication ability; the three test-takers who received a larger number of positive comments on demeanor (Anita, Asad, and Chulsoo) were perceived to have higher communication ability than the other test-takers (Marie, Aida, and Chloe).
Table 4.7

*Frequency of Segments on Demeanor Given to the Cambridge Exams Test-takers*

<table>
<thead>
<tr>
<th></th>
<th>CAE</th>
<th></th>
<th>FCE</th>
<th></th>
<th>PET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anita</td>
<td>27</td>
<td>Marie</td>
<td>13</td>
<td>Chloe</td>
<td>31</td>
</tr>
<tr>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>4</td>
<td>28</td>
<td>10</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Neither</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>35</td>
<td>25</td>
<td>32</td>
<td>33</td>
</tr>
</tbody>
</table>

4.4.5 Differences among Raters

The number of segments referring to the speakers’ demeanor mentioned by each rater was calculated (Table 4.8). The results showed that there were large individual differences in terms of the frequency of comments. Raters B and V provided only three segments regarding demeanor (accounting for 1.2% and 1.1% of their comments), whereas Rater A produced 63 segments (accounting for 15.5% of his comments). The average number of segments was 16.1.

Table 4.8

*Number of Segments on Demeanor Mentioned by Raters*

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>63</td>
<td>3</td>
<td>17</td>
<td>11</td>
<td>9</td>
<td>11</td>
<td>7</td>
<td>19</td>
<td>24</td>
<td>27</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>%</td>
<td>15.5</td>
<td>1.2</td>
<td>8.2</td>
<td>5.4</td>
<td>2.9</td>
<td>4.9</td>
<td>1.8</td>
<td>5.4</td>
<td>6.7</td>
<td>13.0</td>
<td>7.5</td>
<td>7.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rater</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>5</td>
<td>10</td>
<td>14</td>
<td>16</td>
<td>7</td>
<td>21</td>
<td>10</td>
<td>16</td>
<td>22</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>%</td>
<td>2.2</td>
<td>3.6</td>
<td>8.2</td>
<td>5.7</td>
<td>3.6</td>
<td>9.1</td>
<td>6.8</td>
<td>6.7</td>
<td>7.1</td>
<td>1.1</td>
<td>4.1</td>
</tr>
</tbody>
</table>

The frequency of comments on demeanor reflects to some extent its effect on the raters’ overall judgments. Raters A and J, whose comments on demeanor accounted for more than 10% of their comments overall, argued for the importance of confidence and anxiety in their judgments. Rater J mentioned the reasons for its importance in his
judgment (Extract 32).

32 Yeah because if as you speak to me in a nervous way, I may feel that you are not sure about what you really talk about. So maybe I will not trust what you talk about 100% because you, yourself, are not confident of yourself. So how can I trust what you talk? (J-Interview)

Although the number of comments on demeanor in some raters’ protocol data was not high, these raters’ responses to the interview showed that they acknowledged the importance of demeanor, in particular confidence. For example, nine raters who produced a small number of comments (Raters B, E, F, G, M, N, Q, V, and W: from 1.1% to 4.9%) stated that confidence strongly affected their judgments and claimed its importance in communication ability. This finding suggests that the frequency of mentions might not necessarily correspond to the strength of rater attention or the saliency of the feature.

4.5 Non-verbal Behavior

Comments related to the speakers’ non-verbal behavior accounted for 9.4% and 9.6% of the verbal protocol data on the CET-SET and the Cambridge Exams respectively (Table 4.9). Non-verbal behavior mentioned by the raters included a variety of visual actions including (a) body movement, (b) eye contact, and (c) other behaviors. Among these, the raters mentioned eye contact of the CET-SET test-takers most frequently, followed by body movement. In contrast, body movement was the most frequently mentioned subcategory of the Cambridge Exam test-takers, followed by eye contact. There was a statistically significant difference in the frequency with which
non-verbal behavior was mentioned between the two tests ($\chi^2(2) = 42.554, p < .01$).

Table 4.9

*Frequency of Segments on Each Subcategory of Non-verbal Behavior*

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET</th>
<th></th>
<th>Cambridge</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>2.1 Body movement</td>
<td>122</td>
<td>3.3</td>
<td>137</td>
<td>5.5</td>
</tr>
<tr>
<td>2.2 Eye contact</td>
<td>180</td>
<td>4.8</td>
<td>59</td>
<td>2.4</td>
</tr>
<tr>
<td>2.3 Other</td>
<td>49</td>
<td>1.3</td>
<td>42</td>
<td>1.7</td>
</tr>
<tr>
<td>Non-verbal behavior Total</td>
<td>351</td>
<td>9.4</td>
<td>238</td>
<td>9.6</td>
</tr>
</tbody>
</table>

4.5.1 Body Movement

Body movements mentioned by the raters included hand movement, head movement, and body movement. The raters made a larger number of references to them when viewing the paired interaction (5.5% of all the comments) than when observing the individual presentations (3.3% of all the comments). They often referred to the speakers’ body movements as *body language* or *(hand) gestures*.

Raters noticed hand movements both that were used as a part of an utterance and that were not. The former type of movements was intentionally employed to convey or reinforce the meaning of utterances. Gestures indicating the point of the arguments the speakers were making at the time were often remarked upon in the presentations. Rater F occasionally suggested that test-takers should have used such gestures to reinforce their message (Extract 33). Those mentioned in the paired interactions included actions describing objects or conditions and pointing to a picture (Extracts 34-35: Appendix 8).

33 She could have just put her ((uses a gesture)) “All right. Second argument.” using her hands and her arms. (F-B’-SR-670)
Around here, he incorporates gestures in a good way, for example, to express *bigger*. (E-Chulsoo-SR-636)

but Chloe expresses herself more in body language because she would point to the pictures and stuff like that. (M-Chloe-SS-4652, 4653)

Many raters commented that meaningful gestures successfully reinforced the message and facilitated their understanding of what the speakers were trying to say. For example, some raters commented that Asad’s gesture successfully described the unity of his family (Extract 36: see also Appendix 8). Although the excessive use of gestures was perceived to be distracting, a lack of gestures was considered negative in communication, even in the individual presentations. The moderate use of gestures drew the raters’ attention and gave them the impression that the test-takers were engaging with the audience or their partner (Extract 37). However, the raters seemed to negatively evaluate the test-takers’ pointing actions that were not accompanied by precise vocabulary (Extract 38).

So the fact that he was using his hands to gesture “the family being together” was quite good. / It strengthened his message there. (V-Asad-SR-3012, 3013)

If they had their hands in their laps and were quite still, it wasn’t as engaging, whereas if they had their hands on the table and moving them around looking at the speaker, then that was more engaging. (Q-Interview)

and I think what drew me a lot was the fact that she was only pointing at what is this, what is that. (P-Marie-SS-4950)
Several raters (Raters A, E, H, and I) also mentioned that gestures facilitated their understanding of the speakers despite a lack of linguistic resources and poor pronunciation. In this regard, gestures appeared to compensate for low English proficiency. Rater A argued for the importance of gestures when there were linguistic problems (Extract 39).

39 It was really important, especially when I had problem with the words or maybe there were some grammatical points or there were some other missing factors, I found these features very important. (A-Interview)

The other type of hand movements was those that were not directly connected to the content of speech and those the speakers unintentionally used. These actions included touching hair, rubbing the nose or leg, scratching the neck, playing with a ring, and so forth (Extracts 40-41). Additionally, the raters referred to the speakers’ hand position, fingers, and the way of holding a cue card. Rater F referred to this type of hand movement as parasite gestures (Extract 42). The hand movements that the speakers unconsciously demonstrated were negatively perceived in most cases. They distracted the raters from the content and made the speakers appear anxious and uncomfortable (Extracts 43-44).

40 She brings her hands up to her hair and her face. (W-C-SR-3121)
41 Chloe starts playing with her ring. (G-Chloe-SR-930)
42 The few gestures that JK does do not support his speech actually. / More parasite gestures in my opinion, which are not help to convey the message he’s actually speaking and overly expressing. (F-Chulsoo-SR-750, 751)
Around here, she tends to touch her hair when she was at a loss for words.  
/ It distracts me a little bit / and I can’t help looking at it. (E-C-SR-548, 549, 550)

But he looks like he’s rubbing his legs quite a lot, which also indicates to me that he’s very uncomfortable. (L-D-SR-1696)

The raters’ comments also included the test-takers’ head nods, particularly in the paired interactions (Extract 45). The raters perceived that head nods reflected an understanding of the partner. In addition, the speakers’ physical movements including swaying or moving back and forth were noticed. These unintentional actions gave the raters the impression that the speakers were nervous and distracted them from the message (Extract 46).

and he nodded his head sometimes to give response to Asad. / It shows that he could understand what Asad is talking about (H-Chulsoo-SR-1147, 1148)  

And when he became more and more nervous, he couldn’t stop shaking his body and make the listeners very uncomfortable. (H-D-SS-4182)

With regard to the importance of body movement in the speakers’ performance, six raters (Raters E, K, L, Q, T, and U) acknowledged it as one of the three features they attended to most. In particular, strategic use of body movements was positively perceived. Rater I mentioned the importance of non-verbal body movement (Extract 47).
For verbal or oral conversation I think it is important, because it is not only help us to understand what the person is trying to say, but also to keep us focused and concentrate on our partner when they say something.

(I-Interview)

Figure 4.4 illustrates the relationships between body movement and other features. The speakers’ body movement affected the judgments of communication ability, engagement, confidence, and anxiety. It also affected comprehensibility and compensated for limitations of linguistic resources.

**Figure 4.4.** Relationships between body movement and other features.

### 4.5.2 Eye Contact

The test-takers’ eye contact was mentioned more frequently in the CET-SET (4.8% of the comments) than the Cambridge Exams (2.4% of the comments). The raters paid attention to what the speakers were looking at. They commented on whether the speakers were looking at the person or audience whom they were addressing (Extracts
48-49). Without exception, the raters positively evaluated eye contact with the addressee. For some raters, it was one of the decisive factors when judging communication ability (Extract 50).

48  She I think she has good eye contact with the person that she is talking to.  
    (A-A\textsuperscript{+}-SR-10)

49  What I liked most was that he made good eye contact whereas in some of  
    the other videos, when people were stuck for words to say they looked off  
    into the distance, (T-Asad-SS-5324)

50  Actually, I think his communication ability is poor or near to poor because  
    when he tries to communicate with other people, he is very nervous and  
    tries to not have any eye contact with the people that he is talking to.  
    (A-D-SS-3356)

In contrast, avoiding eye contact was viewed negatively. More specifically, the  
raters mentioned the test-takers’ eye gaze to the ceiling, the far end of the room, their  
lap, the script on the table, the window, and so forth. Additionally, a frequent change of  
eye gaze was rated negatively (Extracts 51-52). Nevertheless, referencing the script or  
cue card during the presentations was not necessarily perceived negatively as long as  
the speakers attempted to make eye contact (Extract 53).

51  She is looking somewhere else, somewhere perhaps the far end of the room  
    and not looking in the eye. (F-B\textsuperscript{+}-SR-672)

52  But her eyes just look up, look down, and look around not look at the  
    teachers or her friends sitting around. (J-B\textsuperscript{+}-SR-1426)
and while he was reading off cards he was also looking up and making eye contact with who he was speaking to. (Q-A-SS-5006)

The raters’ comments suggested that eye contact was connected to the perceived level of confidence and anxiety. Eye contact was generally perceived as indicating that the speakers were highly confident and comfortable, whereas avoiding eye contact was thought to be a sign of nervousness or unfamiliarity with the subject (Extracts 54-55). In addition, it gave the raters the impression that the speakers were engaging with the listener and establishing rapport (Extract 56).

he gives eye contact when he is putting his point across, which means to me well, or indicates to me that he is obviously confident in the fact that he knows what he is trying to say (L-D-SS-4538)

She’s flicked her eyes briefly, / and she’s flicked her eyes away / and … from a Kiwi perspective, that’s not coming across as shifty but more … unsure and not knowing topic. (U-C^+SR-2788, 2789, 2790)

Okay, one last thing the one thing that affects me is also the eye contact because eye contacts are very important in communication. Yeah, you establish rapport, or you don’t establish rapport. (M-Interview)

Six raters (Raters M, N, Q, T, U, and W) stated that eye contact was one of the most salient features influencing their judgments of communication ability (Extract 57).

Well, they were interacting with the other person because the ones that weren’t interacting stood out more to me when they are busy looking out
here and just not … well, that person is here, I am going to look over here
gave the really poor impression. (U-Interview)

Figure 4.5 displays the relationships between eye contact and other features.
Eye contact affected the judgments of the speakers’ communication ability, anxiety,
topical knowledge, engagement, and confidence.

![Diagram](image)

*Figure 4.5. Relationships between eye contact and other features.*

### 4.5.3 Other Non-verbal Behavior

The raters mentioned non-verbal behavior other than body movement and eye contact. The frequency of the mentions of these behaviors was not large, accounting only for 1.3% and 1.7% of all the comments on the CET-SET and the Cambridge Exams. In both tests, the major features or behaviors mentioned by the raters were (a) posture, (b) facial expressions, and (c) laughter. Features mentioned less frequently included the speakers’ physical appearance (e.g., clothes, age) and mirroring the partner.
The raters referred to the test-takers’ posture or how they sat during the performance. In the individual presentation, speakers who were sitting up were evaluated positively. In the paired interaction, sitting in a relaxed state gave the raters a positive impression. Although posture was not considered to facilitate an understanding of the message, it appeared to influence the impression of willingness to communicate or comfortableness (Extract 58). Rater W pointed out that posture was one of the most influential features influencing his judgments (Extract 59).

58 I mean the whole posture was something showed that maybe she is not willing to talk. (A-Chloe-SR-140)

59 A second thing would probably be body language. Where were their eyes looking, were they looking off into space, was their chest up, or were they slouched over. (W-Interview)

Facial expression was another feature of non-verbal behavior frequently mentioned by the raters. Some raters claimed that facial expression showed the level of engagement with the listener and drew their attention to the content (Extracts 60-61). In particular, several raters mentioned the test-takers’ smiles (Extract 62), although the way this was evaluated was not consistent among the raters. The judgment of speakers’ smiles might depend on what kind of smile they exhibit. Additionally, evaluations may vary as a function of the type of discourse; smiles may be perceived more negatively in formal presentations than in casual conversations.

60 Here Chloe shows facial expression which I think helps what she’s saying. (K-Chloe-SR-1571)
It’s very important because firstly, yes, you are speaking with your mouth but then you are holding other person’s attention with your face. (P-Interview)

I think his smile kind of helps express what he wants to say / and … yeah it helps the conversation to flow. (T-Chulsoo-SR-2697, 2698)

The raters also mentioned laughter demonstrated by the speakers, and again their evaluation varied. The raters’ comments suggested that laughter was occasionally considered appropriate in paired interactions but not individual presentations (Extracts 63-64). As negative aspects of laughter, some raters mentioned that speakers who laughed looked defensive and uncomfortable. Rater L provided her general perception of laughter (Extract 65).

and what’s more he was also laughing while talking. / So that wasn’t a good sign for this activity. (K-D-SS-4444, 4445)

I liked that they laughed together (T-Asad/Chulsoo-SS-5333)

And laugh always indicates to me that you’re uncomfortable with the situation and sort of trying to break the awkwardness that you feel. (L-NA-SR-1756)

Figure 4.6 shows the relationships between other non-verbal behavior and other features. Posture, facial expressions, and laughter affected the raters’ judgments of communication ability, demeanor (e.g., anxiety, confidence), and engagement.
4.5.4 Differences among Proficiency Levels

The number of segments referring to the speakers’ non-verbal behavior—body movement, eye contact, and other non-verbal behavior—was examined (Table 4.10). The direction of the evaluative judgments suggested that the raters’ judgments of non-verbal behavior were related to English proficiency to some degree. The highest CET-SET score achievers (CETs A$^+$ and A) received a positive evaluation more often than did the lower score achievers (CETs B$^+$, B, C$^+$, C, and D). The number of comments provided to CET A$^+$ was the smallest (N=23). This could be because the raters paid greater attention to the content of her speech and/or she did not use many gestures. Rater M stated that this test-taker did not have to rely on non-verbal cues to express her thoughts and draw the listener’s attention (Extract 66).

   66 In this case here you note down that remember I said that this girl in
Chinese she wasn’t gesturing much, but she was making a very weighted kind of a speech that people are drawn to her speech to what she is trying to express, and you are saying that this girl is impressive. Things are well thought of and stuff like that. So in this case hand gesture, she doesn’t need that kind of, she doesn’t need a lot of hand gesture and stuff like that.

(M-Interview)

Table 4.10

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>A⁺</th>
<th>A</th>
<th>B⁺</th>
<th>B</th>
<th>C⁺</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>18</td>
<td>37</td>
<td>16</td>
<td>14</td>
<td>3</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Negative</td>
<td>4</td>
<td>12</td>
<td>49</td>
<td>22</td>
<td>40</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td>Neither</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>52</td>
<td>70</td>
<td>38</td>
<td>45</td>
<td>59</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 4.11 shows the frequency of comments on non-verbal behavior given to the Cambridge Exams test-takers. In particular, the PET test-takers received many evaluations compared with the other test-takers. The evaluation of non-verbal behavior was not necessarily linked to the test results that reflected their English proficiency. It was revealed that Anita, Aida, Asad, and Chulsoo attracted a larger number of positive comments than negative ones. In the post-session interview, five raters (Raters A, D, K, M, and O) referred to Asad’s gestures, saying that his use of gestures facilitated message conveyance. Although Asad’s English proficiency was not high among the Cambridge Exams test-takers, he appeared to successfully incorporate gestures in the interaction.
Table 4.11
*Frequency of Segments on Non-verbal Behavior Given to the Cambridge Exams Test-takers*

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>CAE Anita</th>
<th>CAE Marie</th>
<th>FCE Aida</th>
<th>FCE Chloe</th>
<th>PET Asad</th>
<th>PET Chulsoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>17</td>
<td>7</td>
<td>11</td>
<td>8</td>
<td>52</td>
<td>47</td>
</tr>
<tr>
<td>Negative</td>
<td>8</td>
<td>33</td>
<td>10</td>
<td>17</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Neither</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>7</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>49</td>
<td>24</td>
<td>32</td>
<td>74</td>
<td>61</td>
</tr>
</tbody>
</table>

4.4.6 Differences among Raters

The number of segments referring to non-verbal behavior mentioned by the raters was calculated (Table 4.12). Although Raters B and C did not make any reference to them in their verbal reports, comments on non-verbal behavior accounted for more than 20% of all the comments made by Raters F and U. Table 4.12 partially reflects the degree of contribution of non-verbal behavior to raters’ judgments. Raters who made a large number of references to non-verbal behavior (Raters J, K, L, T, U, and W) stated that body movement, eye contact, and other non-verbal behavior were one of the three features most strongly affected their overall judgments. Among others, Rater U, who provided 73 segments related to non-verbal behavior, explicitly pointed out the saliency of both physical movements and eye contact in her judgments of communication ability. At the same time, Raters D and R argued for the importance of non-verbal behavior and eye contact in general although they did not frequently mention them in their verbal protocol. Furthermore, Rater B stated that she did not focus on non-verbal behavior in the tests but would like her interlocutor to use them in daily life.
Table 4.12
Number of Segments on Non-verbal Behavior Mentioned by Each Rater

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>28</td>
<td>57</td>
<td>23</td>
<td>37</td>
<td>27</td>
<td>35</td>
<td>28</td>
<td>46</td>
</tr>
<tr>
<td>%</td>
<td>9.3</td>
<td>0.0</td>
<td>0.0</td>
<td>1.5</td>
<td>9.1</td>
<td>25.3</td>
<td>5.9</td>
<td>10.6</td>
<td>7.6</td>
<td>16.9</td>
<td>17.4</td>
<td>14.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rater</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>21</td>
<td>18</td>
<td>8</td>
<td>19</td>
<td>18</td>
<td>1</td>
<td>18</td>
<td>30</td>
<td>73</td>
<td>10</td>
<td>51</td>
</tr>
<tr>
<td>%</td>
<td>9.1</td>
<td>6.5</td>
<td>4.7</td>
<td>6.7</td>
<td>9.2</td>
<td>0.4</td>
<td>12.3</td>
<td>12.5</td>
<td>23.5</td>
<td>3.7</td>
<td>10.9</td>
</tr>
</tbody>
</table>

4.6 Pronunciation

Comments referring to the test-takers’ pronunciation accounted for 11.5% and 10.2% of the comments given to the CET-SET test-takers and Cambridge Exams test-takers, respectively (Table 4.13). In both tests, pronunciation was the fifth largest category. The emerging themes within pronunciation included (a) pronunciation, (b) L1 accent, and (c) prosody and paralanguage. There was no statistically significant difference in the frequency between the two tests at the level of $p=.01$ ($\chi^2(2) = 8.979$).

Table 4.13
Frequency of Segments on Each Subcategory of Pronunciation

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET</th>
<th>Cambridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>3.1 Pronunciation</td>
<td>207</td>
<td>5.5</td>
</tr>
<tr>
<td>3.2 Accent</td>
<td>71</td>
<td>1.9</td>
</tr>
<tr>
<td>3.3 Prosody &amp;Paralanguage</td>
<td>154</td>
<td>4.1</td>
</tr>
<tr>
<td>Pronunciation Total</td>
<td>432</td>
<td>11.5</td>
</tr>
</tbody>
</table>

4.6.1 Pronunciation

The raters referred to the test-takers’ pronunciation in both tests (5.5% and 3.9% of all the comments, respectively) making overall judgments of pronunciation. They often evaluated global pronunciation using adjectives such as good, decent, clear,
poor, and bad (Extract 67). Only a few raters (Raters A, B, and T) referred to pronunciation saying whether it was correct or wrong although the frequency of such comments was quite small. There was only one segment referring to the nativeness of pronunciation (Extract 68). In fact, two EFL raters (Raters A and B) frequently mentioned that they would pronounce a word differently than some test-takers did (Extract 69).

67 This speaker, I think his pronunciations were quite good (D-A-SS-3637)
68 The way she pronounced pretty much similar to the native speaker because the way she pronounced is really good and she also knows how to stress some key words in the sentence when she speaks. (B-B⁺-SS-3451)
69 Obviously, I noticed some different pronunciations which were different my way of talking in English, (A-Asad/Chulsoo-SS-3407)

The test-takers’ pronunciation mostly impacted the comprehensibility of the word or message. The raters often attributed the incomprehensibility of speech to the test-takers’ pronunciation (Extract 70). In fact, many raters acknowledged the importance of pronunciation in judging communication ability; 11 raters mentioned that it was one of the strongest features affecting their judgments. Nevertheless, although the importance of pronunciation was often acknowledged, the raters did not always negatively rate problematic pronunciation as long as the speakers’ overall message was comprehensible (Extract 71).

70 I think the main barrier to understanding what he was saying was his pronunciation of words. (T-D-SS-5299)
And pronunciation for me is also important. Maybe two speakers I really had hard time understanding. But most of the speakers, maybe they were just one or two words that they didn’t say correctly, but it’s fine. I could still understand them. (K-Interview)

Many raters pointed out specific features of pronunciation that they thought problematic. The phonetic features most frequently considered problematic were /θ/ and /ð/. Many raters referred to the test-takers’ substitution of /s/ and /d/ for /θ/ and /ð/, and considered them as pronunciation difficulties or errors (Extracts 72-74). Those who pointed out the pronunciation of these distinctive sounds were mostly ESL and ENL raters; EFL raters rarely made reference to these pronunciation features. The raters’ comments generally indicated that they did not misinterpret words due to the test-takers’ pronunciation of these distinctive sounds. In particular, substitution of /d/ for /ð/ was mostly observed in the pronunciation of definite articles and pronouns (the, this, that) and did not cause serious misunderstanding. Five raters pointed out that CET B’s pronunciation of /θ/ made the word health sound like hairs. However, all of them correctly interpreted the word through context (Extract 75). Several comments referred to other distinctive sounds including /f/ and /p/; /w/ and /v/; and /l/ and /r/ (see Appendix 9).

“Something” sounds like somesing. (V-C-SR-2946)

So that’s first example that she says I here sink instead of “think.”

(R-A-SR-2394)

She says da da da and da da da. / It’s not very strong pronunciation.

(W-C-SR-3123, 3124)
The raters also pointed out some problematic pronunciation for some specific vocabulary. The words most frequently mentioned were *flood(ing)* and *causes* although the pronunciation of these words did not cause any misunderstanding (Extracts 76-77). The raters pointed out some pronunciation features occurred in even high-proficiency test-takers such as Anita (CAE) and CET A. Some other vocabulary items pointed out by the raters are presented in Appendix 9.

76 I think she meant “flood” not *float*. (G-Anita-SR-918)
77 He said “causes” and sounded like *courses*. (K-A-SR-1513)

Although it was possible for the raters to understand the words, some phonological features in the output made it hard to understand what the test-takers intended to say. In some cases, the raters had to make an extra effort to comprehend the message and were distracted by particular features of the test-takers’ pronunciation (Extracts 78-79).

78 But when he says “I like” sounds a little bit more like *I rike*, which you know is not as … not as easy to understand. (V-Chulsoo-SR-3019)
79 What drew my attention was, although I understood completely, she spoke with a trill and added a vowel after “I think,” like *I thinku*. Such pronunciation distracts me a little bit, / but it doesn’t impede my understanding. (E-C*-SR-538, 539, 540)
A few raters referred to their familiarity with the test-takers’ pronunciation or their own English learning experience, which was considered to affect comprehensibility (Extracts 80-81). These comments suggest that the raters did not necessarily attribute the incomprehensibility of the utterance to the test-takers’ pronunciation. The raters’ characteristics contributed to their understanding of pronunciation.

80 So it could be that if I talk to him for a long period of time I might come to understand how he pronounces words differently how I expect them to.
   (T-D-SR-2645)

81 What impressed me on the two last sentence is that he use … he had the same mistake on pronunciation that I used to make. / That’s why I can easily understand what he say. (B-A-SR-205, 206)

Figure 4.7 illustrates the relationships between pronunciation and other features. The speakers’ pronunciation affected the raters’ judgments of communication ability and the comprehensibility of the message. At the same time, problematic pronunciation did not necessarily negatively affect the judgments of communication ability; unintelligible pronunciation negatively affected the judgments. The raters’ familiarity with the speakers’ pronunciation facilitated comprehensibility.
4.6.2 Accent

The raters referred to the test-takers’ L1 accent in their output (accounting for 1.9% and 1.5% of the comments in the CET-SET and the Cambridge Exams). While accent is considered as part of pronunciation, the raters made explicit comments on pronunciation influenced by the L1 of the test-takers. Even though the test-takers’ L1 accent was recognized, it was not always perceived negatively. The raters often guessed the test-takers’ L1 background from their accent (Extracts 82-83).

82  I think it seemed to me that it was a French accent, (F-Marie-SS-3902)
83  It’s easy to recognize that that is a Korean accent because I think there are problems with /ɛ/. (D-Chulsoo-SR-479)

In a few cases, the raters made positive comments on accented English as long
as it was clear and intelligible. Such accent was perceived easily to understand (Extract 84). In contrast, strong accent was evaluated negatively in most cases because it hampered the comprehensibility of the message and sounded that the test-takers were speaking their L1 (Extracts 85-86). These comments indicate that the judgment of accented English depended on its strength and intelligibility.

84 and I found their accent quite easy, yeah to understand the words that they were trying to express. (T-Aida/Chloe-SS-5314)

85 She also has, in my opinion, a pretty strong Asian accent, which makes it hard in this case for me to understand what she is actually saying. (F-C⁴-SS-3881)

86 Again, I sometimes feel like she speaks in Chinese as well. (B-A⁴-SS-3429)

The raters’ comments also indicated that their familiarity with the test-takers’ L1 accent affected the comprehensibility of the message. Strong accent did not seem to negatively affect those who were accustomed to the accent and able to comprehend the message. For example, Rater D, who had been learning Mandarin Chinese for 12 years, understood CET C⁺’s strong accent (Extract 87). Additionally, Rater E mentioned that accent did not draw her attention because she was used to strong foreign accents (Extract 88).

87 however, she had a very strong Chinese accent. / I … personally I could understand it (D-C⁺-SS-3668, 3669)

88 Basically they didn’t catch my attention overall. Their pronunciation or
accent didn’t impede my understanding. There are people who speak with a strong accent and I am used to it. [Researcher: You’ve heard many speeches with Chinese accent. Did it affect your judgments negatively?] Not particularly. (E-Interview)

With regard to the importance of accent, five raters (Raters F, I, K, O, and S) stated that accent was one of the three features affecting their overall judgments of communication ability the most. As stated above, accent might strongly and negatively affect their judgments only when it hampered understanding and imposed a heavy burden on the raters (Extract 89). At least, the raters did not necessarily idealize Standard English or ENL accents and seemed to accept L1 accent of non-native speakers. Rater U (from New Zealand) and Rater H (from China) stated their views on accent (Extracts 90-91).

89  And the third thing I would say is maybe the accent. But I find you could mostly understand what they were saying from the context but that is something which affects whether you can understand somebody properly. (S-Interview)

90  So yeah, accents, everybody has got them. I apparently have a strong one but it doesn’t bother me, so yeah. Yeah, it doesn’t affect whether they can communicate or not. (U-Interview)

91  It is normal to have accent for non-native speakers, (H-C^+SS-4145)

Figure 4.8 illustrates the relationships between accent and other features. The speakers’ accent affected the raters’ judgments of communication ability and the
comprehensibility of the message. In addition, the raters often guessed the speakers’ L1 background based on their accent. Nevertheless, accent did not necessarily negatively affect the judgments of communication ability as long as it was intelligible. Strong accent negatively affected the judgments. The raters’ familiarity with the speakers’ accent facilitated the comprehensibility of the message.

![Figure 4.8. Relationships between accent and other features.](image)

4.6.3 Prosody and Paralanguage

Comments referring to prosody and paralanguage were 4.1% and 4.8% of all the comments given to the CET-SET and Cambridge Exams test-takers, respectively. This subcategory included prosodic features: stress, intonation, and rhythm. It also contained paralinguistic features including voice dynamics (e.g., loudness, voice quality) and articulation.

The raters commented on how the test-takers stressed or emphasized key words
in utterances. Stressing key words was evaluated positively although an extra stress on words that did not deserve an emphasis was rated negatively (Extracts 92-93). Rater B, who made a large number of comments on stress, often claimed that stress affected her understanding of the main points acknowledging its importance in communication. Some raters also noted lexical stress. The primary stress of desert (/dɛzərt/) uttered by Anita (CAE) pronouncing as dessert (/dɛzərt/) was the most frequently mentioned one. Five raters pointed it out although all of them understood her intention to pronounce desert through context (Extract 94). Therefore, idiosyncratic lexical stress did not cause any serious misunderstanding.

92 But she did a good job to stress the word in the sentence / like when she wanted to emphasize some keyword, some main point in the sentence, she raised her voice / and it made me easy to follow. (B-A+SS-3435, 3436, 3437)

93 She says first of all in the beginning / and then her pitch is a little bit off for the way that she says first of ALL. / She is emphasizing words that I don’t believe need emphasis. (W-A+SS-5558, 5559, 5560)

94 There she says dessert instead of “desert,” which changes obviously the meaning of what she’s saying, (S-Anita-SR-2565)

Intonation was another prosodic feature mentioned by the raters. Using appropriate intonation was evaluated highly. The raters also considered that the conveyance of message and emotion was facilitated by changing certain intonation properly in the speech (Extracts 95-96). In contrast, a lack of intonation was often perceived negatively as flat or monotonous intonation and considered to make the
speech less attractive and hamper understanding (Extract 97). This problem was found even in the performance of high-proficiency test-takers (e.g., CET A\textsuperscript{+}, Anita). Three raters (Raters A, I, and T) referred to intonation as one of the most influential features in their judgments of communication ability. Rater P also argued for the importance of intonation in communication in general.

95 And again you can hear he goes down in the end of his sentence so that like the intonation of every sentence is quite good and easy to follow.

(T-A-SR-2598)

96 Chloe sort of demonstrates a bit of voice modulation just to show off what she feels about each of the jobs. (P-Chloe-SR-2262)

97 She is very monotonous, which makes this speech slightly boring and less attractive. (F-A\textsuperscript{+}-SS-3852)

There were only two segments referring to the rhythm of test-takers’ spoken output. Rhythm did not seem to draw strong attention from the raters.

Voice dynamics and articulation were large components of this subcategory. The raters frequently commented on the loudness of voice. Speech with a loud and clear voice was highly evaluated, whereas mumbling or speaking in a low voice was perceived otherwise (Extracts 98-99). The quality of voice was also mentioned, such as clear, strong, soft, sleepy, nasal, trembling, and settled voices. The voice quality gave the raters impressions of comfort, nervousness, and confidence (Extract 100). Accordingly, a clear and loud voice not only facilitated understanding but also led to a positive evaluation of demeanor. Unclear articulation seemed to make the speech more slurred and make the boundary between words unclear. Eight raters stated that voice
dynamics and articulation were important and strongly affected their judgments. This indicated that audibility through clear articulation was, among prosodic and paralinguistic features, the most decisive factor in communication ability (Extract 101).

98 The thing that really stands up very, very starkly to me is the way they project their voice. / Anita projected very well, / I can hear it very clearly. (M-Anita/Marie-SS-4637, 4638, 4639)

99 and she spoke quite quietly and didn’t really project her voice that well, which hindered understanding. (S-C-SS-5214)

100 You can also compare here that Anita has a more clear voice, which gives me the impression that she’s more self-confident whereas Marie was more soft and gave me the impression that she wasn’t sure if she wanted to say that. (O-Anita/Marie-SR-2117)

101 But I think the one that really upsets me also is sort of their projection. You don’t mumble your words, because if you mumble your words, I can’t hear. [...] Yeah clarity of voice is so important, projection of voice.

(M-Interview)

Figure 4.9 illustrates the relationships between prosody/paralanguage and other features. Prosody and paralanguage affected the raters’ judgments of communication ability and demeanor. They also influenced comprehensibility of message and the degree of the raters’ attention to the speech.
4.6.4 Differences among Proficiency Levels

The number of segments on Pronunciation given to each CET-SET test-taker was examined (Table 4.14). The number of positive comments given to CETs A and B$^+$ was larger than that of negative comments. In contrast, the four low-proficiency test-takers (B, C$^+$, C, and D) received a larger number of negative comments than positive ones. Surprisingly, CET A$^+$, whose communication ability and English proficiency were evaluated most highly, received more negative comments than positive comments (25 and 36). This suggests that the number of comments on pronunciation was not necessarily related to raters’ judgments of communication ability and English proficiency as measured by the test. The negative evaluation of CET A$^+$ included the pronunciation of /θ/ and /ð/, stressing unnecessary words, monotonous intonation, and her L1 accent. However, no raters complained that her pronunciation was unintelligible. In contrast, CET D’s pronunciation seemed to seriously hamper his message conveyance; the raters frequently mentioned that his pronunciation and accent made it
hard to understand what he was trying to say.

Table 4.14

<table>
<thead>
<tr>
<th></th>
<th>A⁺</th>
<th>A</th>
<th>B⁺</th>
<th>B</th>
<th>C⁺</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>25</td>
<td>21</td>
<td>53</td>
<td>8</td>
<td>8</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>36</td>
<td>18</td>
<td>17</td>
<td>49</td>
<td>65</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>Neither</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>44</td>
<td>73</td>
<td>62</td>
<td>81</td>
<td>55</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 4.15 shows the number of segments on pronunciation given to the Cambridge Exams test-takers. Except for Chloe (FCE), all the test-takers received a larger number of negative comments than positive ones. This finding also suggests that the number of comments on pronunciation was not necessarily aligned with judgments of communication ability and English proficiency. Anita (CAE) obtained the largest number of comments on her pronunciation, and she had more negative comments regarding her pronunciation of some individual words (flood, pets, kids), lexical stress (desert), and flat intonation. In contrast, negative evaluations given to Marie (CAE) widely varied, including her pronunciation of /θ/ and /ð/, pronunciation of some individual words (flooding, tropical), a lack of stressing words, mumbling, voice quality (soft, small, trembling), and L1 accent.

Table 4.15

<table>
<thead>
<tr>
<th></th>
<th>CAE</th>
<th>FCE</th>
<th>PET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anita</td>
<td>Marie</td>
<td>Aida</td>
</tr>
<tr>
<td>Positive</td>
<td>28</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Negative</td>
<td>36</td>
<td>44</td>
<td>22</td>
</tr>
<tr>
<td>Neither</td>
<td>6</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>63</td>
<td>38</td>
</tr>
</tbody>
</table>
4.6.5 Differences among Raters

Table 4.16 shows the number of segments on *Pronunciation* provided by each rater. The frequency widely varied among raters. Pronunciation accounted only for 1.5 to 1.9% of the comments made by Raters C and Q, but more than 30% of the comments made by Raters B and R. Raters who made a relatively large number of references to pronunciation (Raters B, D, E, R, and T) also mentioned that pronunciation (including accent, prosody, and paralanguage) was one of the most influential features in their judgments. On the other hand, some raters’ comments showed that the frequency did not necessarily reflect how influential pronunciation was. In fact, Rater R frequently pointed out the test-takers’ use of distinctive sounds (e.g., /ð/ and /d/) the most, but these features did not necessarily affect her judgments of communication ability negatively (Extract 102).

102 [Researcher: Did their use of “th” sounds or so irritate or bother you?]

   No, it’s just something I picked up on. Yeah, just something I noticed.
   (R-Interview)

Table 4.16

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>26</td>
<td>95</td>
<td>4</td>
<td>31</td>
<td>64</td>
<td>23</td>
<td>25</td>
<td>26</td>
<td>50</td>
<td>23</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>%</td>
<td>6.4</td>
<td>36.5</td>
<td>1.9</td>
<td>15.1</td>
<td>20.8</td>
<td>10.2</td>
<td>6.4</td>
<td>7.4</td>
<td>14.0</td>
<td>11.1</td>
<td>8.7</td>
<td>9.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rater</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>32</td>
<td>9</td>
<td>6</td>
<td>17</td>
<td>3</td>
<td>70</td>
<td>6</td>
<td>45</td>
<td>12</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>%</td>
<td>13.9</td>
<td>3.2</td>
<td>3.5</td>
<td>6.0</td>
<td>1.5</td>
<td>30.3</td>
<td>4.1</td>
<td>18.8</td>
<td>3.9</td>
<td>13.4</td>
<td>8.1</td>
</tr>
</tbody>
</table>

4.7 Linguistic Resources

The raters referred to linguistic resources in their verbal reports. Linguistic
resources included (a) vocabulary and wording, (b) grammar, and (c) other linguistic resources. Comments on this category accounted for 10.4% of the comments on the CET-SET test-takers and 12.7% of the comments on the Cambridge Exams test-takers (Table 4.17). This category was the sixth largest category in the CET-SET and the third largest category in the Cambridge Exams. There was a statistically significant difference in the frequency with which linguistic resources were mentioned between the tests ($\chi^2(2) = 19.814, p < .01$); specifically, in the frequency of comments on grammar.

Table 4.17

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET</th>
<th>Cambridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>4.1 Vocabulary &amp; Wording</td>
<td>188</td>
<td>5.0</td>
</tr>
<tr>
<td>4.2 Grammar</td>
<td>158</td>
<td>4.2</td>
</tr>
<tr>
<td>4.3 Other</td>
<td>45</td>
<td>1.2</td>
</tr>
<tr>
<td>Linguistic Resources Total</td>
<td>391</td>
<td>10.4</td>
</tr>
</tbody>
</table>

4.7.1 Vocabulary and Wording

Comments on vocabulary and wording accounted for 5.0% and 4.8% of the comments given to the CET-SET and Cambridge Exams test-takers, respectively. The raters commented on the test-takers’ lexical choices, sophistication of vocabulary, and range of vocabulary.

Lexical choices were one of the features within this subcategory mentioned by the raters. They remarked the accuracy, precision, and appropriateness of lexical choice in test-takers’ output. In most cases, their comments on lexical choices were negative, and the raters provided with an alternative to express the intended message more precisely (Extracts 103-104). As these extracts show, the test-takers’ lexical choices did not cause serious misunderstanding or incomprehensibility since the raters were able to
interpret what the test-takers intended to say through context. Negative comments were
given not only to low-proficiency but also high-proficiency test-takers’ outputs. Usage
of vocabulary or phrases was also pointed out (Extract 105). Those who pointed out
test-takers’ lexical choice were, in most cases, the ESL and ENL raters. More comments
on lexical choices are presented in Appendix 10.

103 What’s a civilized citizen? / Probably he says like “big cities” or “urban
areas.” / Probably he was thinking about that, not civilized citizen. /
((laughs)) I believe that over 90% of population is civilized. (G-A-SR-777,
778, 779, 780)
104 Yeah, human body. / I don’t think it’s the human body. / It’s just “human”
or “the human factors.” (M-C-SR-1852, 1853, 1854)
105 The first of the consequences is well of course. / Of course what? / I mean,
“Of course” is said when something is evident. / So you cannot have it as a
given. (G-B\textsuperscript{+}-SR-799, 800, 801, 802)

The raters also occasionally referred to the sophistication of vocabulary or
phrases observed in the test-takers’ performances. They evaluated this aspect of
vocabulary commenting on whether vocabulary was general, basic, simple, descriptive,
advanced, and complicated (Extract 106). Some raters pointed out that the test-takers
used only simple adjectives to describe their feelings (Extract 107). The raters also
mentioned the test-takers’ frequent use of demonstrative pronouns in the paired
interactions (Extract 108). In CAE and FCE, the test-takers occasionally pointed to
some pictures by saying “this” or “that” without using the word of the referent. This
seemed to give the raters the impression that the test-takers did not know the word.
She’s talking about how something is *illustrated*, how she would *prefer* things for so and so reasons. / She’s definitely using a lot more descriptive words than Chloe has. (N-Aida-SR-2032, 2033)

I think adjectives are the most important. Because I noticed a lot of them would say *good, nice* and just would repeat those adjectives or say *very good, really nice* instead of just using other words to describe something (V-Interview)

But when she starts referring everything as *this this*, not basically means that she’s not being really able to explain the situation, like put the word on that weather. / I think that is the biggest issue with me I guess. (P-Marie-SR-2233, 2234)

The range of vocabulary was another aspect within this subcategory mentioned frequently. The raters referred to the breadth of vocabulary or the number of lexical types used in the performances (Extracts 109-110). From the range of vocabulary used in performances, they made an inference about the test-takers’ vocabulary knowledge (Extract 111). The repetitive use of the same words or phrases was one of the factors in judging vocabulary knowledge (Extract 112). Repeating the same vocabulary or phrase in a speech or interaction often distracted the raters from the content.

and all the way through they used very few words. (D-Aida/Chloe-SS-3715)

So Asad has get use of a lot of different words (W-Asad-SR-3257)

Yeah Chloe is not having sufficient vocabulary to basically support her original idea which was to contradict at least. / And she is basically lacking
vocabulary (F-Chloe-SR-732, 733)

112 His vocab was clearly limited because he would return to talking about the
gas or gases which would harm the environment, (V-B-SS-5475)

A lack of vocabulary knowledge was attributed to the test-takers’ difficulties in message conveyance and content elaboration (Extract 113). Furthermore, it was perceived to be a reason for dysfluent utterances (Extract 114). The raters felt that the test-takers’ lack of vocabulary made the utterances less fluent because they required more time to retrieve vocabulary to express their thoughts.

113 He has a pretty poor vocabulary, which towards the end of the presentation makes it hard for him to elaborate and to … he had to elaborate on his arguments. (F-B-SS-3873)

114 It seems this guy has, for me, it seems he has a problem with vocabulary that is what … that is affecting him when he is speaking in terms of his fluency in the language. (D-B-SS-3652)

With regard to the importance of vocabulary and wording, two raters (Raters P and W) mentioned that it was one of the most decisive factors in their judgments. However, the raters’ perception of its importance varied. Some raters (Raters A, D, I, J, and Q) stated that vocabulary considerably affected their impressions (more than grammar) whereas other raters (Raters B, F, G, H, O and R) did not consider it salient.

Among the three aspects of vocabulary (choices, sophistication, and range), the range of vocabulary used or vocabulary knowledge seemed to be a salient factor in raters’ judgments of communication ability. Many raters mentioned that a lack of
vocabulary knowledge would seriously impede the conveyance of message or emotion (see Extract 113 above). Repeating the same words was also considered to be partly due to poor vocabulary knowledge. Since the range of vocabulary was considered closely related to the ability to express the message, the raters probably considered it salient compared with other aspects of vocabulary. In fact, even though the test-takers could not precisely utter the word of the referent, conveying their intended message using circumlocution was evaluated positively (Extract 115).

115 That sometimes you could understand what they were saying even if they didn’t have the vocabulary. I am thinking of those two girls who were talking about the jobs. Sometimes they would do a description of the job rather than use the word. So in that sense vocabulary wasn’t an issue.

(S-Interview)

In contrast to vocabulary range, sophistication and lexical choices were considered peripheral. Some raters acknowledged that sophisticated vocabulary gave them a better impression and positively affected their judgments of communication ability. However, it was often claimed that message conveyance should be prioritized rather than using low-frequency sophisticated vocabulary (Extract 116). Lexical choices were considered to affect overall judgment mostly when it would seriously impede precise message conveyance (Extract 117).

116 I mean especially … obviously better if you can be more specific and use upgrade a variety of words. But I think the first step is just having any word to say what you want to say and then. (C-Interview)
Yeah, it depends on the context actually. As in, if *human body* is going to be a keyword, yes I am going to assess the plot very badly but if it’s not going to be a keyword, I am probably not going to assess, I am just going to let it pass. (M-Interview)

Figure 4.10 illustrates the relationships between vocabulary/wording and other features. The raters perceived that knowledge of vocabulary affected fluency, message conveyance, and elaboration of content. Vocabulary used in the performance also affected comprehensibility of the message and the judgments of communication ability and English proficiency. Lexical choices that made the message incomprehensible particularly affected the raters’ judgments of communication ability.

![Diagram showing relationships between vocabulary/wording and other features](image-url)

*Figure 4.10. Relationships between vocabulary/wording and other features.*
4.7.2 Grammar

Comments on grammar accounted for 4.2% and 7.2% of all of the comments given to the CET-SET and Cambridge Exams test-takers, respectively. The raters referred to grammatical aspects of the performance in the paired interactions more frequently than in the individual presentations. Comments within this subcategory can be divided into grammatical accuracy and complexity.

The raters frequently commented on the accuracy of grammatical forms in the performance. They mentioned their global impressions of the test-takers’ grammatical accuracy (Extract 118). In addition, the raters noted specific grammatical features that they thought problematic; such features were called mistakes, errors, bad constructions, less structured sentences, improper usages, and problems. The types of grammatical feature most frequently noted were singular/plural marking, subject-verb agreement, word order, the use of pronouns and articles, and omitted parts of speech (Extracts 119-123). Features mentioned less frequently were double negative, the wrong use of tense, pluralizing uncountable nouns, and so forth (see Appendix 11). The raters’ negative comments on grammatical accuracy were given not only to low-proficiency test-takers but also to high-proficiency test-takers.

118 Although I still think that, you know, his sentence structure has room for improvement. (O-Asad-SS-4859)

119 So he said all the house are too big when it should be “all the houses are too big.” (R-Asad-SR-2486)

120 Here she says affect the way people lives instead of “the way people live.” (R-Anita-SR-2457)

121 Marie says it can affect a lot your mood. / That’s not something that
grammatically doesn’t make sense in my opinion. (W-Marie-SR-3167, 3168)

122 He said *you can’t meet your family all the day* when it probably should be “you can’t meet your family all day.” (R-Asad-SR-2487)

123 Yeah, *many many people concern. / “ARE concerned.” /* So she does a few mistakes various points of the speech. (F-C-SR-696, 697, 698)

Those who observed specific grammatical features were mostly Raters F, G, R, V, and W (two ESL raters and three ENL raters); other raters rarely made references to specific grammatical accuracy even though they commented on the global grammatical accuracy of the performances. Rater H (Chinese, 6.5 in IELTS) admitted that she did not notice any grammatical mistakes from CET A+.

Even though some raters noticed particular grammatical features as errors or problems, many of them were perceived minor in the sense that they did not impede comprehensibility. Such grammatical features were often referred to as *minor errors* (Extract 124). In contrast, some grammatical features including incomplete sentences and wrong word order (Extracts 125-126) impeded the raters from interpreting the message correctly.

124 So *lives* instead of “live.” / Just a minor thing. (V-Anita-SR-2983, 2984)

125 so they weren’t in sentences but just the fragmented ideas, which made it harder to understand. (R-D-SS-5131)

126 he would start sentences with verbs sometimes, which just meant well, kind of signified that he was well, obviously English was not his first language / and it made his message harder to get across I think.
Grammatical accuracy directly influenced the raters’ judgments of communication ability and English proficiency. Raters K, S, and W claimed that inaccurate grammar was a direct reason that they did not rate Anita (CAE) as Excellent.

There was a large difference in the perceived importance of grammatical accuracy among the raters. Nine raters (Raters E, H, I, J, L, M, N, P, and T) thought that it was a peripheral factor in their judgments, whereas four raters (Raters F, R, V, and W) perceived that accuracy largely affected their impressions. In fact, many raters explicitly commented that grammatical accuracy was not a primary factor, so long as the test-takers were able to convey the message (Extract 127). This indicated that the seriousness of grammatical errors seemed to affect their judgments, and minor errors were likely to be overlooked (Extract 128). However, the perceived seriousness of errors was different among raters. For example, Rater W considered error in singular/plural marking as “a big issue” although other raters tended to perceive it minor. Furthermore, for some raters (in particular the ENL raters), minor grammatical issues appeared to be distracting (Extract 129).

127 I mean it’s obviously important to speak correctly, but I think it’s probably more important just to be able to make a point or two to be understood or at least that’s the first step and then maybe the grammar can be improved, … (C-Interview)

128 I would say strong when a mistake was obvious. I mean there are some minor mistakes like saying “person” instead of “persons.” You don’t really notice that or you let it pass when the rest of the English is correct.
129 … sometimes even if you are not completely grammatically accurate you can kind of guess where they are getting at. But it does require a bit more mental effort. So in that sense it’s harder to understand. (S-Interview)

Some raters made reference to the relationships between grammar and other features. Raters E and F stated that utterances would be incomprehensible if grammatical errors were accompanied by dysfluency and strong L1 accent (Extract 130). In this sense, grammatical accuracy was not necessarily the sole factors affecting comprehensibility. Similarly, inaccurate grammar might not be critical if the test-takers were able to maintain fluency and intelligible pronunciation (Extract 131).

130 If a speaker can speak sentences smoothly, I can understand even though grammar is wrong. But if a speaker speaks with pauses in the middle of sentence and with incorrect grammar, it is hard to follow what she is talking about. (E-Interview)

131 … the attempts to be accurate hindered my ability to understand. Because they were thinking so much about it and umming and erring that the message got a bit lost. I think I prefer bad grammar if it’s fluent than slow convoluted speech that’s grammatically correct. (T-Interview)

The raters’ expectations of accuracy also seemed to influence their attitudes to inaccurate grammatical structures. Some raters (Raters L, M, N, and S) claimed that they tolerated a lack of accuracy because the speakers were NNESs.

The second aspect of grammar mentioned by the raters was grammatical
complexity. The frequency of grammatical complexity was lower than that of grammatical accuracy. The raters commented on the complexity of sentence structures by pointing out that a sentence was *simple, short, complex, long, complicated,* or *confusing* (Extract 132). Simple sentence structures helped some raters understand the message but were not necessarily perceived positive (Extract 133).

132 Even though Asad (didn’t) you know use the complicated sentence as well, he try to make every information that he want to say in a simple sentence as possible. (B-Asad-SR-295)

133 And he does not try to build complex sentences, which makes it easy to understand but which demonstrates I guess pretty poor communication ability because his thought have to be much more elaborate, much more complex than what he actually is able to say. (F-Chulsoo-SR-757)

Figure 4.11 illustrates the relationships between grammar and other features. Grammatical forms in the performance affected comprehensibility and the judgments of communication ability and proficiency. Grammatical structures impeding comprehensibility were particularly perceived negatively. Raters’ expectation of accuracy influenced the extent to which they tolerated grammatical errors.
4.7.3 Other

Features categorized into other linguistic resources accounted for 1.2% and 0.8% of all the comments given to the CET-SET and Cambridge Exams test-takers, respectively. Linguistic features within this subcategory included discourse markers, sentence variations, circumlocution and paraphrasing, and sentence characteristics that could not be categorized into either vocabulary or grammar.

The most frequent comments were on discourse markers used in the individual presentations indicating the sections of the speech. The use of discourse markers was rated positively since it gave the raters the impression that the speech was well organized (Extract 134). Some raters also referred to sentence variations to express the speakers’ thoughts. Reiterating the speakers’ thoughts with different expressions was perceived positively (Extract 135). The raters also highly evaluated the test-takers’ attempts to express their thoughts using various expressions when they got stuck with
their speech.

134 I noticed that she used phrases like *first of all, second of all* gives me the impression that she organizes her thoughts very well. (O-A’-SR-2071)

135 He’s able to use a few different ways to describe what he thinks the government should do, / so that he says “the responsibility is tremendously big.” / He also said “they should take more strenuous efforts.” / So be able to describe in two different ways and emphasize his point. (Q-A-SS-2310, 2311, 2312, 2313)

### 4.7.4 Differences among Proficiency Levels

Table 4.18 shows the number of segments on *Linguistic Resources*—vocabulary and wording, grammar, and other—given to each CET-SET test-taker. The positive and negative comments showed that the raters’ judgments of linguistic resources seemed to be related to the CET-SET scores. The two most highly proficient test-takers (A’ and A) received more positive comments, whereas the other lower-proficiency test-takers received more negative comments. The raters made fewer comments on linguistic resources of CET D compared with the other speakers (18 segments).

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>A’</th>
<th>A</th>
<th>B’</th>
<th>B</th>
<th>C’</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>27</td>
<td>35</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Negative</td>
<td>12</td>
<td>30</td>
<td>44</td>
<td>42</td>
<td>39</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td>Neither</td>
<td>5</td>
<td>14</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>79</td>
<td>59</td>
<td>68</td>
<td>58</td>
<td>65</td>
<td>18</td>
</tr>
</tbody>
</table>
Table 4.19 shows the number of segments on linguistic resources given to each Cambridge Exams test-taker. Except Anita (CAE), all the test-takers received a larger number of negative comments than positive ones. That is, linguistic resources of even intermediate- to high-level English speakers (Marie, Aida, and Chloe) might be perceived negatively overall. Furthermore, even though Asad’s (PET) communication ability was perceived the second highest following Anita, his linguistic resources were evaluated quite negatively. The raters negatively commented on Asad’s linguistic resources most frequently (53 segments).

Table 4.19

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>CAE</th>
<th>FCE</th>
<th>PET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anita</td>
<td>Marie</td>
<td>Aida</td>
</tr>
<tr>
<td>Positive</td>
<td>28</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>Negative</td>
<td>23</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td>Neither</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>43</td>
<td>63</td>
</tr>
</tbody>
</table>

4.7.5 Differences among Raters

Table 4.20 presents the number of segments on linguistic resources provided by each rater. The frequency of mentions widely varied ranging from 8 (Rater Q: 4.1%) to 75 (Rater R: 32.5%). Raters F, G, R, V, and W made a relatively large number of references to linguistic resources. They pointed out specific lexical and grammatical features as problematic. Other raters rarely mentioned them unless the features impeded comprehension.
Table 4.20  
**Number of Segments on Linguistic Resources Provided by Each Rater**

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>57</td>
<td>31</td>
<td>38</td>
<td>51</td>
<td>12</td>
<td>34</td>
<td>66</td>
<td>19</td>
<td>10</td>
<td>15</td>
<td>13</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14.0</td>
<td>11.9</td>
<td>18.4</td>
<td>24.9</td>
<td>3.9</td>
<td>15.1</td>
<td>17.0</td>
<td>5.4</td>
<td>2.8</td>
<td>7.2</td>
<td>8.1</td>
<td>9.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rater</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>17</td>
<td>17</td>
<td>28</td>
<td>8</td>
<td>75</td>
<td>14</td>
<td>13</td>
<td>11</td>
<td>66</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.3</td>
<td>6.1</td>
<td>9.9</td>
<td>9.9</td>
<td>4.1</td>
<td>32.5</td>
<td>9.6</td>
<td>5.4</td>
<td>6.5</td>
<td>24.6</td>
<td>15.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**4.8 Fluency**

The rater made references to fluency of the test-takers’ utterances, including (a) overall fluency, (b) rate and amount of speech, (c) pause phenomena, and (d) repair phenomena. Fluency in the individual presentations appeared to draw the raters’ attention more than that in the paired interaction (Table 4.21). Comments on this category accounted for 13.5% and 6.6% of all the comments given to the CET-SET and Cambridge Exams test-takers, respectively. It was the third and eighth largest category in each test. There was a statistically significant difference in the frequency of the segments between these two tests ($\chi^2(3) = 34.131, p < .01$). In particular, a significant different was found in repair phenomena.

Table 4.21  
**Frequency of Segments on Each Subcategory of Fluency**

<table>
<thead>
<tr>
<th></th>
<th>CET-SET</th>
<th>Cambridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>5.1 Overall fluency</td>
<td>56</td>
<td>1.5</td>
</tr>
<tr>
<td>5.2 Rate &amp; Amount of speech</td>
<td>65</td>
<td>1.7</td>
</tr>
<tr>
<td>5.3 Pause phenomena</td>
<td>300</td>
<td>8.0</td>
</tr>
<tr>
<td>5.4 Repair phenomena</td>
<td>86</td>
<td>2.3</td>
</tr>
<tr>
<td>Fluency Total</td>
<td>507</td>
<td>13.5</td>
</tr>
</tbody>
</table>
4.8.1 Overall Fluency

The raters commented on global fluency in the performances (1.5% and 0.9% in the tests, respectively). Comments within this subcategory did not show any specific features of fluency but simply indicated the raters’ overall impressions of fluency. The raters often mentioned if the speaker speaks fluently, effortlessly, smoothly, fluidly, continuously, or with ease (Extracts 136-137).

136 But she spoke smoothly, (E-A^+SS-3737)
137 and in general his language doesn’t flow very well. (S-B-SS-5199)

Some raters claimed that overall fluency was important and largely affected their judgments. Two raters (Raters B and D) mentioned that fluency was one of the three features that most strongly influenced their judgments, and three raters (Raters J, Q, and S) acknowledged its saliency in their impressions of the speakers (Extract 138). Some raters (Raters E and M) implied that fluency helped to maintain their concentration on the speech and understand the content better (Extract 139).

138 I would say a lot because I think you can give the impression of being a very good communicator through fluency even if what you are saying is not necessarily grammatically yours and tactically correct. Yeah, so I think fluency is a big factor in communication. (S-Interview)
139 Without fluency, I have to make an effort to understand the speech and can’t concentrate on the content. (E-Interview)
4.8.2 Rate and Amount of Speech

The raters referred to the rate and amount of speech when commenting on how fast and how much the test-takers spoke within the given time. The comments within this subcategory accounted for 1.7% and 1.3% of the comments in the CET-SET and the Cambridge Exams, respectively.

The direction of their evaluations of the rate and speed was not consistent. Fast utterances were not always perceived positively or negatively (Extracts 140-141). Instead, the raters seemed to appreciate a moderate speech rate.

140 She is too fast. / She is like singing. (I-C -SS-4268, 4269)
141 She was obviously very slow in what she said, (R-C-SS-5124)

Rater I (Indonesian; 7.0 in IELTS) most frequently commented that test-takers spoke too fast and perceived it quite negatively (see Extract 140 above). Other raters including ESL and ENL raters (Raters F, N, and T) also mentioned that some test-takers’ utterances were too fast. The fast rate of speech often hindered the raters’ comprehension (Extract 142). However, the rate of speech was not the sole problem for comprehensibility. Rater F pointed out that some test-takers spoke fast in spite of their low English proficiency. Their attempt to maintain a fast pace had a detrimental effect because it made the test-takers hesitate and stutter (Extract 143).

142 the only things that I would say she could improve on was that she spoke quite quickly, which made it difficult to understand sometimes
(T-C -SS-5282)
143 He should … he should speak much more slowly because it’s … the fact
that he wants to speak too fast is makes him sort of hesitate on a few words and makes him resay again the the government, the the, should should.

(F-A-SR-664)

In contrast, the slow utterances speed distracted the raters from the content and made conversation less interesting. Raters E and K commonly mentioned that Aida and Chloe (FCE) spoke too slowly, leading to negative evaluations of the speakers’ communication ability (Extract 144). Slow utterances appeared to require an extra effort to maintain the raters’ concentration. Furthermore, slow utterances were not perceived positively even though they were grammatically accurate. Speaking with a moderate rate seemed important even if co-occurred with some grammatical errors (Extract 145). Some raters also mentioned that the rate of speech reflected confidence and anxiety. They believed that the more confident the test-takers were, the faster the speech became. The importance of rate of speech was acknowledged by five raters (Raters F, I, O, V, and U).

144 Even though they sound relaxed, I think they talk a bit slow. / So it makes it a bit less interesting to listen to them. (K-Aida/Chloe-SR-1567, 1568)

145 In terms of communication I feel the fast one and the ungrammatically correct one is better because it means your ideas are getting across faster. (R-Interview)

The rater also referred to the amount of speech within the given time, though less frequently than the rate of speech. Low amounts of speech were generally evaluated negatively. In particular, the raters seemed to have difficulty judging the communication
abilities of CET D and Chulsoo (PET) because they did not produce enough spoken output.

Figure 4.12 shows how the relationships between the rate and amount of speech and other features. The rate and amount of speech affected the judgments of communication ability, comprehensibility, and the raters’ concentration on the speech. It was believed that the rate affected the speakers’ grammaticality and reflected confidence and anxiety.

Figure 4.12. Relationships between the rate/amount of speech and other features.

4.8.3 Pause Phenomena

Within the fluency category, pause phenomena were the features mentioned most frequently. They accounted for 8.0% and 4.3% of all the comments given to the CET-SET and Cambridge Exams test-takers, respectively. Pause phenomena particularly drew the raters’ attention in the individual presentations. This subcategory included
unfilled pauses and filled pauses.

The raters often indicated unfilled pauses or silences in the performance. They described these phenomena as *pauses*, *silences*, *interruptions*, and *stops* and evaluated them quite negatively in most cases. In particular, the raters pointed out prolonged and frequent pauses in the middle of a sentence. Stimulated recall data revealed that pauses prolonged for at least 2.0 seconds were perceived long (Extract 146: the italicized texts are utterances of the test-taker). Rater A stated that he lost the whole meaning of utterance when pauses lasted approximately 4.0 seconds. In contrast, pauses less than 1.0 second were not perceived as long or not even commented on by the raters; Rater A considered 0.2 to 1.0-second unfilled pauses as *minor silences* (Extract 147).

146 [CET C: *will err (1.0) harm our (.2) health? err (1.0) it can: (1.5) it will it will mm (4.0) (touches hair) I think it will err (3.0) it will (1.0)]

So here is a long pause right in the middle of the video, which is … it will pause it will pause. (W-C-SR-3119)

147 [CET B: *err mm the population is very (.2) you know the world is heavily populated. and err mm err the space and err resources are limited. and err mm (1.0) it will. err (.2) in the in the life of human being. it will mm it also (.8) mm produce many mm (.8) mm many gas. which is also]*

There are some minor interruptions and some minor silences in the sentence, / but I think it’s not that long (A-B-SR-46, 47)

Pauses were perceived to be distracting if they occurred frequently. The raters felt it difficult to follow the content of speech if it contained a number of short pauses (Extract 148). Additionally, unfilled pauses gave the impression that the test-taker was
not prepared for the speech, lacked topical knowledge, or lost what to say.

148 [Aida: *tch! I think it it's a great job? yes? but err (.5) the: hhh. probably I would like to: to do? (.2) more this one? to:: to give the: *tch! the prizes to: the: (. ) champions? so (. ) because (.2) I think you can err (.2) you can see the champions err on the: on the sportive.*]

There’s also quite a few pauses between there the sentences and the words, which makes it hard. (V-Aida-SR-3002)

Although pauses were generally perceived negatively, some raters (Raters F and I) occasionally suggested that some CET-SET test-takers should have taken pauses at the end of the sentence (Extract 149). Their comments indicated that the test-takers’ attempts to avoid silence with fillers were considered negatively. Instead, it would be better to use pauses effectively and strategically. Short silence at the end of sentence seemed to be evaluated highly.

149 And she should pause a little bit, use a silence, and try to think get one or two extra seconds to think “oh well” what she wants to say to make it much clearer and much more efficient in her communication with the audience. (F-C*-SR-694)

The raters made a large number of references to filled pauses or pauses accompanied by voiced fillers. The types of fillers most frequently mentioned were non-lexical fillers such as *err, mm, uhm, um,* and *uh.* These fillers were noticed when they occurred frequently in the performance and negatively evaluated without exception
Accordingly, those who produced only a few fillers were judged positively although even high-proficiency speakers were not able to avoid filled pauses completely. The test-takers also frequently used lexical fillers, which were also evaluated negatively. The type of lexical filler most frequently noted was *you know* observed in CET A and Asad’s utterances (Extract 152). Similar to non-lexical fillers, the frequent use of lexical fillers was perceived negatively in the performance. Other types of lexical fillers mentioned were *I think, it’s like, maybe, for example,* and *I don’t know.*

150 He says sort of uhm or mm a lot, which kind of breaks up the sentences, (Q-B-SR-2326)

151 First thing, a lot of hmm, hmm, hmm, / there was no way he could complete two or three words in straight ahead. (G-D-SS-4032, 4033)

152 and he kept on saying *you know,* which is a phrase that’s distracting in any conversation to use it often. (K-Asad-SS-4471)

Both non-lexical and lexical fillers distracted the raters and made it difficult to comprehend the message (Extracts 153-154). In addition, some lexical fillers gave particular impacts on the rater’s impressions of the test-takers. For example, the excessive use of *you know* gave the impression that the speakers were constantly asking for recognition or influenced by North American culture. Similarly, the phrase *I think* made the raters feel that the test-takers were opinionated. These fillers greatly impacted overall judgments of communication ability (Extract 155). Nevertheless, some raters (Raters L and T) showed high tolerance for fillers. Although fillers were considered distracting, they were not necessarily a crucial problem (Extract 156).
153 I think fluency is quite important because a lot of them, talking about the environment, they started speaking and then they just “uh uh uh uh,” so you just get bored and lose concentration. (D-Interview)

154 I think her ums and uhs did get a little in the way of me understanding what she was saying, (T-C-SS-5294)

155 One of the reasons that I marked him down is he says *you know* and he gets into a lot of uhs multiple times. (W-A-SS-5577)

156 I didn’t mind the ums and uhs. [...] And that, yeah, overall the ums and uhs didn’t really faze me in my understanding. (T-Interview)

Pause phenomena often gave the raters the impression that the test-takers were struggling for words and considering what to say. In many cases, the raters commented that the test-takers were struggling for, searching for, or at a loss for words when they took pauses or used fillers (Extract 157). This comment was given mostly to low-proficiency CET-SET test-takers (CETs B, C\(^+\), C, and D) and the FCE test-takers (Aida and Chloe). Filled pauses were also considered one of the factors contributing to the judgments of overall English proficiency. Similarly, the raters assumed that the test-takers were thinking about what to say when they observed pause phenomena (Extract 158).

157 [CET C: *err firstly*? *is the:: human (. ) body. err* (1.0) *first*? *they: the: air pollution will err (1.0) harm our (.2) health? err (1.0) it can* (1.5) *it will will mm]*

So she pauses a lot, / … / This is I think the point when she starts to really struggle with trying to find the words to explain her point. (L-C-SR-1672,
And it becomes more obvious that she’s thinking too long what she will say. (K-C-SR-1550)

Within the fluency features, pause phenomena were considered most influential to the raters’ judgments. Eight raters (Raters H, K, M, O, Q, S, V, and W) referred to pauses as one of the most influential features (Extract 159). In addition, six raters (Raters A, C, D, G, K, and U) mentioned the saliency of pause phenomena when asked how important fluency was in their judgments (Extract 160).

First of all the hesitation. If they hesitated a lot or stopped and started and begun new sentences, that was probably the biggest thing. (V-Interview)

Yeah, I think that’s quite important. [...] Or I think lot of them stopped, stopped a lot and maybe you tend to get lost as a listener. (C-Interview)

Figure 4.13 displays the relationships between pause phenomena and other features. Pause phenomena affected the raters’ judgments of communication ability, English proficiency, preparedness, and topical knowledge. They also influenced comprehensibility and the extent to which the raters focused on the speech.
4.8.4 Repair Phenomena

Repair phenomena accounted for 2.3% of the comments given to the CET-SET test-takers. The raters did not provide any comments on these phenomena to the Cambridge Exams test-takers. Repair phenomena mentioned by the raters included repetition of part of utterances, self-correction, and stuttering.

The repetition of part of utterances was the most frequently mentioned within repair phenomena. The raters frequently commented on the continuous repetition of the same word or phrase (Extract 161). This phenomenon appeared to distract the raters from the content and was evaluated negatively.

161 [CET A⁺: unfortunately they couldn’t reach the decision. so I think uhm (.2) err the president of (xxx) the country should be should err should be aware of]

She repeated should many times. / It distracts me (E-A⁺-SR-492, 493)
The raters also commented on stuttering or repeating the first sound of words. CET B and CET C⁺ stuttered several times in their speeches, which were evaluated negatively and distracted the raters from the content (Extract 162).

162 [CET C⁺: mm we can er use err less err plast plastic plastic bags as as as]

Plast pla pla plastic bag as as as. / It’s something surely difficult to keep track of the conversation. (G-C⁺-SR-864, 865)

Another feature within repair phenomena was self-correction. The raters pointed out self-correction of the content and grammatical forms (Extract 163). They were mostly perceived negatively, as the phenomena confused the raters and made the speech sound less fluent. However, in one case, self-correction was perceived positively because it reflected the test-taker’s high grammatical knowledge (Extract 164).

163 He keeps on correcting himself / so it can get confusing like he said reduce then reduce and product and production, (K-B-SR-1528, 1529)

164 and she was able to correct herself when she started out a sentence incorrectly. / So, for example, she started out saying he said / and she corrected it to she said. (Q-A⁺-SS-5001, 5002, 5003)

4.8.5 Differences among Proficiency Levels

The number of segments related to fluency—overall fluency, the rate and amount of speech, pause phenomena, and repair phenomena—given to each CET-SET test-taker was examined (Table 4.22). Except for CET B⁺, all the test-takers received a larger number of negative comments than positive comments. In particular,
low-proficiency test-takers (CETs B, C\(^{+}\), C, and D) received quite a large number of negative comments on fluency (ranging from 61 to 91). This shows that English proficiency might be linked to the raters’ judgments of fluency. At the same time, the fluency of high proficiency test-takers (CETs A\(^{+}\) and A) was also evaluated negatively.

<table>
<thead>
<tr>
<th>Table 4.22</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequency of Segments on Fluency Given to the CET-SET Test-takers</strong></td>
</tr>
<tr>
<td>A(^{+})</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Neither</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Table 4.23 shows the frequency of segments on fluency given to the Cambridge Exams test-takers. Anita (CAE) received the smallest number of comments on fluency, and there were no negative comments on her performance. In contrast, with regard to the other test-takers, the number of negative comments was larger than that of positive comments. In particular, the fluency of the two FCE test-takers and Marie (CAE) was harshly evaluated since there were only a few positive comments. Compared with them, the PET test-takers’ fluency was more positively evaluated. These results indicated that English proficiency was not necessarily aligned with the raters’ judgments of fluency.

<table>
<thead>
<tr>
<th>Table 4.23</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequency of Segments on Fluency Given to the Cambridge Exams Test-takers</strong></td>
</tr>
<tr>
<td>CAE</td>
</tr>
<tr>
<td>An</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Neither</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
4.8.6 Differences among Raters

The number of segments on fluency provided by each rater was examined (Table 4.24). Rater C, whose comments on fluency were the largest and accounted for 23.2% of her comments, acknowledged that one of the fluency features—pause phenomena—was quite important in her judgment of communication ability (see Extract 160 above). This supported the high frequency of comments on fluency. However, Rater R, who made the least number of references to fluency, did not necessarily perceive that fluency was unimportant. She made a large number of comments on linguistic resources (accounting for 32.5% of her entire comments) but claimed that she preferred fast ungrammatical utterances to slow grammatical ones (see Extract 145 above). This adds evidence to the notion that the frequency of comments does not strongly reflect the strength of influence or the raters’ belief about the importance of communication ability features generally.

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>29</td>
<td>23</td>
<td>48</td>
<td>17</td>
<td>34</td>
<td>24</td>
<td>40</td>
<td>19</td>
<td>53</td>
<td>12</td>
<td>20</td>
<td>47</td>
</tr>
<tr>
<td>%</td>
<td>7.1</td>
<td>8.8</td>
<td>23.2</td>
<td>8.3</td>
<td>11.0</td>
<td>10.7</td>
<td>10.3</td>
<td>5.4</td>
<td>14.8</td>
<td>5.8</td>
<td>12.4</td>
<td>14.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rater</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>26</td>
<td>22</td>
<td>15</td>
<td>32</td>
<td>32</td>
<td>9</td>
<td>22</td>
<td>17</td>
<td>30</td>
<td>44</td>
<td>55</td>
</tr>
<tr>
<td>%</td>
<td>11.3</td>
<td>7.9</td>
<td>8.8</td>
<td>11.3</td>
<td>16.3</td>
<td>3.9</td>
<td>15.1</td>
<td>7.1</td>
<td>9.7</td>
<td>16.4</td>
<td>11.8</td>
</tr>
</tbody>
</table>

4.9 Content

The raters referred to the content of the test-takers’ speeches and the ideas held by the test-takers. This category was thus not directly related to the language observed in the performances or the manner of speech. More specifically, the raters mentioned (a) the ideas presented by the test-takers, (b) the framing or organization of ideas, and (c)
the test-takers’ topical knowledge (Table 4.25). Content was the second largest main category of both tests (accounting for 15.1% and 13.7% of all the comments). There was a statistically significant difference in the frequency of mentions on the subcategories between the CET-SET and the Cambridge Exams ($\chi^2(2) = 75.287, p < .01$) with ideas presented by the Cambridge Exams test-takers mentioned more frequently than statistically expected. In contrast, the framing of ideas was significantly more frequently commented on in the individual presentations than in the pair interactions.

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET #</th>
<th>CET-SET %</th>
<th>Cambridge #</th>
<th>Cambridge %</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Ideas</td>
<td>336</td>
<td>9.0</td>
<td>287</td>
<td>11.6</td>
</tr>
<tr>
<td>6.2 Framing of idea</td>
<td>126</td>
<td>3.4</td>
<td>11</td>
<td>0.4</td>
</tr>
<tr>
<td>6.3 Topical knowledge</td>
<td>103</td>
<td>2.8</td>
<td>41</td>
<td>1.7</td>
</tr>
<tr>
<td>Content Total</td>
<td>565</td>
<td>15.1</td>
<td>339</td>
<td>13.7</td>
</tr>
</tbody>
</table>

### 4.9.1 Ideas

The raters commented on the ideas presented by the test-takers most frequently (9.0% and 11.6% of the comments given to the CET-SET and Cambridge test-takers, respectively). The comments within this subcategory included the raters’ confirmatory checking of an idea, the quality of an idea, content elaboration, relevance to the topic and coherence, and content repetition.

The raters frequently confirmed or summarized the content or ideas given by the test-takers. The confirmation of the idea was the largest component of the Ideas subcategory in terms of the frequency of the segments. Although their confirmations were non-evaluative, this type of comment reflected their understanding of the message.
No comment indicated that the raters had misinterpreted the message. The raters confirmed what the test-takers had said by summarizing or explaining the content with their own words (Extracts 165-166). The raters also gave their interpretation of what the test-takers had said. For example, when Chloe (FCE) was not able to successfully convey her feeling about a particular job, Rater P interpreted what she intended to convey (Extract 167). Even though there were some linguistic forms and pronunciation features that made the message unclear, the raters attempted to understand the speakers’ intention from the context.

165 So he said “My house is really big. We’re not always together.” / And he went on to say that you know “we’re in all individual rooms and we’re not always together.” (T-Asad-SR-2688, 2689)

166 So this point Anita is literally just reviewing each every weather condition. (N-Anita-SR-1998)

167 [Chloe: yes. I: I agree with you. but (.2) I think it's a a good job but it's not really: (1.0) tch! I don't know its say (.)]

Yeah. He wanted … she wanted to say it’s not interesting.

(P-Chloe-SR-2249)

In addition to the confirmation of the ideas, the raters provided evaluative comments on the quality of ideas presented in the test-takers’ performances. The raters commented on the clarity of the content most frequently. For instance, they pointed out whether the content or point was clear (Extract 168). Furthermore, the raters often provided general judgments of the content of the speech, saying that the content was good or poor (Extract 169). Specificity was another aspect of content; specific and deep
content was more highly evaluated than shallow and superficial material (Extract 170). In particular, giving concrete examples or cases to support the main point was highly evaluated. For example, CET B⁺ discussed the sky in Tibet to argue air pollution in China, and CET A⁺ discussed an international conference on pollution recently held in Copenhagen. In contrast, Aida (FCE) was not able to provide concrete advantages of garbage collectors, which affected the raters’ judgments negatively. The raters also mentioned that arguments or thoughts given by some speakers were strong, mature, solid, reasonable, convincing, and strange (Extracts 171-172). Finally, some raters expressed their interest in the content (Extract 173).

168 Her points are very, very clear. (L-A⁺-SS-4481)

169 That was a good point. (G-C⁺-SR-856)

170 Her reason around here is relatively too simple. (E-Aida-SR-600)

171 She gives possible measures and the way she would tackle problems. / So it’s quite convincing. (G-A⁺-SR-762, 763)

172 She made a strong point about how to promote it, (N-C⁺-SS-4726)

173 and I found myself interested in what her thoughts were on the consequences of air pollution aside from just trying to give some feedback on her English language skills. (W-B⁻-SS-5590)

The raters most highly evaluated the content given by CET A⁺, CET A, and CET B⁺. These speakers gave clear and specific examples to support their arguments on the topic. In contrast, the ideas presented by low-proficiency test-takers (CET C and CET D) were rarely commented on. Among the Cambridge Exams test-takers, the content given by Asad, Chulsoo, and Anita was positively evaluated. They typically
made concrete description or explanation related to the topic. In contrast, Chloe and Aida failed to discuss the concrete advantages and disadvantages of the jobs they would like to engage with, and the raters negatively judged their content. Some of Asad’s content was criticized since his points were occasionally not clear.

Content elaboration was another major issue noted by the raters. They often made references to how much the speakers elaborated on their thoughts. Without exception, a high degree of elaboration was positively perceived (Extract 174). Speakers who elaborated on their message provided examples or reasons that supported their main point (Extract 175). Furthermore, detailed description was perceived positively in both monologic and dialogic tasks (Extract 176). In contrast, test-takers who were thought of as lacking in elaboration often failed to explain their message with examples, reasons, and detailed descriptions, and tended to proceed quickly to their next points. In the paired conversation, the raters negatively rated speakers who changed the subject without elaborating on or following the partner’s opinions (Extract 177).

174 So he’s actually trying elaborate from the point in his card, which is good.
   (M-A-SR-1810)

175 She gives a couple of reasons for why she likes that job. / And it’s good.
   (G-Aida-SR-959, 960)

176 In terms of contents, Asad conveys more clear, a clearer and more detail depiction of his hou ... his home and the things he talked about,
   (J-Asad-SS-4396)

177 And you can tell with Chloe specifically when she doesn’t know something and she doesn’t know how to explain, she kind just avoids it / and she just moves on / and she picks something else to elaborate on rather
than giving reasons why she doesn’t like a specific sort of job in the situation, whereas Aida is able to sort of give answers why she likes a job, why she like the same job. (L-Chloe-SR-1763, 1764, 1765)

The raters mentioned whether the content was relevant to the topic or coherent throughout the speech. The speakers who successfully answered the given question were positively evaluated rather than those failed to answer (Extract 178). Directness to the topic was also mentioned. With regard to this, several raters pointed out a contrast between Asad and Chulsoo (PET); the former elaborated on the content that was not necessarily relevant to the question, whereas the latter was more direct to the question though he lacked elaboration (Extract 179). Although staying with the subject was thought necessary, a contrasting opinion was also given (Extracts 180-181). Moreover, the raters mentioned the coherence of content including logicality and flow of arguments (Extract 182).

178 She’s talking a lot about the advantages, / but she hasn’t talked about the disadvantages, / so she’s missing half of the answer.
(Q-Aida-SR-2378,2379, 2380)

179 The differences were with Asad, it seemed that he was more rambling / and he wasn’t sort of getting straight to the point of the answer, whereas Chulsoo was able to clearly state “Okay, I have got two chairs, a television, this and that”. / So he was able to clearly answer the questions.
(Q-Asad/Chulsoo-SS-5069, 5070, 5071)

180 Okay he’s giving a measure. / Focus on causes not measures. (G-A-SR-783, 784)
And for factory, he not only mentioned why these factories cause air pollution but also provided some solutions for air pollution. He mentions government in this part and I think for the topic, even if he should talk about the causes but he also gave his own opinion about some effective solutions to this question. (H-A-SS-4106, 4107, 4108)

“Life of human beings because it produces gases and that is bad for environment.” He’s mixing things and there’s not a natural flow of the exposition of this idea. (G-B-SR-835, 836, 837)

The raters referred to the repetition and amount of content, although the frequency at which this was mentioned was smaller than the other aspects of ideas. The raters mentioned that some speakers repeated their opinions or kept stating similar thoughts during their performances (Extract 183). These behaviors were negatively evaluated since repetition distracted the raters’ attention and made them wonder about the speakers’ point. Instead, including many different ideas in a speech was rated positively (Extract 184).

I felt like the first two points are quite similar. Instead of a bit of repetition in saying about the harmful gases, which made me think that perhaps he couldn’t think of another point. (Q-B-SR-2324, 2325)

In the short time she has been able to communicate a lot of different ideas and a lot of points. And she’s done it pretty well. (D-A-SR-417, 418)

Three raters (Raters C, L, and M) respectively stated that an aspect of ideas—relevance, coherence, and repetition—affected their overall judgments the most.
In addition to these opinions, other raters provided their perspectives on the ideas shown by the speakers. Some raters felt that the content did not strongly affect their judgments because all the test-takers discussed similar subjects, and their content made few differences. For other raters, however, some specific aspects of ideas—elaboration of ideas and amount of content—seemed to strongly contribute to their judgments of communication ability (Extract 185).

185 And a lot of like, for instance, the air pollution talks only went to a certain point saying that, you know, “The carbon dioxide is bad for your health” and “This is not good” and “That’s not good,” but they rarely talked about why, and you know, without that content yeah, the impression I get is that their ability to communicate is limited. (V-Interview)

It was also commented that language proficiency determined how much the raters focused on the ideas presented by the speakers. The raters had difficulty in concentrating on the content given by speakers with poor fluency, pronunciation, and linguistic resources. In particular, Rater W argued that it was necessary to exceed a certain threshold level to make the listener attend to the content (Extract 186). Meanwhile, some raters claimed that good content could draw their attention to the performance in spite of low English proficiency. In this sense, as some raters commented on, the quality of an idea could play an important role in compensating for a speaker’s linguistic limitations (Extract 187)

186 If your pronunciation is terrible and your grammar is terrible, then you have already lost me before you get to content. So you need to have those
things as sort of baseline fundamentals before you can … I think start thinking about content. […] So content is, for me, it’s a number of things have to come before having good content. (W-Interview)

187 If it’s interesting it’s easier to listen to you know, some … if it’s interesting you can sort of forgive just about any you know, muck up of the English language or whatever. It’s interesting you sort of fight to understand it. (U-Interview)

Figure 4.14 summarizes the relationships between the speakers’ ideas presented and other features. The speakers’ ideas affected the raters’ judgments of communication ability and the degree to which they attend to the performance. The ideas presented compensated for limitations of English proficiency although a lack of English proficiency distracted the raters from the ideas.

**Figure 4.14.** Relationships between ideas and other features.

**4.9.2 Framing of Idea**

The raters mentioned the framing or schematic organization of ideas (3.4% and 0.4% of the comments given to the CET-SET and Cambridge test-takers, respectively).
Specific features sorted into this subcategory included schematic organization of speech and arrangement of ideas. Comments on this subcategory were mostly made when the raters viewed the individual presentations. This suggests that the framing of idea did not draw the raters’ attention in the paired interactions.

The raters attended to schematic organization of the whole speech. Including an introduction and conclusion was highly evaluated in the presentations. The kind of introduction often rated positively was a statement of the topic and an outline of the speech (Extracts 188-189). These types of introduction gave the raters the overall picture of a speech and made it easy to follow the subsequent content. The kind of conclusion evaluated positively was a summary of the main points with concluding remarks, which helped the raters remember the content and gave the impression that the speech was well organized (Extract 190). However, Rater G felt that a speaker’s summary of the argument was redundant, which suggests that simply restating the argument might not guarantee positive evaluation. The overall speech structure that contained an introduction, body, and conclusion was referred to as a sandwich effect by Rater W (Extract 191).

188 I think it was good that she introduced her topic so that you’ll be able to understand what she’s going to discuss. (K-A^[−]-SR-1508)

189 That’s a very good way to start. / “I have a couple of ideas. I will illustrate them in two dimensions.” / That’s a very very good structure. (G-A-SR-772, 773, 774)

190 And she also rounds off nicely and ends up her speech, which gives the whole thing some sort of structure, which is good. (C-B^[−]-SR-329)

191 and he has a nice sort of sandwich effect on his response where he has got
an introduction in the body and sort of the conclusion. (W-A-SS-5576)

Another feature related to the overall schematic structure of speech was the division of arguments. Most of the CET-SET test-takers divided their arguments into several parts and discussed each in turn using discourse markers and signposts. The raters positively commented on this, stating that the speech was well structured and laid out (Extract 192). Well-structured speech gave the raters the impression that the content was considered and preplanned well (Extract 193). In addition, well-structured speech was deemed to facilitate the understanding of the speech (Extract 194).

192 Her second … she laid her response out pretty well in that she had a first thought and then a second thought, (W-B⁺-SS-5593)
193 Another thing I forgot to talk about is that this girl, she is structuring the speech / and so it gives people the sense that her speech are all well thought off. (M-A⁺-SR-1794, 1795)
194 So he said “I’m gonna, I think, that’s four answers to this. This is the first one and this is the second one.” helps me to follow even though I might not quite understand all the words that he’s saying. (T-B-SR-2618)

How the test-takers arranged their ideas was another feature mentioned within this subcategory, though less frequently. The raters made positive comments on some test-takers who gave their main point first and supporting evidence or examples subsequently, as often observed in compositions (Extract 195). This feature was also mentioned when the raters rated the paired interactions, in particular the utterances of Anita (CAE) and Asad (PET). Some raters (Raters A, Q, and T) mentioned that these
test-takers said a broad idea first and gave specific examples later (Extract 196).

195 and with different … with every single part he helped the main sentence to
mention about it and other sentence to make the main point clearer.

(B-A-SS-3449)

196 It is interesting that she first of all gives a broad picture, big picture, and
then tries to add some examples and each part. / I mean, the whole general
ideas as presented first / and then it’s break down into parts.

(A-Anita-SR-110, 111, 112)

In the post-session interviews, the raters stated how strongly framing of idea
affected their overall judgments. Schematic organization of speech was one of the three
features affecting the judgment of communication most strongly for Raters A, C, G, and
J. When asked to what extent speech organization had affected their judgments, many
raters acknowledged the strength of its influence (Extract 197).

197 It [speech organization] actually influenced my understanding of this. I
think the way they organized their talk or not, strongly maybe 80% or 75%.

(A-Interview)

Two intriguing perspectives on the framing of ideas were explored in the
interview data. First, some raters (Raters K, O, and S) stated that schematic organization
was important in the individual presentation but not quite so in the pair conversations.
In particular, Rater S mentioned that people do not usually organize utterances when
they converse with others in real-world contexts (Extract 198). This may explain the
low frequency of comments on speech organization given to the Cambridge Exams test-takers. Second, the relationships with language proficiency were mentioned. Rater E commented that good organization would not be helpful if presented with low English proficiency (Extract 199). At the same time, Rater L claimed that good organization could compensate for the speakers’ low English proficiency (Extract 200).

198 Sometimes if it’s too organized, it seems a little bit false because when you do communicate and speak to people in general you don’t say, “Okay, first I am going to speak to you about this, and second it is…” (S-Interview)

199 But organization depended on English proficiency. Fluent speakers could speak without problem and it was not bad. But even though organization was good, I didn’t understand very well if the speaker can’t speak smoothly. (E-Interview)

200 Even if … because if you float into the next point and maybe your English or your pronunciation was not the best or your fluency, sometimes for me, due to the fact that you are on another point, so to some extent … and probably it also helps them to be like “Okay, I finished the first one, I need to you know, start on the second one, yeah.” So to some … it helped them and it helped … (L-Interview)

Figure 4.15 shows the relationships between the framing of idea and other features. The framing of ideas affected the raters’ judgments of communication ability and preparedness of the speech. It also to some extent compensated for a lack of English proficiency. Although well-structured speech facilitated comprehensibility, it was not helpful if the speakers’ proficiency was quite low.
4.9.3 Topical Knowledge

The test-takers’ background knowledge on the topic was noted by the raters (2.8% and 1.7% of the comments given to the CET-SET and Cambridge test-takers, respectively). Comments within this subcategory were related to knowledge and ideas held by the speakers: topical knowledge and ideas in mind.

The raters referred to the test-takers’ knowledge of or familiarity with the topic (Extract 201). The topical knowledge of CET A⁺, CET A, and Anita was commented on positively, but that of CET C and CET D was commented on otherwise. Although the raters did not explicitly state the reason underlying their judgments of topical knowledge, one of the factors was seemingly the content of the speech. Some raters praised CET A⁺’s deep knowledge of air pollution when she talked about an actual international conference on pollution held in Copenhagen. Additionally, some raters’ comments suggested that they inferred topical knowledge based on the passion and facial expression of the speakers. A specific language feature (the use of *et cetera*) also contributed to judgments of topical knowledge (Extract 202). Pauses and fillers also
seemed to make the raters assume that the speakers were not familiar with the topic (Extract 203).

201 and in addition, I think her knowledge regarding the topic was much better compared to the previous speakers. Yeah. (D-A^+SS-3636)

202 but he ended badly when he said and many other reasons so it’s like blah-blah-blah, and when I get lost, “yeah and etcetera”. / You don’t say “etcetera” in expositions / and it looks like you don’t know what you are talking about. (G-B-SS-3996, 3997, 3998)

203 [CET C: mm divided into: two parts. err firstly? is the:: human (.) body. err. (1.0) first? they: the: air pollution will err (1.0) harm our (.2) health? err (1.0)]

I think from this part, it shows some indication that she do not really understand about the topic. (I-C-SR-1262)

The speakers’ topical knowledge was considered to influence their demeanor. Some raters assumed that confidence or anxiety was closely related to the knowledge of the topic (Extract 204). Topical knowledge was also a decisive factor in the judgment of communication ability since the speakers could not convey a clear message without it (Extract 205).

204 But somehow I can sense that actually English is not really the main problem on here / but I just sense that probably the reason of her performance in here are just because lack of understanding about the topic that / and it lead her to become more confused and nervous / and it impact
to the whole performance of her speech. (I-C-SR-1274, 1275, 1276, 1277)

205 But she just doesn’t know … she’s coming across as poor communication
cuz she doesn’t know the topic or can’t tell us about her topic.

(U-C-SR-2805)

The raters also mentioned whether or not the speakers had ideas to discuss in
their mind. The raters considered that long pauses in the performances were because the
speakers did not know what to say or forgot the content of the speech (Extract 206). In
the paired interactions, the amount of contribution was thought as an indicator of ideas
that the speakers had in mind. For example, some raters thought that Chulsoo (PET) did
not speak much because he had little to say. Additionally, some raters mentioned that
Marie’s (CAE) small contribution to the conversation was related to her incapability of
finding things to say (Extract 207). These comments suggest that the raters attributed
long pauses or the small amount of contribution to a lack of ideas in mind not only a
lack of English proficiency. A lack of ideas in mind was considered related to
communication ability since it was perceived that the speakers were not able to come up
with the content on the spot. For example, CET C’s long pause led Rater R to perceive
that she could not find anything to discuss (Extract 208).

206 She has either forgotten the speech she’s prepared / or if she was, in my
opinion, if she was given a topic right there, then she’s run out of ideas. /
She doesn’t know what to say. (D-C-SR-448, 449, 450)

207 But maybe she was not that successful in trying to find new ideas and most
of the time just confirming what the other person was talking about.

(A-Marie-SS-3374)
In terms of communicating her ideas, she was obviously very, very nervous and couldn’t think of the right thing to say (R-C-SS-5120)

The raters also referred to whether the test-takers fully understood what they were asked to talk about (Extract 209). They seemed to judge the understanding of the question mostly through the content presented by the test-takers (Extract 210).

She’s able to firstly … first understand clearly what the instructor’s asking of her, (Q-Anita-SR-2356)

It’s obviously Chulsoo understands this question a lot better because he answered the question more directly than Asad did. (L-Chulsoo-SR-1788)

Figure 4.16 displays how topical knowledge was related to other features. On the one hand, the speakers’ topical knowledge affected the raters’ judgments of communication ability and was considered to influence demeanor and message conveyance. On the other hand, the content of the speech, fluency, and vocabulary in the performance affected the raters’ judgments of topical knowledge.
4.9.4 Differences among Proficiency Levels

The number of segments related to content—ideas, framing of idea, and topical knowledge—given to each CET-SET test-taker was examined (Table 4.26). Three high-proficiency test-takers (CET A⁺, CET A, and CET B⁺) received a larger number of positive comments than negative ones. In particular, the content presented by CET A⁺ was rarely criticized. In contrast, the rest of the test-takers (CET B, CET C⁺, CET C, and CET D) received a greater number of negative comments on the content of their speech. This result suggests that the raters’ judgments of content are related to English proficiency as gauged by the test. The raters made fewer comments on the content of CET D’s presentation.

Figure 4.16. Relationships between topical knowledge and other features.
The frequency of comments on content given to the Cambridge test-takers was also examined (Table 4.27). The content of the four speakers (Anita, Aida, Asad, and Chulsoo) appeared to be evaluated more positively, whereas the other two speakers’ (Marie, Chloe) content was evaluated more negatively. This suggests that the judgment of content was not necessarily related to the speakers’ English proficiency. A large number of neutral comments on Asad’s content were due to the raters’ frequent confirmation of his ideas.

<table>
<thead>
<tr>
<th></th>
<th>CAE</th>
<th>FCE</th>
<th>PET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anita</td>
<td>Marie</td>
<td>Aida</td>
</tr>
<tr>
<td>Positive</td>
<td>27</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Negative</td>
<td>6</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Neither</td>
<td>0</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>26</td>
<td>61</td>
</tr>
</tbody>
</table>

4.9.5 Differences among Raters

The number of comments on content mentioned by each rater was examined (Table 4.28). The frequency greatly varied; content accounted only for 1.8% of the comments made by Rater F but 31.6% of the comments made by Rater Q. Raters A, C, G, J, L, and M claimed that content was one of the features most strongly affecting their...
judgments and made a moderate number of references to speech content (from 10.6% to 27.8%). The finding shows that there was a large individual difference in the frequency and suggests that the frequency did not necessarily indicate the strength of attention to the feature.

Table 4.28
*Number of Segments on Content Provided by Each Rater*

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>43</td>
<td>12</td>
<td>26</td>
<td>16</td>
<td>42</td>
<td>4</td>
<td>108</td>
<td>77</td>
<td>45</td>
<td>39</td>
<td>26</td>
<td>53</td>
</tr>
<tr>
<td>%</td>
<td>10.6</td>
<td>4.6</td>
<td>12.6</td>
<td>7.8</td>
<td>13.6</td>
<td>1.8</td>
<td>27.8</td>
<td>22.1</td>
<td>12.6</td>
<td>18.8</td>
<td>16.1</td>
<td>16.4</td>
</tr>
<tr>
<td>Rater</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>33</td>
<td>81</td>
<td>25</td>
<td>67</td>
<td>62</td>
<td>12</td>
<td>12</td>
<td>29</td>
<td>28</td>
<td>31</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>14.3</td>
<td>29.1</td>
<td>14.6</td>
<td>23.7</td>
<td>31.6</td>
<td>5.2</td>
<td>8.2</td>
<td>12.1</td>
<td>9.0</td>
<td>11.6</td>
<td>7.1</td>
<td></td>
</tr>
</tbody>
</table>

4.10 Interaction

The raters mentioned the test-takers’ interactional behaviors or features, (a) interaction and engagement and (b) interactional patterns, mostly when they rated the Cambridge Exam test-takers (Table 4.29). The comments on this category accounted for 12.0% of all the comments given to the Cambridge Exam test-takers, making Interaction the fourth largest category. In contrast, the raters made only a few references to interaction in the presentations on the CET-SET. This finding is not surprising since the performances on the CET-SET were non-interactional individual presentations.

Table 4.29
*Frequency of Segments on Each Subcategory of Interaction*

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET #</th>
<th>CET-SET %</th>
<th>Cambridge #</th>
<th>Cambridge %</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Interaction and Engagement</td>
<td>28</td>
<td>0.7</td>
<td>175</td>
<td>7.1</td>
</tr>
<tr>
<td>7.2 Interactional pattern</td>
<td>0</td>
<td>0.0</td>
<td>128</td>
<td>5.2</td>
</tr>
<tr>
<td>Interaction Total</td>
<td>28</td>
<td>0.7</td>
<td>303</td>
<td>12.2</td>
</tr>
</tbody>
</table>
4.10.1 Interaction and Engagement

Comments on interaction and engagement accounted for 0.7% and 6.8% of all the comments given to the CET-SET and Cambridge Exams test-takers, respectively. Emerging themes or behaviors within this category were engaging with the partner or audience, asking questions, giving responses to the partner, supporting the partner, listening to and understanding the partner, and directness and humor.

The raters commented on the test-takers’ attempts to involve themselves in the interaction and engage with their partner, referring to such behaviors as interactive or communicative. In particular, they positively judged contingent interaction between pairs of test-takers but negatively judged individual message conveyance that did not accompany paying attention to the partner (Extracts 211-212). To a lesser degree, the raters commented on some CET-SET test-takers’ interactional behaviors. Addressing all the members of the audience (the examiner and fellow test-takers) made the raters perceive that the test-taker was fully engaged (Extract 213).

211 So like I said, the second speaker attempts to make some sort of conversation and to interact with the other speaker. (C-Chulsoo-SR-392)

212 Actually both sides I think they just want to say their own ideas and rarely pay attention to the other persons’ reaction to what they say. (A-Aida/Chloe-SR-149)

213 She’s engaging with the person all the time / Interaction with her fellow panels. (U-A⁺-SR-2723, 2724)

One of the specific behaviors related to interaction and engagement was asking questions of the partner, including general and follow-up questions. In most cases,
asking the partner questions was perceived positively (Extract 214). Asking for the partner’s opinions was often considered as a sign of engagement with the partner and considered to create interactions between test-takers (Extract 215). Additionally, the raters thought that those asking follow-up questions understood what the partner had previously said. For example, Raters H and Q positively commented on Chulsoo’s question, “How many people are in your family?” followed by Asad’s story about his large family in his hometown (Extract 216). In contrast, a lack of follow-up questions was considered unnatural in conversation (Extract 217).

214 She also gives the speaking to Marie, which demonstrates her communication ability. / She actually asks for Marie’s opinion, which is a good thing. (F-Anita-SR-720, 721)

215 And then she passes the conversation onto her with saying what do you think?, which is I mean the first part of conversation to include the other person in it. / So becomes a two-way channel. (N-Aida-SR-2009, 2010)

216 I think it was good for Chulsoo to give a question to Asad because according to what Chulsoo said, I think Chulsoo really understood and he could ask some questions according to Asad’s answer. (H-Chulsoo-SR-1150)

217 Now that I think about, it … it’s not very conversation because they don’t really ask each other follow-up questions. (K-Asad/Chulsoo-SR-1584)

The raters’ comments also included the test-takers’ responses to the partner, such as responses to questions asked by the partner or feedback on what the partner had mentioned. For example, Rater Q referred to Asad’s feedback on Chulsoo’s description
of simple furniture in his room (Extract 218). In addition, the raters made reference to
the test-takers’ agreement and disagreement with the partner (Extract 219). Finally,
responses particularly highly evaluated were backchannels or verbal signs showing
active listening. Using backchannels was seen as encouraging the partner to talk and
gave the raters the impression that the test-takers were listening to the partner (Extract
220).

218 Here Asad’s communicating more of an idea, which is good because it
shows that he’s understanding what Chulsoo said and then running with
that by saying “oh you know simple design you know might be more
relaxing or easier.” (Q-Asad-SR-2393)

219 and the only good point is at some point she challenged one of the
comments of Anita, / that was good. (G-Marie-SS-4052, 4053)

220 And Chulsoo is nodding and going mm mm. / And I think that encourages
Asad to speak. (T-Chulsoo-SR-2684, 2685)

Listening to and comprehending what the partner said was also considered as
part of communication ability (Extract 221). The test-takers’ behaviors such as nodding,
giving responses related to the partner’s utterances, responding with backchannels, and
asking follow-up questions gave the raters the impression that they were actively
listening to the partner. In particular, Chulsoo’s active listening was frequently
commented on. Among the CET-SET test-takers, there were those who received help
from the examiner. The raters attended to the listening ability through their response to
the examiner (Extract 222).
221 Aida does actually demonstrate good communication abilities right here because she’s listening pretty well to what Chloe is saying right now. (F-Aida-SR-738)

222 So she demonstrates some kind of communication ability because she’s able to listen, which is part of, as far as I’m concerned, of the communication ability. (F-C-SR-705)

The raters also positively considered the test-takers’ support when their partners could not maintain their speech. For example, Rater E positively considered Anita’s support when Marie seemed to have difficulty in continuing her utterance (Extract 223).

223 [Marie: *uhm*: *I think this one? it err must be very difficult for this old man or woman? he can’t barely walk? (.5) and err*  
Anita: *yes I think it’s really extreme situation. so you can’t]*  
Around here, she supported her partner and continued conversation successfully when her partner was at a loss for words. / Maybe for her partner, the support was very effective, (E-Anita-SR-588, 589)

Finally, to a lesser extent, some raters mentioned the test-takers’ politeness strategies and humor observed in the interactions. Some raters pointed out ways of showing disagreement with the partner. When Chloe (FCE) was asked for her opinion about a particular job, she agreed with her partner and attempted to express her disagreement. Two raters exhibited contrasting opinions about her way of showing disagreement. Rater U perceived that the utterance showed Chloe’s strategy not to be
direct and offensive, whereas Rater I felt that she should have been more direct
(Extracts 224-225). A few raters commented on humor observed in the PET speakers’
performance. They positively perceived humorous exchanges with laughter between the
two (Extract 226).

224 [Chloe: yes. I: I agree with you. but (.2) I think it’s a a good job but it’s not
really: (1.0) tch! I don’t know its say (. it’s:)]
She’s coming across as she doesn’t wanna offend somebody by saying that
she thinks it’s a dumb job but more on the lines of she doesn’t wanna do it.
(U-Chloe-SR-2863)

225 But in my opinion if she think that she disagree, she can just say she
disagree. / Doesn’t have to say she agree and then saying disagree.
(I-Chloe-SR-1358, 1359)

226 [Asad: but I think simple furniture is (.2) smart. (. because make you
maybe: relax. there is no lot of things you (.2) make you can feel you see
and you hear. %you know%]
Chulsoo: yes. I can make it clean easy. ((laughs))=
Asad: =yeah! that’s that’s important point.=
Chulsoo: =yes. hhh! ((laughs))]
And I like the bit of humor at the end. (T-Asad/Chulsoo-SS-5332)

Two raters (Raters H and N) referred to test-takers’ interactional behaviors
showing their engagement as one of the three features most strongly affecting their
judgments (Extract 227).
227 One was in the conversation pieces, how they reacted and responded and interacted with each other. If one person had more conversation connectors like “what do you think” versus the other person trying to just talk to themselves that affected my impression. (N-Interview)

4.10.2 Interactional Pattern

The raters mentioned interactional patterns exhibited by pairs of test-takers in the Cambridge Exams. Comments within this subcategory accounted for 5.2% of all the comments. Behaviors and features falling into this subcategory included initiating interaction and the balance of contribution.

The raters made a large number of comments on the test-takers’ initiating moves in the paired interactions. They often referred to and positively evaluated such behaviors as taking control, taking initiative, taking the lead, playing a leadership role, and so forth in most cases (Extract 228). Specific initiating behaviors pointed out by the raters were asking questions and facilitating conversation. For example, Rater O commented on Anita, who asked Marie for her opinions about the main themes of the conversation (Extract 229). Leading conversation by asking questions was considered to facilitate conversation and assist in making a better interaction (Extract 230). In each pair of test-takers, the one who was considered to initiate the interaction was Anita (CAE), Aida (FCE), and Asad (PET). The raters noticed that these test-takers often took the lead in the interactions (Extract 231).

228 And I notice that she has a more leading position between the both of them.

(O-Aida-SS-4850)

229 Here you see that Anita is playing the leadership role again by guiding the
conversation between the both of them. (O-Anita-SR-2124)

230 Throughout this conversation, she is the one who was facilitating the topic in a sense. (K-Anita-SR-1562)

231 It seems in this conversation structure is more of Asad talks first and Chulsoo talks after, (K-Asad/Chulsoo-SR-1582)

In contrast to those taking the initiative, their partners—Marie (CAE), Chloe (FCE), and Chulsoo (PET)—seemed to be perceived to be followers. The raters often negatively commented that these test-takers were passive, following the partner, waiting for the partner to talk, taking a back seat, and so forth (Extract 232). The most evident behavior giving the raters such an impression was agreeing with everything mentioned by the partner (Extract 233). In fact, the three passive test-takers seldom challenged their partner or elaborated on the partner’s utterances. These behaviors negatively influenced the raters’ judgments of the speakers’ ability to convey their thoughts (Extract 234). This might explain why Marie’s challenge to Anita was considered as a positive aspect of her performance (see Extract 219 above).

232 But maybe just during the conversation, he was the follower,

(A-Chulsoo-SS-3421)

233 She’s just repeating or saying whatever like just nodding and yes.

(P-Marie-SR-2240)

234 Marie seems to just be going along with whatever Anita suggests. / So that kind of makes it look as if she’s not as able to communicate what she is thinking. (Q-Marie-SR-2359, 5340)
The raters also commented on the relative size of contribution between pairs of test-takers in the interaction. They noticed how much each test-taker spoke in an interaction. First, the test-takers’ dominance in the interactions was mentioned. In particular, Anita (CAE) and Asad (PET) were considered to be dominating the conversation (Extracts 235-236). These speakers were perceived negatively since they did not necessarily give the partner a chance to talk. These test-takers were referred to as *outspoken, aggressive, and talkative*. At the same time, those who failed to contribute to the interaction were also perceived negatively as *less active, less forthcoming, shier, quieter, passive, and more reserved*. Marie (CAE) and Chulsoo’s (PET) contributions were particularly thought not to be enough (Extracts 237-238). The difference in the size of contribution between Aida and Chloe (FCE) was not so evident. Rater K commented positively on the similar amount of contribution between these speakers (Extract 239).

235 and she dominates the conversation talking potentially 75%, 80% of the 3 minutes and 20 seconds or more. (W-Anita-SS-5691)

236 But I feel like, I have to say though Asad talks about himself and goes on at length, not necessarily giving Chulsoo a chance, who has to kind of carve out his own speaking space. (N-Asad-SS-4714)

237 she doesn’t contribute as much as Anita does. (N-Marie-SS-4780)

238 and his response’s a lot shorter than Asad’s. (K-Chulsoo-SR-1583)

239 and I think they gave the equal amount of responses more or less, which I think is fine, because it’s … they were giving opinions. (K-Aida/Chloe-SS-4466)
Although no raters referred to interactional patterns in the interview as an influential feature, two raters (Raters V and W) mentioned that initiating behaviors affected their judgments of communication ability (Extract 240).

240 I thought Marie’s communication ability was little on the poor side, just below fair because she just … well, first of all she didn’t take as much initiative in answering the questions or keeping up with the conversation.

(V-Marie-VP-5515)

4.10.3 Differences among Proficiency Levels

Table 4.30 shows the frequency of comments on interaction given to each CET-SET test-taker. Only a few comments on interaction and engagements were given to the test-takers.

<table>
<thead>
<tr>
<th></th>
<th>A+</th>
<th>A</th>
<th>B+</th>
<th>B</th>
<th>C+</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Neither</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 4.31 presents the frequency of comments given to the Cambridge Exams test-takers. The evaluation of the comments suggests that the judgment of interactional features might not necessarily be linked to the level of English proficiency.

Low-proficiency test-takers (Asad and Chulsoo) received a larger number of positive comments than negative ones. Overall, the test-takers’ interactional features and
behaviors were evaluated positively, although the number of negative comments given to Chloe surpassed the positive comments given to her.

Table 4.31

Frequency of Segments on Interaction Given to the Cambridge Exams Test-takers

<table>
<thead>
<tr>
<th></th>
<th>CAE</th>
<th></th>
<th></th>
<th>FCE</th>
<th></th>
<th></th>
<th>PET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anita</td>
<td>Marie</td>
<td></td>
<td>Aida</td>
<td>Chloe</td>
<td></td>
<td>Asad</td>
</tr>
<tr>
<td>Positive</td>
<td>61</td>
<td>24</td>
<td></td>
<td>27</td>
<td>12</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>24</td>
<td></td>
<td>5</td>
<td>15</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Neither</td>
<td>19</td>
<td>16</td>
<td></td>
<td>9</td>
<td>7</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>64</td>
<td></td>
<td>41</td>
<td>34</td>
<td></td>
<td>52</td>
</tr>
</tbody>
</table>

4.10.4 Differences among Raters

Table 4.32 shows the number of comments on interaction mentioned by each rater. The number widely varied ranging from 0 (Raters B and R) to 47 (Rater H). The comments on interaction accounted for 14.6% of the comments made by Rater O.

Table 4.32

Number of Segments on Interaction Provided by Each Rater

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>23</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>9</td>
<td>23</td>
<td>21</td>
<td>47</td>
<td>22</td>
<td>6</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>%</td>
<td>5.7</td>
<td>0.0</td>
<td>3.9</td>
<td>1.5</td>
<td>2.9</td>
<td>10.2</td>
<td>5.4</td>
<td>13.5</td>
<td>6.2</td>
<td>2.9</td>
<td>6.8</td>
<td>5.0</td>
</tr>
<tr>
<td>Rater</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>13</td>
<td>20</td>
<td>25</td>
<td>10</td>
<td>15</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>25</td>
<td>3</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>5.7</td>
<td>7.2</td>
<td>14.6</td>
<td>3.5</td>
<td>7.7</td>
<td>0.0</td>
<td>3.4</td>
<td>4.6</td>
<td>8.1</td>
<td>1.1</td>
<td>3.2</td>
<td></td>
</tr>
</tbody>
</table>

4.11 Overall Impression

The raters often provided the overall evaluation of the speakers’ speech performance and ability. Comments within this category were related to the raters’ unspecified and holistic judgments of communication ability, which could not be coded
into any other analytic categories. The subcategories of *Overall Impression* were (a) overall performance and global ability, (b) overall message conveyance, (c) overall comprehensibility, and (d) overall English proficiency (Table 4.33). This was the largest category in both tests (accounting for 21.1% and 19.4% of all the comments, respectively). There was a statistically significant difference in the frequency of the segments between the two tests ($\chi^2(3) = 14.092, p < .01$).

Table 4.33  
*Frequency of Segments on Each Subcategory of Overall Impression*

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET</th>
<th>Cambridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>8.1 Overall performance &amp; Global ability</td>
<td>261</td>
<td>7.0</td>
</tr>
<tr>
<td>8.2 Overall message conveyance</td>
<td>173</td>
<td>4.6</td>
</tr>
<tr>
<td>8.3 Overall comprehensibility of message</td>
<td>312</td>
<td>8.3</td>
</tr>
<tr>
<td>8.4 Overall English proficiency</td>
<td>45</td>
<td>1.2</td>
</tr>
<tr>
<td>Overall Impression Total</td>
<td>791</td>
<td>21.1</td>
</tr>
</tbody>
</table>

**4.11.1 Overall Performance and Global Ability**

The test-takers’ overall performance and global ability were frequently mentioned (accounting for 7.0% and 7.4% of the comments). More specifically, the comments were regarding individual’s overall communication ability, overall impression of individual’s performances, impression of conversation as a whole, and unspecified features.

In their summary statements on the reasons for the ratings, the raters frequently mentioned their ratings first (Extract 241). Likewise, the raters also referred to their holistic judgments of individual test-takers’ communication ability (Extract 242).

241 I rated her communication ability average. (Q-C+SS-5027)
In providing their overall judgments of the speakers’ communication ability, the raters frequently referred to other test-takers and made comparisons. The CET-SET test-takers were compared with other CET-SET test-takers whom the raters had witnessed previously (Extract 243), although comparisons were not frequently observed. In contrast, many raters compared the communication abilities of two test-takers who interacted with each other in the Cambridge Exams. In most cases, Anita and Aida were judged more highly than their partners in CAE and FCE were, respectively (Extracts 244-245). With regard to Asad and Chulsoo (PET), a few raters made comparisons, but their judgments contradicted each other. The speakers’ communication ability was rarely compared with other speakers across different tests.

243 So he’s better than the other few who have done this presentation in this area because he’s got a bit grasp on some of his wording and pronunciation. (U-A-SR-2741)

244 Of the two, Anita and Marie, Anita has better communication abilities. (N-Anita-SS-4773)

245 but Aida has a slightly better edge than Chloe. (O-Aida-SS-4843)

The raters also referred to their overall impressions of the individual speakers’ performances. The raters frequently made reference to the quality of the performance itself rather than to the speakers’ communication ability (Extract 246). They also commented on whether they felt comfortable watching the performance or whether the performance was interesting (Extracts 247-248).
246 I think … I think her performance is very good. (J-A\textsuperscript{+}-SS-4332)

247 It was uncomfortable to watch. (W-D-SS-5674)

248 ((Laughs)) That is very interesting. / … / This is really funny. / This was really funny. / Interesting. (P-D-SR-2226, 2228, 2229, 2230)

The verbal reports also included comments on overall interactions between two speakers in the Cambridge Exams. These comments were not related to individual speakers’ ability but the raters’ overall impressions of the interactions or conversations themselves (Extract 249). Two raters (Raters I and V) also pointed out the overall flow of the conversation (Extract 250). In addition, several raters indicated that some interactions were contrived and did not look like real conversations. They claimed that a lack of topic development and eye contact made the interactions look less realistic (Extract 251). At the same time, Rater T referred to a part of interaction between Asad and Chulsoo as less scripted (Extract 252). In that part of interaction (see Extract 226 above), there were multiple short turns commenting on previous turns and several conversational features such as latched utterances and overlapping talk.

249 But I think this, it can be effective conversation because both of them can express their thoughts (H-Asad/Chulsoo-SS-4238)

250 and in general the conversation went very slowly, (V-Aida/Chloe-SS-5534)

251 One person is talking about something, / the other person interrupts that person and adds something or maybe give information about another topic. / It is less … than a successful two-side conversation. / These people maybe they lack interaction between each other as well. (A-Aida/Chloe-SR-152, 153, 154, 156)
It feels less like … a scripted less / and more like … I guess it like real conversation. (T-Asad/Chulsoo-SR-2699, 2670)

The comments within this subcategory also contained the raters’ impressions of unspecified features in a part of the performance. During the stimulated recall session, the raters stopped the video-recordings to make comments on the features that had affected their judgments. They addressed a particular part of the performance but made a holistic judgment without specifying the feature they attended to (Extracts 253-254).

I think she started out all right, (K-C+SR-1540)

So I think this is where she starts to deteriorate in her communication skills. (S-C+-SR-2537)

4.11.2 Overall Message Conveyance

The raters referred to the test-takers’ overall message conveyance in the CET-SET and Cambridge Exams (accounting for 4.6% and 4.4% of the comments, respectively). Comments within this subcategory were simply related to whether the speakers conveyed or failed to convey their thoughts to the listener or partner. Thus, the comments were not related to the content of the message or the judgment of the content (c.f., Section 4.9) but related to the speakers’ action of conveying a message in general. The raters often commented that the speakers were communicating or expressing their thoughts (Extract 255). The test-takers’ successful conveyance of content was positively perceived. The raters also judged highly those who were able to convey their feelings or emotions (Extract 256).
I think that he is pretty well able to communicate his ideas and to express exactly what his thoughts are. (F-Asad-SS-3934)

She shows the excitement she feels when she’s talking about champions and … yeah. (P-Aida-SR-2259)

Some raters mentioned that some speakers (in particular, intermediate English speakers such as CET C and Aida) conveyed the message successfully despite problems with fluency, linguistic resources, and pronunciation (Extract 257).

Even though his sentences like the grammar isn’t perfectly constructed or anything, he’s still conveying really clearly what he’s trying to say by just being confident, speaking loudly, engaging through eye contact and hand gestures. (Q-A-SR-2309)

In contrast, unsuccessful message conveyance was perceived negatively. The raters made reference to some speakers’ inability to convey their message (Extract 258). Furthermore, they commented on specific behaviors related to unsuccessful message conveyance, such as struggling with expressing thoughts, giving up expressing, and receiving help from others. These features were interrelated, since speakers who struggled with expressing their thoughts occasionally gave up talking and eventually received help from an examiner or a peer test-taker. In particular, many raters harshly judged Chloe (FCE) when she attempted to disagree with her partner but ended up agreeing because she could not express her thoughts (Extract 259). In addition, CET C and CET D gave up talking during their presentations and were interrupted by the third person, which was perceived quite negatively by the raters (Extract 260).
258 And he just is pretty much unable to say what he thinks. (F-D-SR-716)

259 [Chloe: I agree with you. but (.2) I think it’s a a good job but it’s not really: (1.0) tch! I don’t know its say (.) it’s: (1.0) I think it’s if people: do: (2.0) no that’s okay. yes I agree with you hhh! ((laughs)))]

See she tries for a while / and then I think she wasn’t able to express herself fully. (O-Chloe-SR-2132, 2133)

260 So this is the first person where we had another person at the table coming and interject because I think he might have been stuck here. / … / But when the other woman comes in and starts speaking and saving him. /

That’s when I marked him down even another notch two out of seven.

(W-D-SR-3150, 3152, 3153)

Some raters mentioned that the speakers’ lack of vocabulary knowledge was a reason for not being able to convey the intended message successfully. For example, Chloe’s failure to express her disagreement (see Extract 259 above) was considered as a consequence of her lack of vocabulary (Extract 261). At the same time, some raters showed difficulty in identifying the causes of unsuccessful message conveyance. They often wondered if the problem was a lack of topical knowledge or a lack of communication skills or English proficiency (Extract 262). Nervousness was considered another contributing factor to failure in conveying the message (Extract 263). This type of comment suggests that the raters considered these (topical knowledge and nervousness) as potential features related to message conveyance. When the raters encountered those who could not convey message, these features (vocabulary knowledge, topical knowledge, English proficiency, and nervousness) were considered as the causes.
I guess she can’t express it because she doesn’t have the vocabulary. (D-Chloe-SR-468)

So my first problem with this would be, it is very hard to distinguish between whether it’s a lack of communication skills or whether it’s the lack of knowledge or the lack of thoughts which she wanted to really put forward. (P-C-SS-4917)

So, I am not really sure maybe he just do not understand the topic or he just lost his focus because of something come up into his mind or his nervous or something. (I-D-SR-1293)

Two raters (Raters C and D) stated that message conveyance was one of the three features affecting their judgment most strongly. In addition, some raters’ comments showed that it was a direct reason for their overall judgments of communication ability (Extract 264).

But I think this, it can be effective conversation because both of them can express their thoughts (H-Asad/Chulsoo-SS-4238)

4.11.3 Overall Comprehensibility

The raters frequently referred to the overall comprehensibility of test-takers’ message or words (8.3% and 5.8% of the comments given to the CET-SET and Cambridge Exams test-takers, respectively). The raters commonly commented on whether they comprehended the message, overall picture, logic, key words, and sentences (Extracts 265-266). Even though some of the comments appeared to be related to other features discussed previously (e.g., pronunciation, content), comments
within this subcategory do not certainly show the specific part that the raters attended to. For example, Extract 265 does not clearly show whether Rater A was referring to CET D’s vocabulary, pronunciation, grammatical structure, or any other features. Therefore, comments including the phrase “I can(not) understand/get/follow …” were sorted into the Overall Comprehensibility subcategory. Furthermore, this subcategory explicitly refers to the raters’ understanding rather than speaker behaviors or speech features.

265 I cannot easily get what he is trying to say. (A-D-SS-3360)

266 But overall I could understand what they were trying to say.

(V-Anita/Marie-SR2989)

Many of the comments within this subcategory included features that had affected the comprehensibility (Extracts 267-275 below). Although these comments indicate specific aspects of performances attended to by the raters, only the main clause was interpreted when coding (see Guideline for Coding Segments in Table 3.18). However, these segments are important to interpret what specific features or behaviors strongly affected the comprehensibility of the message.

The raters most frequently referred to pronunciation, L1 accent, prosody, and paralanguage as features affecting whether the message and words were comprehensible. Overall, the raters attributed incomprehensibility of CET C+ and CET D to their pronunciation and strong L1 accent (Extract 267). In contrast, the pronunciation of high-proficiency test-takers (CET A+, CET A, and Anita) was highly evaluated since it made their message highly comprehensible (Extract 268). With regard to prosodic and paralinguistic features, the quite voices of some speakers (CET C, Marie, and Aida) made their utterances hard to understand (Extract 269). On the other hand, the raters
mentioned that the clear articulation of CET A and Chulsoo facilitated the understanding of their message.

267 and I found it quite difficult to understand what he was saying because things were rolling off his tongue a lot differently (L-D-SS-4544)
268 I thought it was easy to understand because he pronounced very clearly overall and the rhythm of his English was very consistent. (E-A-SS-3744)
269 and she also … she’s also sometimes hard to understand simply because she speaks sort of quite low voice and mumbles bit and probably also due to nerves. (C-C-SR-356)

The second most frequently mentioned feature influencing comprehensibility was fluency. A moderately slow pace of utterances facilitated understanding of message (Extract 270), but utterances with many pauses, fillers, and self-corrections were hard to comprehend (Extract 271). Some of the raters’ comments revealed that incomprehensibility was caused by both pronunciation and fluency (Extract 272).

270 Also maybe because the reason that she speak every single word really slowly, so it’s easy to follow what she want to say. (B-Aida-SR-290)
271 I think for the start of his speech, it takes a while to grasp what he’s going to talk about because he said uhm a lot and he kept correcting himself, which kind of actually make it harder to understand. (K-B-SR-1527)
272 So when they are speaking in English with their accent and talking faster, so it’s quite difficult to fully understand what she is trying to say. (I-Asad/Chulsoo-SS-4325)
Linguistic resources—vocabulary and grammar—were also mentioned as influential features but not as frequently as pronunciation and fluency. Some raters claimed that lexical choice and vocabulary knowledge affected comprehensibility (Extract 273). Only a few raters claimed that grammatical features affected their understanding (Extract 274).

273 As I said earlier, Asad is more descriptive / and he’s using adjectives, / so you can really understand what he’s describing like his room, (K-Asad-SR-1578, 1579, 1560)

274 [CET C+: mm err first of all I think er we should er do some promotion to let everyone to know how worse is er air pollution.] See this is where it’s quite difficult to understand because of grammatical things. (S-C−-SR-2533)

Speech content also contributed to comprehensibility. Elaborated and well-organized speech was easier to understand than that lacking elaboration on content and clear organization (Extract 275). Features influencing comprehensibility mentioned by only a few raters were confidence, gesture, and amount of speech. The raters’ language learning experience and familiarity with the speakers’ English were also considered to impact their understanding of messages.

275 Also, he suddenly finished his first point / and I didn’t get what he wanted to say. (E-D-SS-3789, 3790)

Some raters stated that the context helped them comprehend the test-takers’
intended messages (Extract 276). Even though pronunciation, fluency, and linguistic resources affected comprehensibility as discussed above, problems with these features did not necessarily hamper the raters’ comprehension (Extracts 277-278). Some of the raters’ comments also revealed that some positive linguistic features compensated for problems with other areas (e.g., slow and clear utterances compensated for strong L1 accents).

276 I mean I understood from the context. / But and it should be very easy for any person to understand what he’s trying to say from the context.

(L-A-SR-1602, 1603)

277 And it’s relatively easy to understand what he’s trying to say even though he doesn’t necessarily always pronounce the word the way it’s supposed to pronounced. (L-A-SR-1611)

278 It’s easy to understand what is trying to say even if there are mistakes.

(D-Asad-SR-476)

Figure 4.17 shows the features affecting comprehensibility of the test-takers’ message. The speakers’ pronunciation, fluency, linguistic resources, speech content, gestures, and confidence affected comprehensibility. At the same time, problems with pronunciation, fluency, and linguistic resources did not necessarily impede comprehensibility since context compensated for these limitations.
4.11.4 Overall English Proficiency

The raters made comments on the test-takers’ overall English proficiency although the frequency was relatively low (1.2% and 1.7% of the comments). Comments within this subcategory did not show any specific aspect of English proficiency but only included the raters’ holistic evaluations of proficiency (Extract 279). Unspecified English errors and struggles with English were also observed in this subcategory (Extract 280).

279 Okay, the speaker in the sixth experiment, obviously he is … I would say his ability to speak English is relatively good. (L-B-SS-4504)

280 You don’t talk … I don’t know … she’s first struggling with the language clearly. (G-Chloe-SR-953)
Some raters (Raters E, O, and L) indicated that English proficiency was a direct reason for their ratings (Extract 281), suggesting that communication ability of low-proficiency speakers would be judged as poor. However, some other raters did not necessarily regard communication ability as equivalent to the English proficiency of the speakers. For example, Rater I rated CET C’s communication ability as average (4 out of 7) although she judged the speaker’s English highly (Extract 282). Instead, the rater attributed her poor performance to other features such as a lack of topical knowledge or nervousness. These contrasting perspectives appear to suggest that a certain level of English proficiency is necessary, but that English proficiency does not guarantee linguistic laypersons’ positive judgments of communication ability.

281 so I felt her communication ability, maybe it’s because of English, is poor. (E-Marie-SS-3807)

282 I would give her a 1, 2, 3, 4 … 4 rate. / I think her problem is not English. / When she speak in … at the beginning, to me his English is quite good. (I-C-SR-4279, 4280, 4281)

Similar to the judgment of overall performance and global ability, the raters made comparisons of English proficiency between speakers. In most cases, the raters compared the speakers who had interacted with each other in the Cambridge Exams (Extract 283).

283 That was my first impression of Chulsoo that he might be a little bit less good English speaker than Asad. (W-Chulsoo-SR-3245)
4.11.5 Differences among Proficiency Levels

The number of comments on *Overall Impression* given to the CET-SET test-takers was examined (Table 4.34). CET D received by far the largest number of comments on this category (168 segments). The number of positive and negative comments suggested that the judgments of *Overall Impression* were closely related to English proficiency as measured by the CET-SET. The three high-proficiency speakers (CET A⁺, CET A, and CET B⁺) received a greater number of positive comments, whereas the low-proficiency speakers (CET B, CET C⁺, CETC, and CET D) received a greater number of negative comments.

Table 4.34

<table>
<thead>
<tr>
<th></th>
<th>A⁺</th>
<th>A</th>
<th>B⁺</th>
<th>B</th>
<th>C⁺</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>73</td>
<td>57</td>
<td>64</td>
<td>35</td>
<td>31</td>
<td>29</td>
<td>16</td>
</tr>
<tr>
<td>Negative</td>
<td>7</td>
<td>14</td>
<td>22</td>
<td>59</td>
<td>64</td>
<td>78</td>
<td>125</td>
</tr>
<tr>
<td>Neither</td>
<td>11</td>
<td>15</td>
<td>20</td>
<td>12</td>
<td>11</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>91</td>
<td>86</td>
<td>106</td>
<td>106</td>
<td>106</td>
<td>128</td>
<td>168</td>
</tr>
</tbody>
</table>

Table 4.35 shows the frequency of comments on *Overall Impression* given to the Cambridge Exams test-takers. All the test-takers except Chloe received more positive comments than negative ones. This finding seems to be aligned with the raters’ overall judgments of communication ability (see Table 4.2 and Figure 4.1 above).
Table 4.35
*Frequency of Segments on Overall Impression Given to the Cambridge Exams Test-takers*

<table>
<thead>
<tr>
<th></th>
<th>CAE Anita</th>
<th>CAE Marie</th>
<th>FCE Aida</th>
<th>FCE Chloe</th>
<th>PET Asad</th>
<th>PET Chulsoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>76</td>
<td>39</td>
<td>58</td>
<td>28</td>
<td>74</td>
<td>62</td>
</tr>
<tr>
<td>Negative</td>
<td>10</td>
<td>28</td>
<td>36</td>
<td>61</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Neither</td>
<td>14</td>
<td>17</td>
<td>17</td>
<td>18</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>84</td>
<td>117</td>
<td>107</td>
<td>101</td>
<td>94</td>
</tr>
</tbody>
</table>

### 4.11.6 Differences among Raters

The number of segments referring to *Overall Impression* mentioned by each rater was calculated (Table 4.36). Overall, comments on this category accounted for 11.3% to 34.2% of all the comments.

Table 4.36
*Number of Segments on Overall Impression Provided by Each Rater*

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>91</td>
<td>73</td>
<td>54</td>
<td>41</td>
<td>100</td>
<td>38</td>
<td>52</td>
<td>59</td>
<td>78</td>
<td>31</td>
<td>26</td>
<td>54</td>
</tr>
<tr>
<td>%</td>
<td>22.4</td>
<td>28.1</td>
<td>26.1</td>
<td>20.0</td>
<td>32.5</td>
<td>16.9</td>
<td>13.4</td>
<td>16.9</td>
<td>21.8</td>
<td>15.0</td>
<td>16.1</td>
<td>16.7</td>
</tr>
<tr>
<td>Rater</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>54</td>
<td>62</td>
<td>42</td>
<td>40</td>
<td>41</td>
<td>37</td>
<td>50</td>
<td>61</td>
<td>63</td>
<td>71</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>23.5</td>
<td>22.3</td>
<td>24.6</td>
<td>14.1</td>
<td>20.9</td>
<td>16.0</td>
<td>34.2</td>
<td>25.4</td>
<td>20.3</td>
<td>26.5</td>
<td>11.3</td>
<td></td>
</tr>
</tbody>
</table>

### 4.12 Other

The raters made a large number of comments on features that could not be coded into any other categories (Table 4.37). Such comments were coded into *Other*, which included (a) miscellaneous speech features and speaker behaviors, (b) comments not related to speaker behaviors, (c) raters’ general perceptions, and (d) fillers, procedures, and unclear comments. There was no statistically significant difference in
the frequency of comments between the two tests ($\chi^2(3) = 10.770, p>.01$).

Table 4.37

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>CET-SET</th>
<th>Cambridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>9.1 Miscellaneous speech features and speaker behaviors</td>
<td>187</td>
<td>5.0</td>
</tr>
<tr>
<td>9.2 Comments unrelated to speaker behaviors</td>
<td>118</td>
<td>3.2</td>
</tr>
<tr>
<td>9.3 Rater’s general perception</td>
<td>64</td>
<td>1.7</td>
</tr>
<tr>
<td>9.4 Fillers, Procedures, &amp; Unclear Comments</td>
<td>96</td>
<td>2.6</td>
</tr>
<tr>
<td>Other Total</td>
<td>465</td>
<td>12.4</td>
</tr>
</tbody>
</table>

4.12.1 Miscellaneous Speech Features and Speaker Behaviors

The raters made comments on speech features or speaker behaviors that did not fall into any categories (accounting for 5.0% and 2.5% of all the comments, respectively). Comments within this subcategory contained a wide variety of features and behaviors whose frequencies were quite low.

In the performances on the CET-SET, the raters commented on the test-takers’ apologies, preparedness, time management, memorization of scripts, recitation of notes, translation, management of nervousness, and familiarity with presentations (Extracts 284-290). The frequency of these kinds of comments was relatively low (10 to 15 segments).

284 She says *sorry* again. (G-C-SR-883)

285 This guy just seem that he is not prepared for anything at all.

(M-D-SR-1873)

286 When he said *in the life of human being*, I also somehow feel that he’s like memorizing the sentence. (I-B-SR-1224)
And he was so dependent on that piece of paper, (N-D-SS-4771)

I think here it’s more obvious that he’s trying to translate something.  
(B-O-SR-2091)

but he has managed to control his nervousness. (A-A-SS-3294)

and she is used to giving these presentations. (P-A\+SS-4868)

In the performance on both tests, the raters mentioned the speakers’ language background, miscellaneous speech features, and speakers’ behaviors. While watching the videos, the raters guessed where the speakers were from based on their performances and appearance (Extract 291). Miscellaneous features and behaviors included a wide variety of behaviors that were mentioned only a few times and did not fall under any category. Appendix 12 presents comments categorized into this subcategory.

I can clearly tell that Chulsoo is from Korea. (H-Chulsoo-SS-4223)

4.12.2 Comments Unrelated to Speaker Behaviors

This subcategory contained comments that were not about the test-takers’ behaviors or speech features (3.2% and 2.8% of all the comments). Nevertheless, some of them were deemed to affect the raters’ overall judgments. Such comments included the raters’ perception of topic, a peer or examiner’s behaviors, visual and sound quality of the videos, non-English placard found in the video recordings, description of the situation, and unrelated episodes (Extracts 292-297). These comments indicated that the raters considered the environment or circumstance in which the performance took place not only the test-takers’ performances per se.
292 Boring topic. (U-Aida/Chloe-SR-2869)

293 See the girls, his friends sitting behind him, are also smiling.
  (J-D-SR-1458)

294 The audio quality in this is a little bit poorer than the Cambridge one, which makes it … which probably makes me grade it a little bit lower.
  (W-A⁺-SS-5557)

295 I also noticed that there was a label for the University of Cambridge,
  (W-Asad/Chulsoo-SS-5735)

296 That’s … it’s a lot like an interview and examination like the Certificate of Advanced English something like that. (G-Aida/Chloe-SR-943)

297 It does sound a bit like he has maybe some marbles in his mouth when I watch it. / Anybody who has ever seen Gossip Girl, one of the girls who talks like she has marbles in her mouth. (W-D-SS-5665, 5666)

Other comments within this category were considered irrelevant to their judgments but indicated that the raters were addressing the researcher. The comments contained the raters’ account of their own reports and messages to the researcher (Extract 298).

298 Or I don’t know this ((laughs)) is relevant. (F-C-SR-700)

4.12.3 Rater’s General Perception

The raters occasionally stated their general perceptions of or beliefs about communication ability (accounting for 1.7% and 1.4% of all the comments). Comments within this subcategory were not directly related to the speakers’ performance or ability
but reflected the raters’ standards or internal criteria used to judge communication ability. The comments included (a) the raters’ beliefs about particular features, (b) knowledge of English words or phrases, and (c) knowledge of foreign accents.

The raters provided their general beliefs about particular features and the importance of the features (Extracts 299-300). These comments arose from the test-takers’ performances but not directly related to them. Such beliefs were considered to underlie their judgments of communication ability.

299 There should be communication during the conversation.

   (H-Anita/Marie-SR-1127)

300 Voice modulation. / Yes, very important. (P-B⁺-SR-2183, 2184)

Three raters (Raters G, V, and W) provided an explanation of how certain words or phrases were supposed to be used. They pointed out the test-takers’ particular lexical choices, and gave their thoughts about the correct usage of such choices (Extract 301). Some raters also mentioned their perceived difficulty of certain words used in the test-takers’ performances (Extract 302).

301 I mean, “of course” is said when something is evident. / You explain something to someone that does not know or even if they don’t know or I mean they are supposed not to know. / So you cannot have it as a given.

   (G-B⁺-SR-801, 802, 803)

302 the word “strenuous” is not … I don’t think really basic English word.

   (W-A-SR-3048)
Some raters also displayed their knowledge of the test-takers’ L1 accent or typical characteristics of NNESs with particular L1 background. For example, Rater D, who had learned Mandarin Chinese for 12 years, showed his knowledge of certain pronunciation features often observed in Chinese speakers’ speech (Extract 303). This was considered to facilitate his comprehension of the CET-SET test-takers’ presentations.

303 Yeah, what from my experience, when Chinese people speak English, they normally have a problem with some /z/ “measure.” (D-A+SR-402)

4.12.4 Fillers, Procedures, and Unclear Comments

Since this study employed the C-unit as the unit of analysis of the verbal reports, fillers produced by the raters were automatically counted as individual segments. Such fillers included “Okay” or “Yeah” with a falling intonation. This subcategory also included the statements on the procedures of the verbal report tasks, such as “I am going to go to the next one” or “We will take another view.” In addition, unclear and incomprehensible comments due to the inaudibility of segment were included in this subcategory. Comments within this subcategory were considered irrelevant to the raters’ judgments of communication ability.

4.13 Summary

This chapter has presented the results of (a) Rasch analysis of the ratings and (b) qualitative analysis of verbal protocol and post-session interview data. Whereas the raters’ impressions of the CET-SET test-takers were correlated with their test scores or English proficiency, high Cambridge Exams score achievers did not necessarily receive
high ratings from the raters. This means that high English language proficiency did not
directly lead to the raters’ high evaluations of communication ability. The verbal
protocol analysis explored nine main categories and 27 subcategories of factors that
appeared to influence the raters’ impressions. The factors were found to be intertwined
and to influence one another. The interview data also indicated the extent to which each
factor had affected the raters’ judgment of communication ability.
Chapter 5 Discussion

5.1 Introduction

This chapter provides a detailed discussion of key findings presented in Chapter 4 to answer the research question—“What features of oral performance or behaviors of speakers affect linguistic laypersons’ impressions of communication ability?”—and explain how the findings contribute to existing theories of L2 communication ability and language testing. The first section briefly summarizes the key findings of the study. The subsequent sections discuss speech features and speakers’ behaviors that affected the raters’ judgments of communication ability. The transferability of the findings is discussed thereafter.

5.2 Restatement of Key Results

Twenty-three linguistic laypersons (raters), proficient speakers of English with different L1 backgrounds, were asked to watch videos of oral performances in two tests, the CET-SET and the Cambridge Examinations, and provided verbal reports on the features that affected their judgments of communication ability. Their verbal reports contained a variety of features and behaviors, which were coded into nine main categories: Demeanor, Non-verbal Behavior, Pronunciation, Linguistic Resources, Fluency, Content, Interaction, Overall Impression, and Other. Table 5.1 shows the ranking of each category in terms of the frequency of segments. Semi-structured interviews were conducted to explore the raters’ perspectives on the importance of each feature. Table 5.2 shows the features and behaviors the raters provided when asked to nominate the three features that had most strongly affected their judgments.
Table 5.1

*Features Mentioned by the Participants and Their Rankings in Terms of Frequency*

<table>
<thead>
<tr>
<th>Ranking</th>
<th>CET-SET (monologue)</th>
<th>Cambridge (dialogue)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Overall impression (21.1%)</td>
<td>Overall impression (19.4%)</td>
</tr>
<tr>
<td>2nd</td>
<td>Content (15.1%)</td>
<td>Content (13.7%)</td>
</tr>
<tr>
<td>3rd</td>
<td>Fluency (13.5%)</td>
<td>Linguistic resources (12.7%)</td>
</tr>
<tr>
<td>4th</td>
<td>Other (12.4%)</td>
<td>Interaction (12.2%)</td>
</tr>
<tr>
<td>5th</td>
<td>Pronunciation (11.5%)</td>
<td>Pronunciation (10.2%)</td>
</tr>
<tr>
<td>6th</td>
<td>Linguistic resources (10.4%)</td>
<td>Non-verbal behavior (9.6%)</td>
</tr>
<tr>
<td>7th</td>
<td>Non-verbal behavior (9.4%)</td>
<td>Other (9.2%)</td>
</tr>
<tr>
<td>8th</td>
<td>Demeanor (5.7%)</td>
<td>Fluency (6.6%)</td>
</tr>
<tr>
<td>9th</td>
<td>Interaction (0.7%)</td>
<td>Demeanor (6.3%)</td>
</tr>
</tbody>
</table>

Table 5.2

*Feature that Raters Regarded as Most Influential*

<table>
<thead>
<tr>
<th>Feature/Behavior</th>
<th>Number of raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation (pronunciation, accent, prosody, paralanguage)</td>
<td>16</td>
</tr>
<tr>
<td>Fluency (flow, speed, pause phenomena)</td>
<td>13</td>
</tr>
<tr>
<td>Non-verbal behavior (gesture, eye contact, facial expression)</td>
<td>11</td>
</tr>
<tr>
<td>Content (organization, relevancy, coherence, repetition)</td>
<td>6</td>
</tr>
<tr>
<td>Grammar (accuracy)</td>
<td>5</td>
</tr>
<tr>
<td>Confidence</td>
<td>4</td>
</tr>
<tr>
<td>Ability to convey message</td>
<td>2</td>
</tr>
<tr>
<td>Interactional behaviors (engagement, response to partner)</td>
<td>2</td>
</tr>
<tr>
<td>Vocabulary (lexical choice)</td>
<td>2</td>
</tr>
</tbody>
</table>

5.3 Factors in L2 Communication Ability

This section provides an analysis of speech features and speakers’ behaviors that affected linguistic laypersons’ judgments with relevant literature.

5.3.1 Overall Impression and Speech Content

The raters made a large number of references to their overall impression of the performance and speech content. *Overall Impression* included overall performance and
global ability, overall comprehensibility, overall message conveyance, and overall English proficiency. **Content** included ideas, framing of ideas, and topical knowledge. These were the most frequently mentioned categories in their verbal reports, accounting for 36.2% and 33.1% of all the comments in the two tests.

Main components of these two largest categories—in particular, message conveyance, comprehensibility, and ideas—were considered closely related to the outcome of communicative performance. The outcome of communication is generally dependent on the goal or purpose of communication. The type of individual presentation in the CET-SET (see Section 3.4.2.5) was informative speech. The purpose of this speech was to inform listeners about issues of air pollution. Beebe and Beebe (2009) state that the goals of an informative speech are to (a) enhance understanding, (b) maintain the interest of the audience, and (c) make the audience remember what was said. Therefore, one of the outcomes of the presentations was successful or unsuccessful conveyance of the information intended by the speakers. Furthermore, the content of the speech is also closely related to the outcome of a presentation since what the audience takes away from a presentation has to do with interest and memorability.

This is also the case for the paired interactions in the Cambridge Exams, in which an elicitation task required the speakers to exchange information and thoughts. The type of the interaction was different from interactional interaction where speakers communicate to create and maintain personal relationship (Goh & Burns, 2012). In this sense, the goal of the paired interactions was conveying intentions; as Rickheit et al. (2010) put it, in terms of purpose of message production: “Speakers (as well as writers and signers) produce language in order to convey certain ideas to their interlocutors” (p. 27). Accordingly, the speakers’ meaning conveyance and mutual understanding were the main outcomes of the paired interactions. At the same time, task achievement was a
crucial part of raters’ judgments. In fact, the speakers were given subjects to discuss (e.g., “Discuss how each weather condition in the picture affects people’s lives and choose two conditions that could have the most harmful effect.”). Thus, whether the speakers successfully discussed the given subject and reached an acceptable conclusion was another outcome of the interactions.

Linguistic laypersons’ strong attention to the outcome of communication and content aligns with previous studies of indigenous assessment criteria derived from domain experts who are also linguistic laypersons (see Section 2.3.3). For example, Elder (1993) demonstrated that subject teachers were concerned most strongly with teacher trainees’ behaviors that would make students’ learning successful rather than with grammatical accuracy. Furthermore, Jacoby (1998) found that many of the criteria used by physicists for assessing peers’ conference presentations were related to the content of informative presentation. Likewise, Galloway (1980) found that linguistic laypersons attended mostly to the content of oral presentations in their judgments of it, and expressed boredom at the content. In this sense, these linguistic laypersons applied the strong approach to performance assessment (McNamara, 1996), focusing on the real-life criteria or successful task achievement. This is clearly distinct from the weak approach or language specialists’ focus on the quality of language in the performance (Brown, 2007; Brown et al., 2005; McNamara, 1990). It was found that the linguistic laypersons in the present study, as domain experts in various studies, evaluated the success of the achievement largely by taking into account the goal of communication.

The raters’ reflections on their judgments suggest that they would indeed be concerned primarily with the outcome of the performance and with its content in the real-life situation. For example, it was shown in Rater S’s account, “when you speak to someone in daily life your goal generally isn’t to understand the level of their
communication ability. Your goal is to understand what they are saying.” Furthermore, it was indicated that linguistic laypersons would not pay strong attention to molecular features in real-life settings. These comments support the claim of Hughes (2011) that “In naturally occurring spontaneous speech interlocutors do not focus on the mechanics of their interaction but on the ideas/emotions/information being conveyed” (p. 85). It is therefore not surprising that the data showed a high frequency of comments on Overall Impression and Content.

It was evident that English proficiency played an important role in the raters’ judgments of the communicative outcome. The relationship between the test scores and the raters’ judgments of Overall Impression and Content showed that high-scoring achievers (high-proficiency speakers) received more positive judgments of these categories, most specifically in the oral presentations (see Tables 4.26 and 4.34). This finding is not surprising because language ability generally enables speakers to create discourse (Bachman & Palmer, 2010) and convey detailed and relevant information (De Jong et al., 2012). In this sense, low-proficiency speakers are less likely to convey intentions successfully and provide a high-quality message. In other words, the outcome of communication does depend on the level of English proficiency to some degree. Furthermore, a certain level of language proficiency was necessary to draw attention to the content of the speech. This is clearly reflected in Rater W’s comment, “If your pronunciation is terrible and your grammar is terrible, then you have already lost me before you get to content” (Extract 186). This supports the finding of Pollitt and Murray (1996) in that raters paid attention to the speech content of higher proficiency speakers and to the problematic linguistic features of lower proficiency speakers. This might be the reason that the raters made considerably fewer comments on the content of CET D’s presentation than other speakers.
At the same time, however, English proficiency is not the sole factor in the raters’ judgments of communicative outcome. This was particularly observed in the raters’ judgments of the paired interactions (see Tables 4.27 and 4.34). For example, Rater H regarded the overall interaction between Asad and Chulsoo (PET) as an “effective conversation” (Extract 249) even though their English proficiency was not necessarily high. Message conveyance and comprehensibility were claimed to be successfully achieved despite the speakers’ language problems. This supports Rickheit et al. (2010), who state that the linguistic representation of utterances does not have to be complete, detailed, and accurate for the interlocutor to understand the speaker’s intended message. Similarly, even though language proficiency is necessary to produce a message, it did not guarantee a high-quality message since the quality of content also depends on topical knowledge or metacognitive strategies (Bachman & Palmer, 2010). It is also related to part of Hymes’s (1972) ability for use such as cognitive ability or general reasoning powers. This explains why the raters’ judgments of content were not aligned with the level of English proficiency.

The raters’ overall impressions of communication ability were correlated with English proficiency as gauged by the CET-SET but not necessarily with proficiency levels measured by the Cambridge Exams. Intermediate to high proficiency speakers—Aida (CAE), Marie (FCE), and Chloe (FCE)—received lower ratings than the test-takers whose English proficiency was lower—Asad and Chulsoo (PET). There are two possible explanations for this finding. First, the raters might have evaluated the co-constructed interaction per se (Jacoby & Ochs, 1995) not only each individual’s communication ability. For example, some of the raters’ comments referred to the flow of conversation per se or confluence (McCarthy, 2010). The test scores might not have captured this aspect. Another possible explanation might be that the raters made
comparisons of two speakers within a pair rather than speakers across different tests. The data show that the raters frequently compared speakers who interacted with each other but rarely compared those who did not (e.g., Anita and Asad). In contrast, in the case of the CET-SET, the raters judged a speaker’s ability compared with other CET-SET test-takers whom they had witnessed previously. This divergence suggests that the raters used different frameworks of reference when judging the monologic and dialogic performances.

5.3.2 Language Proficiency

As discussed previously, one of the criteria used by linguistic laypersons in judging communication ability was related to the outcome of communicative performance. At the same time, the components of language proficiency—fluency, pronunciation, and linguistic resources (vocabulary and grammar)—were mentioned by the raters. These language-related categories accounted for 35.4% and 29.5% of the comments given to the CET-SET and Cambridge Exams test-takers, respectively.

5.3.2.1 Fluency

Fluency was found to affect the raters’ judgments of communication ability considerably. Fluency discussed here is related to fluency in the narrow sense in Lennon’s (2000) term, which is “the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing” (p. 26). Some language proficiency models (e.g., De Jong et al., 2012; Hulstijn, 2011) have included fluency (processing skills) as a component of core L2 speaking proficiency. However, what observable features of fluency or utterance fluency (Segalowitz, 2010) would affect linguistic laypersons’
perception of communication ability and how important fluency is in relation to other linguistic and non-linguistic features were unknown. The verbal protocol data showed that fluency was the third and eighth largest category in the monologic and dialogic tasks, respectively (this large difference is discussed below), and 13 raters selected utterance fluency as one of the most salient features.

The raters frequently claimed that dysfluency was a serious cause of lack of comprehensibility of message among language-related categories. Such comments are consistent with previous research findings regarding the role of fluency in people’s perception. Researchers claimed that perceived fluency considerably influences not only comprehensibility of message (Derwing, Rossiter, Munro, & Thomson, 2004) but also listeners’ concentration on or attention to the content (Lennon, 1990, 2000). It was frequently mentioned that the raters were particularly distracted from the message when they encountered excessively slow speech, pause phenomena, and repair phenomena. Among the subcategories of fluency (speech rate, pause phenomena, and repair phenomena), pause phenomena were mentioned by far the most frequently, and eight raters recognized them as one of the most influential features in their judgments. This indicates that pause phenomena—unfilled pauses, filled pauses, and hesitations, which are regarded as breakdown (dys)fluency (Skehan, 2009)—considerably influenced the linguistic laypersons’ perceptions of communication ability.

In contrast, fluent utterances gave the raters positive impressions and made them focus on the content. Moreover, it was shown that grammatical mistakes were overlooked if the speakers could maintain fluency. These findings support Lennon (1990), who argues that “fluency reflects the speaker’s ability to focus the listener’s attention on his or her message by presenting a finished product, rather than inviting the listener to focus on the working of the production mechanisms” (pp. 391-392). It was
also suggested that fluency or flow of speech was perceived to be more important than grammatical accuracy. This was particularly reflected in Rater R’s comment: “In terms of communication I feel the fast one and the ungrammatically correct one is better because it means your ideas are getting across faster” (Extract 145). This might be because fluency is more closely related to comprehensibility and message conveyance. This strong attention to fluency rather than linguistic forms is a unique characteristic of linguistic laypersons and contrasted previous studies investigating language specialists’ judgments of speaking proficiency (e.g., Brown, 2007; Brown et al., 2005; McNamara, 1990; Zhang & Elder, 2011).

As stated above, fluency was the third largest category in the CET-SET but the second smallest category in the Cambridge Exams. This might be due to the raters’ different expectations of preplanned speech and impromptu conversation. The CET-SET test-takers were given one minute of preparation time prior to their speech. Even though they needed to construct most of the discourse on the spot, the raters assumed that their presentations were planned speech. Some raters considered a prompt card held by the test-takers as notes they had prepared beforehand. In general, planned speech contains fewer filled pauses and hesitations than spontaneous talks since speakers are supposed to know what to say (Guillot, 1999). Therefore, raters can be considered to expect individual presentations to flow without undue pauses and hesitations. In contrast, the raters might have regarded the performances of the Cambridge Exams test-takers as impromptu conversations. As Hughes (2011) states, it is normal for speakers to be extremely hesitant when giving their opinions spontaneously, and listeners in this situation do not pay much attention to pauses or filled pauses to any great extent. Accordingly, the raters might not have penalized dysfluency in dialogues because of their expectations about fluency (or the lack thereof) in spontaneous conversation.
Another possible reason that the raters did not frequently comment on fluency in the paired interactions is that the speakers displayed a number of interactional behaviors that masked a lack of fluency. The raters often commented on the Cambridge Exams test-takers’ backchannels and supports for their partner. This sort of interlocutor’s involvement in conversation might make an individual speaker’s utterances sound more fluent since it filled silences. In addition to this, as McCarthy (2010) suggests, each participant in conversation contributes to conversational flow or *confluence* by employing various turn-taking strategies (including turn-opening and turn-closing). As some of the rater comments showed, the raters attended to confluence, not only individual speakers’ temporal fluency.

### 5.3.2.2 Pronunciation

Pronunciation was found to be another influential language-related category affecting the raters’ judgments of communication ability. Phonological knowledge is included in most theoretical models of communicative competence and speaking language proficiency as an essential component (e.g., Bachman & Palmer, 2010; Canale & Swain, 1980; De Jong et al., 2012; Hulstijn, 2011). In this study, pronunciation features were frequently recognized as one of the most influential features. Nevertheless, it was the fifth largest category in the performances on both tests, and the frequency of comments on pronunciation was not particularly large. A possible explanation for this gap is that the raters’ comments on incomprehensibility due to pronunciation were categorized into *Comprehensibility* rather than *Pronunciation*. Since pronunciation was the most significant reason for incomprehensibility, this categorization affected the frequency of segments on pronunciation more than other categories. Another reason might be that the raters let pass some problematic pronunciation features, an allowance
often observed in ELF communication (Firth, 1996), and did not stop the videos to comment on them. Thus, although the frequency of mentions was moderate, the influence of pronunciation on overall judgment was considered strong.

The raters considered that the test-takers’ pronunciation (including L1 accent) strongly affected their judgments of communication ability because it largely affected the comprehensibility of messages. In this regard, pronunciation was one of the major causes of incomprehensibility of messages. Thus, unintelligible pronunciation permeating a performance was a fundamental problem and automatically became the sole reason for severe judgment of communication ability. In fact, CET D’s poor pronunciation and strong L1 accent were often perceived to be unintelligible and directly led to a low overall rating. Furthermore, his content and linguistic resources were not strongly attended to. This role of pronunciation supports a claim made by Iwashita et al. (2008): “If a listener cannot make out the words, then they are not in a position to judge the ideas, syntax, and so on; pronunciation may therefore be acting as a sort of first level hurdle” (p. 44). Therefore, for low-proficiency speakers, achieving intelligible pronunciation should be prioritized. The crucial role of pronunciation for low-proficiency speakers is also corroborated by some past research findings (De Jong & Van Ginkel, 1992; Higgs & Clifford, 1982; Pollitt & Murray, 1996).

Although unintelligible pronunciation severely lowered the raters’ judgments, pronunciation deviant from standard or native-like pronunciation did not necessarily lead to negative judgment of communication ability. L1 accent was perceived in the same way. Accent per se did not lower overall judgment as long as it was intelligible. Furthermore, having an L1 accent was considered a normal trait, and did not automatically suggest unintelligibility. This supports Munro (2008), who states that “listeners have an underlying awareness of the distinction between speech that is merely
‘different’ from native speaker output and speech that is difficult to understand” (p. 203). It is thus suggested that once intelligibility is ensured, accented pronunciation does not affect people’s judgment and distract people from the content. This perception seemed to be common among raters with different L1 backgrounds, although ENL raters were more sensitive to and affected by the accuracy of segmental features than ESL and EFL raters, as has also found by some researchers in the past (Kim, 2009b; Riney et al., 2005). This finding overall suggests that the raters did not necessarily idealize NES pronunciation and espouse the standard language ideology, which is claimed to be popular among the general public (Jenkins, 2007; Seidlhofer, 2011; Widdowson, 2003). This could be due to their familiarity with NNES pronunciation and accent as well as their extensive experience in engaging with ELF communication in daily life.

There are two points to be discussed regarding pronunciation. First, the raters’ L1, English learning experience, and familiarity with the speakers’ pronunciation potentially influenced how much they could comprehend the message. This study confirmed some previous studies (e.g., Bradlow & Bent, 2008; Gass & Varonis, 1984) in that listeners’ familiarity with speakers’ accent is likely to facilitate comprehension. The result also suggests that interlocutors who share the same L1 as speakers may find it easier to understand them than those who do not, a notion which contrasts with the findings of some studies (Harding, 2011; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Munro, Derwing, & Morton, 2006). In fact, since these speaker-independent factors were based on the raters’ self-reports, it is still unknown about how strongly the factors affected the raters’ comprehension and the judgments of communication ability. At least, this study indicated that (a) intelligibility is not the sole responsibility of speakers, (b) the degree of intelligibility differs depending on interlocutors, and (c) the first impression of a speaker may change as interlocutors
become familiar with him or her.

Second, this study explored a speech feature that has not been emphasized in communicative competence and L2 language proficiency models so far: vocal paralanguage. This feature includes characteristics of voice dynamics (e.g., loudness, voice quality) and articulation (e.g., clear enunciation). These affected the raters’ impressions of comfort, nervousness, and confidence as well as intelligibility. In particular, moderately loud voice and clear articulation were perceived positively. This aligns with Spitzberg and Hurt (1987), who included vocal tension in their model of interpersonal competence as a skill related to composure (relaxation and confidence). The results also revealed that voice dynamics is a molecular feature related to people’s overall impression of interpersonal skills. In fact, vocal paralanguage has likely not been regarded as a component of L2 communicative competence because it is not a skill exclusive to L2. However, given the raters’ perception, it should be regarded as a crucial skill of L2 oral communication that has an impact on interlocutors’ impressions of demeanor.

5.3.2.3 Linguistic resources

Comments on linguistic resources mainly included vocabulary and grammar, which have previously been elucidated by most theoretical models of L2 communicative ability and L2 speaking proficiency (e.g., Bachman & Palmer, 2010; Canale & Swain, 1980; De Jong et al., 2012; Hulstijn, 2011). Linguistic resources are also one of the main features in an oral performance that language teachers strongly attend to (e.g., Brown, 2000, 2007; Brown et al., 2005; Galloway, 1980; Kim, 2009b; McNamara, 1990; Zhang & Elder, 2011). However, this study found that vocabulary and grammar did not necessarily affect the overall judgments of communication ability.
strongly, as compared with other linguistic and non-linguistic features. Linguistic resources were the sixth largest category in the CET-SET, and seven raters mentioned that they were one of the most influential features. Nevertheless, the raters frequently commented on these aspects of the paired interactions. Linguistic resources were the third largest category, which was larger than pronunciation and fluency.

Accuracy of linguistic forms was found in the comments on linguistic resources. However, the speakers’ inaccurate vocabulary use and grammatical structure in a performance rarely hampered the comprehensibility of the message. This might be because the test-takers were intermediate English speakers who could produce meaningful utterances with their linguistic resources although the forms were not necessarily accurate. Therefore, it might be that they had minimum linguistic resources to convey message, and the raters attended to other aspects of performances (Pollitt & Murray, 1996).

Most of the raters acknowledged that inaccuracy was not crucial as long as the content was decipherable. This suggests that comprehensibility was the primary criterion to determine error gravity or the seriousness of linguistic errors (James, 1998). For example, idiosyncratic but comprehensible grammatical forms were perceived as a “minor thing” (Extract 124). Similar to the test-takers’ pronunciation, the raters did not necessarily idealize the NES standard forms and demand conformity. Since the raters placed a strong emphasis on the outcome of communicative performances and the content as discussed above, it is not surprising that linguistic forms did not strongly affect judgments of communication ability if they understood the message. Furthermore, some raters tolerated grammatical forms deviant from NES standards because of their general expectations about English spoken by NNESs.

However, many comments on linguistic resources were related to
comprehensible but idiosyncratic lexical choices and grammatical forms. Raters who pointed out specific lexical choices and grammatical forms were mostly ESL and ENL raters. This might be due to their high sensitivity to linguistic resources, as in the case of pronunciation. In fact, several lexico-grammatical features that ELF researchers consider as varietal differences or innovations (Cogo & Dewey, 2012; Lowenberg, 1993, 2002; Seidlhofer, 2011) were perceived as errors and distracted the ESL and ENL raters’ attention. For example, Rater W referred to pluralization of an uncountable noun (hair) as “a big issue that tends to be noticed.” In fact, pluralization is often observed in English spoken by EFL and ESL speakers (Lowenberg, 1993, 2002) and can even enhance functional effectiveness (Seidlhofer, 2011). However, the use of English deviations from the norm of ENL speakers was not overlooked, even though they did not seriously impede comprehensibility. This means that comprehensibility was not the sole criterion for perceived error gravity, and thus that the raters (subconsciously) used another criterion: conformity (James, 1998).

The raters made a larger number of references to linguistic resources in the paired interactions than the individual presentations. The frequency of comments on Grammar was the cause of the gap; the raters made a significantly larger number of references to grammar of the Cambridge Exams test-takers than the CET-SET test-takers (p < .05). One of the possible reasons is that the raters could not judge the grammatical aspect of CET D’s performance due to his unintelligible pronunciation. In this sense, he did not pass the “first level hurdle” (Iwashita et al., 2008, p. 44) and failed to draw the raters’ attention to his grammar. Another possible explanation is that the performances of the Cambridge Exams test-takers might contain more grammatical errors than those of the CET-SET test-takers.

The raters’ verbal protocols showed two important relationships among the
language-related categories. First, fluency was emphasized more strongly than linguistic accuracy. A large number of raters mentioned that they were more distracted by features of dysfluency than by linguistically inaccurate forms. Moreover, as stated above, some raters explicitly expressed their preference for fluent speech production even at the cost of ungrammaticality. This may be because dysfluency results in loss of interlocutors’ attention to the content (Lennon, 1990, 2000), but native-like accuracy is not required for comprehensibility of messages (Widdowson, 2003). Furthermore, fluent utterances might distract interlocutors from linguistic forms (Lennon, 1990). This could explain why linguistic self-correction was not perceived positively even though it led to more-accurate linguistic forms. Second, the raters had greater difficulties in comprehending the content when the speakers had problems with more than one linguistic feature. This suggests that even minor problems in grammar, vocabulary, pronunciation, and fluency may be crucial if they occur simultaneously.

5.3.3 Interactional Competence

Rather unsurprisingly, interactional features and behaviors were noted mostly in the dialogic performances between two test-takers. This category is the fourth largest one in the paired dialogue, and two raters acknowledged that the test-takers’ interactional behaviors (responses to and engagement with the partner) affected their overall judgments most strongly. The findings suggest that successful interactional behaviors are one of the key components of communication ability as defined by linguistic laypersons.

Many of the behaviors within this category were those that had been explored by empirical studies and recognized as aspects of interactional competence (Kramsch, 1986). The raters mentioned how interactive or engaging the test-takers were. Directly
addressing the partner was generally perceived to be highly interactive and engaging. More specifically, the raters positively perceived interactive behaviors such as asking questions, responding to the partner, listening attentively, using backchannels, and supporting the partner. These interactional behaviors were considered as part of interactional competence contributing to co-construction of interaction (e.g., Brown, 2005; Ducasse, 2010; Galaczi, 2013; Lazaraton, 2002; May, 2011a). In addition, some of them (asking questions or listening attentively) are recognized as interaction management skills in Spitzberg and Hurt (1987) and Wiemann’s (1977) theoretical models of interpersonal competence. In contrast, interactions that lacked these behaviors were considered unnatural and unsuccessful. For example, the first part of the interaction between Asad and Chulsoo (PET) did not look like an authentic conversation because they lacked follow-up questions. This shows the importance of other-orientation or altercentrism, which requires a communicator to be open to the message given by the partner (Spitzberg & Cupach, 2002; Wiemann, 1977).

The raters also noted test-takers’ interactional patterns, pointing out the speakers’ contribution to interactions, including initiation and passiveness. The raters positively perceived a collaborative interactional pattern characterized by high mutuality and high equality (Galaczi, 2008). In all pairs of interaction, however, an asymmetric interactional pattern was observed to some degree. That is, one speaker was dominant and the other speaker was passive. When the raters noticed this interactional pattern, they tended to judge the dominant speaker positively and the passive speaker negatively (relatively speaking). The interactional managing skills (Spitzberg & Hurt, 1987; Wiemann, 1977) demonstrated by the former speaker were positively perceived as actions that facilitated conversation. In contrast, passive speakers typically simply expressed their minimal acknowledgement of the partner’s utterances, which was
interpreted as meaning that they could not express their opinions and contribute to the interaction. Nevertheless, the dominant speakers were occasionally criticized by the raters since they did not provide their partners chances to state their opinions, which indicated a lack of interpersonal competence (Spitzberg & Cupach, 2002). As Galaczi (2008) argues, dominance could signify inappropriate interactional behavior on the part of the speaker or might be a consequence of partners’ lack of contribution. The raters seemed to perceive dominance in both ways blaming the speaker in some cases and to the interlocutor in other. At least, this finding suggests that linguistic laypersons would make positive judgments of speakers who are capable of leading a conversation without depriving their interlocutor of the chance to talk.

There were two types of speakers’ behaviors that were not categorized into Interaction but were nevertheless considered highly relevant to engagement and interaction: (a) non-verbal behavior and (b) topic development. Although these were assigned to different categories (Non-verbal Behavior and Content, respectively), some of the features gave the raters the impression that the speakers were engaging with their partner and contributed to co-construction of interaction. Non-verbal behavior relevant to interaction and engagement included gestures, head nods, eye contact, and facial expressions. These behaviors are noted by Spitzberg and Hurt (1987) and Wiemann (1977), who claim that they contribute to showing affiliation with the partner and altercentrism. This finding is also consistent with the results of some previous studies (Ducasse, 2010; May, 2011a; Orr, 2002) in that gestures and eye contact were acknowledged as behaviors contributing to interactional effectiveness. Young (2008) also regards gestures and facial expressions as components of interactional resources.

Another interactional behavior not categorized into Interaction was topic development. This is related to the degree of development of the topic initiated by the
partner, which is also referred to as *contingency* with each other’s ideas (Gan, 2010) or *horizontal management* (Ducasse, 2010). This feature is related to interactional patterns since topic expansion is one of the parameters of collaborative, parallel, and asymmetric interactional patterns (Galaczi, 2008). Some studies have demonstrated that high-proficiency L2 speakers can follow the partner’s previous utterance by using various language functions such as agreeing, suggesting, disagreeing, explaining, and challenging, whereas low-proficiency speakers tend to exhibit a parallel interactional pattern by initiating a new topic (Galaczi, 2013; Gan, 2010). In this study, FCE test-takers often failed to elaborate on the utterance given by the partner and exhibited a parallel interactional pattern. Rater evaluations suggested that speakers need not only to acknowledge their partner’s utterances but also to contribute to the interaction by elaborating on their thoughts as expressed in previous turns.

The literature has suggested that interactional competence is highly relevant to language proficiency. Researchers (Galaczi, 2013; Gan, 2010; Lazaraton, 2002) found that high-proficiency English speakers can develop other-initiated topics, take initiatives appropriately, and display natural turn allotment. However, in this study, the interaction and engagement of some of the low-proficiency test-takers were perceived more positively than those of higher-proficiency test-takers. In particular, the FCE test-takers were criticized for their parallel interactional pattern, whereas the PET test-takers were perceived to more successfully display their interactional competence. One of the reasons might be because of the different task requirements of each test. The PET test-takers were asked to discuss their living places freely whereas the FCE test-takers were required to discuss a particular subject by going through given pictures to make a decision. Thus, the latter might have felt pressure to discuss all the pictures and could not elaborate on topics given by the partner. Another possible reason might be that some
interactional behaviors (e.g., attentive listening, using backchannels) are less relevant to language proficiency. Some raters praised those who asked questions even with a simple grammatical structure (e.g., *What do you think?*) or gave responses with a single word. This shows that even low-proficiency speakers may be able to successfully demonstrate interactional competence.

As stated above, the communication abilities of some lower-proficiency test-takers were judged more highly than that of those with higher proficiency. The abilities of the PET test-takers were judged as being higher than those of a CAE test-taker and of the two FCE test-takers. A plausible reason is that the speakers’ interaction and engagement played a more decisive role in the judgments of communication ability than linguistic resources including vocabulary and grammar did. In dialogic interactions, simply displaying high language proficiency and achieving one’s own communicative goal does not guarantee an observer’s positive evaluation. Rather, co-construction of interaction and achievement of mutual satisfaction (i.e., altercentrism) appears crucial (Jacoby & Ochs, 1995; Wiemann, 1977).

5.3.4 Non-linguistic Factors

The raters’ verbal protocol and interview data indicated that the speakers’ demeanor and non-verbal behavior (NVB) affected their overall judgments of communication ability. Although the frequency of segments was not particularly large, the importance of these features in oral performances was mentioned (see Table 5.2). This suggests that these factors were perceived as important components of communication ability.

Although the frequency of verbal comments on NVB—body movement, eye contact, and others—was not large, the important role of NVB in the judgment of
communication ability was confirmed by the interview data and by its strong impacts on other factors. The raters often acknowledged that NVB facilitated message conveyance and drew attention to the content. This view is consistent with the role and importance of NVB stated in many theoretical models of communication. For example, Canale (1983a) regarded NVB as a communication strategy that (a) compensates for communication breakdown and (b) enhances communicative effectiveness. In addition, Celce-Murcia et al. (1995) provided a list of non-verbal communicative features in their pedagogically oriented model of communicative competence, arguing that these features carry a large part of social meaning and inappropriate use of NVB may result in miscommunication. NVB is also regarded as part of interactional competence, which contributes to co-construction of interaction (Young, 2008). Accordingly, it is not surprising that the linguistic lay raters considered NVB as a component of communication ability.

One of the NVBs positively perceived by the raters was gesture, which refers to movements of the arms and hands that can carry meanings (McNeill, 1992, 2005). They are different from body movements that speakers unconsciously exhibit when they are nervous, such as touching their hair or playing with rings (Kendon, 2004). Gestures that the test-takers demonstrated in their performances were often positively perceived because they reinforced verbal messages and facilitated message conveyance. This finding indirectly supports Gullberg (1998) and Nambiar and Goon (1993), who discovered that observed gestures played an important role in raters’ judgments of communicative performances. Although the present study did not compare the raters’ judgments of the same performance with and without gestures, the data suggested that using gestures contributed to positive evaluation of communication ability. The type of gesture perceived helpful in the performance was iconic gestures, which “present
images of concrete entities and/or actions” (McNeill, 2005, p. 39). The raters often referred to Asad’s iconic gestures that facilitated communication. This finding appears to support Gullberg’s (1998) finding that the use of iconic gestures was positively correlated with rater evaluations of proficiency. It is thus suggested that the ability to effectively and strategically use iconic gestures should be considered a part of communication ability.

The speakers’ eye contact was another aspect of NVB noted by the raters. It expressed the speakers’ engagement or rapport with the partner and considerably affected the raters’ judgments of anxiety and confidence. Unfortunately, L2 communicative and interactive competence theories have neither specifically explicated eye contact nor stressed its importance in communication (e.g., Bachman & Palmer, 1996, 2010; Canale & Swain, 1980; Young, 2008). However, some empirical studies have demonstrated that eye contact affects raters’ judgments of interactional success (Ducasse, 2010; May, 2011a; Orr, 2002). Furthermore, communicative and interpersonal competence models (Morreale et al., 2013; Spitzberg & Hurt, 1987; Wiemann, 1977) include eye contact as a skill related to affiliation, interactional management, and composure (i.e., confidence and relaxation). These models align with the raters’ comments on eye contact in this study. Absence and intentional avoidance of eye contact—gaze omission or gaze avoidance (Hargie, 2011)—were negatively perceived without exception. This finding was intriguing, since appropriate eye contact differs depending on cultures and social norms (Harris, Moran, & Moran, 2004). However, it may be safe to conclude that gaze omission in oral presentations (or public speaking) is likely to induce negative impressions in the audience. Regarding eye contact in dialogues, the results may not always be applicable to natural conversation because in the tests the test-takers were asked to discuss pictures placed in front of them; this
interactional setting differs from authentic conversation.

The speakers’ demeanor (mainly related to confidence and anxiety) was considered as a feature influencing people’s judgments of communication ability. In particular, confidence seemed to compensate for linguistic limitations and enhance comprehensibility of the message. The role of demeanor has rarely been discussed in communicative competence and interactional competence models. An exception is Canale (1983a), who states essentially that confidence and willingness to communicate can compensate for L2 speakers’ difficulties in grammatical accuracy. The results of the present study support his claim. In fact, social relaxation and composure are acknowledged as components of communicative and interpersonal competences (Spitzberg & Hurt, 1987; Wiemann, 1977), suggesting that being relaxed and confident is part of skills of effective communication. Similarly, in an LSP context, Douglas and Myers (2000) found that professional demeanor was considered as a criterion by veterinary professors in assessing prospective vets’ performance. Given the role of demeanor mentioned in the literature and empirical studies, it can be considered that speakers’ confidence and anxiety are components of communication ability.

Demeanor was judged mostly through the speakers’ NVB; these categories appeared to be highly intertwined. This observation is largely supported by the interpersonal competence models (Morreale et al., 2013; Spitzberg & Hurt, 1987; Wiemann, 1977), which regard NVB (e.g., body movement, posture, gaze) as skills that reflect composure including relaxation and confidence. May (2011a) found a similar phenomenon, namely that test-takers’ body language was linked with rater impressions of confidence. This study also found that some language-related factors such as pauses and repair phenomena (hesitation, pauses, and stuttering) and articulation affected judgments of demeanor. This is in keeping with notions of interpersonal competence,
which view speech rate and vocal tension as skills contributing to composure (Spitzberg & Hurt, 1987). Although a speaker’s demeanor is not necessarily an L2-exclusive factor, it should be acknowledged as an important factor considerably influencing people’s judgments of L2 communication ability.

**5.3.5 Other Factors**

The raters mentioned a wide variety of miscellaneous behaviors or features observed in the test-takers’ performances. Some of them were found in previous studies as indigenous criteria used to assess communicative performances in specific domains (e.g., time management as found by Jacoby (1998)). In addition, many features were found to be task-specific (e.g., memorization of scripts, recitation of notes, and management of nervousness in oral presentations). This study revealed that the frequency of comments on each feature was quite low and that no rater mentioned any of these features in the post-session interviews. However, it would be premature to conclude that the miscellaneous features played only a minor role in the raters’ judgment of communication ability, given the nature of verbal protocol and self-report data (Barkaoui, 2011; Cohen, 1994; Lumley, 2005).

It was also found that some factors not relevant to the speakers’ behaviors nonetheless affected the raters’ judgments. They included factors that might be applicable to real-life settings (e.g., interestingness of the subject matter and audience reactions) and those unlikely to be applicable (e.g., non-English placards in the video recordings). This finding shows that the speakers’ behavior and speech features are not the only factor affecting laypersons’ judgments. However, such a speaker-independent factor cannot be regarded as indicative of a speaker’s communication ability.

Comments on these factors (miscellaneous features and those not relevant to
the speakers) indicate the complexity of linguistic laypersons’ judgments of L2 oral performances. They appeared to pay attention to a wide range of speakers’ behavior and speech features as well as external factors, and there were large individual differences. This complexity is not surprising given the fact that even trained and accredited raters have been shown to attend to various features not stated in assessment criteria even though they assess performances using fixed assessment criteria (Brown et al., 2005; Orr, 2002; Zhang & Elder, 2013). In particular, Orr (2002) found that trained raters of FCE seemed to heed 12 non-criterion features although the FCE criteria were composed of only four criteria. Furthermore, some of these were not directly related to the test-takers’ ability to speak the L2, such as age and gender of candidates. Similar to Orr’s (2002) study, this study suggests that raters’ judgment is determined by a large number of factors, including those that cannot be regarded as relevant to a speaker’s communication ability.

5.4 Transferability of the Findings

After providing verbal protocols on all the test-takers, the raters were asked a question as to whether the features they mentioned in this study would affect their impressions of L2 speakers in daily life. The question was “You talked about the things that affected your impression. Do you think the same things affect your impression of people in daily life too?” Most of the raters (20 out of 23 raters) reported that the features they mentioned in the protocol sessions would also affect their impressions of L2 speakers in everyday life. For example, Rater N answered: “Yeah, absolutely I do. … So these [features that she mentioned in her verbal reports] are things that I look at. So I used that here.” This is a quite positive result from this study since it suggests high transferability of its findings to similar communicative settings in real-world contexts.
It must be acknowledged however that there were three factors likely to undermine this transferability. First, it should be noted that the experimental setting in the study was different from that of everyday interactions. In real-world settings, linguistic laypersons are not likely to consciously judge an interlocutor’s ability to communicate. Instead, they focus primarily on understanding the message the interlocutor is trying to convey. For example, Rater S noted: “when you speak to someone in daily life your goal generally isn’t to understand the level of their communication ability. Your goal is to understand what they are saying. So you don’t so much focus on how they are saying things but on the content of what they are talking about.” Accordingly, linguistic laypersons might not attend to the manner of communication as strongly as the raters did in the study. It might be that for some raters (in particular ENL raters) it was easy to comment on obviously problematic linguistic features in the stimulated recall session in which they were required to stop the videos and comment on the performances as frequently as possible (see Brown, 2007).

Second, being asked to judge oral performances in experimental settings might have affected the raters’ judgments. This is particularly applicable to raters’ judgments of the paired interactions. In the case of the paired interactions of the Cambridge Exams test-takers, the raters were asked to watch the interactions without actually participating in the conversation. This point was made by Rater S, who stated that “I think it’s very different in daily life because when you are speaking to someone directly, it’s a completely different experience from viewing them talking to somebody else over a computer screen.” Figure 5.1 illustrates a dyadic interaction with an observer, showing that the observer judges the actor, his or her interlocutor (the co-actor), and the interaction itself (Spitzberg, 2000). As the figure suggests, the observer’s judgments are not necessarily identical with the actor and co-actor’s judgments of each other and their
own interaction. This study only explored linguistic laypersons’ judgments of communication ability as an observer.

Figure 5.1. Target and attributor loci of competency judgments (Spitzberg, 2000, p. 113).

Finally, the difference between communicative tasks used in the study and everyday communication should be noted. Since the communication setting of the study was a high-stakes oral examination, most of the test-takers were under stress and appeared extremely nervous. Rater Q mentioned this point: “It’s a bit of a different setting because it’s not just an everyday conversation, they are kind of being tested. So I think they probably have a bit more pressure on them than in everyday conversation.” Thus, the findings of the study are most likely to be transferable to various public speaking where presenters are under pressure. In contrast, the transferability of the findings to a variety of casual conversation might be weaker than individual presentations because participants in conversation are in general not likely to feel extremely nervous. In this study, therefore, the raters might have paid more attention to
confidence or anxiety than they would normally do in casual communication in the daily life. Moreover, as discussed above, the task-oriented nature of the paired interactions was considered to influence the raters’ judgments of communication ability. The results of the study might not be readily transferred to casual conversation in which participants are not motivated by a concrete pragmatic purpose (Egging & Slade, 1997) or communication to maintain social relationships with others (Goh & Burns, 2012; Morreale et al., 2013).

5.5 Summary

This chapter has discussed the findings of the present study in terms of the research question, “What features of oral performance or behaviors of speakers affect linguistic laypersons’ impressions of communication ability?” The study revealed nine categories of features that had affected the raters’ judgments of communication ability and the degree of their influence through a verbal protocol and interview data.

The results showed that the outcome of communicative performances, including successful message conveyance and the quality of content, was one of the main criteria used to judge communication ability. Although linguistic features such as fluency, pronunciation, and linguistic resources were recognized as factors affecting the outcome of communication, raters neither considered them crucial nor penalized errors harshly unless comprehensibility was seriously impeded. Interactional features such as engagement and the size of contribution affected raters’ judgments in the paired interactions. Their impressions were also influenced by the test-taker’s non-verbal behavior and perceived level of confidence, which were found to be closely related.

In conclusion, the findings of this study have suggested that linguistic laypersons consider some factors that existing theories of L2 communication ability
have overlooked. Their perspectives on language proficiency also seem to differ from language specialists’ perspectives. Finally, this study has found a wide variety of factors in linguistic laypersons’ judgments and the complexity of subjective judgments, supporting McNamara (1996), who states that “Judgements that are worthwhile will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement” (p. 117).
Chapter 6 Implications

6.1 Introduction

This chapter provides implications of this study and discusses how this study contributes to the field of applied linguistics. The first section addresses the contribution of the current study to the theory of L2 communication ability by discussing how it differs from existing theoretical models of communicative competence, interactional competence, and L2 speaking proficiency. The second section deliberates implications for language testing, in particular for the construct definition of general-purpose oral proficiency tests.

6.2 Theoretical Implications

This study used empirical data to define L2 communication ability. This ability was defined by identifying its components (knowledge, skills, and abilities), as has been attempted by a range of applied linguists (Purpura, 2008). The components explored in the study offer important insights for theories of L2 communication ability and for deeper understanding of communication ability.

In this study, the empirical data on which the definition of L2 communication ability was based were the linguistic laypersons’ subjective judgment of L2 speakers’ oral performance. This study has explored speech features and speakers’ behaviors that positively and negatively affect the linguistic laypersons’ judgment of communication ability. Importantly, the explored components of communication ability should be regarded as those affecting the degree to which an L2 speaker is viewed as a competent communicator rather than as those precisely determining L2 communication ability, as illustrated in Figure 6.1. In other words, the identified features and behaviors do not
guarantee positive or negative evaluation of the speaker; however, they are likely to 
impact judgments of the speaker. The features and behaviors are most likely to gain the 
linguistic laypersons’ attention in L2 speakers’ oral performances. The appropriate 
demonstration of the important features and behaviors increases the probability that the 
audience, or interlocutor, will judge the speaker as a competent communicator.

Figure 6.1. Theoretical implication of the present study.

As Figure 6.1 shows, the speakers’ speech features and behaviors are one of the 
components that affect the linguistic laypersons’ judgment of communication ability. It 
is important to note that other factors would also affect their judgment. This study has 
revealed that the perspectives of the linguistic layperson were closely intertwined with 
communicative contexts, tasks, and objectives. In addition, the laypersons’ familiarity 
with NNESs’ English and personal beliefs of the laypersons were found to be factors 
affecting their judgment. This suggests that communication ability should not be 
regarded as a static ability that is generalizable to different contexts; this study thus 
supports the interactional view of L2 communication ability because the context of 
language use is inseparable from the ability to communicate (Chalhoub-Deville, 2003;
This study’s approach to L2 communication ability has an obvious advantage because it offers an explanation of how people actually perceive L2 communication in real-world contexts. The approach to theorizing L2 communication ability based on the linguistic laypersons’ views is beneficial for applied linguists, language teachers, and language testers since the theory can inform them of the knowledge, skills, and abilities that L2 speakers should possess to be a competent communicator. In particular, this theoretical description of communication ability can largely contribute to L2 education that aims to teach and assess a set of abilities considered important by real interlocutors or audiences outside the teaching and testing milieu. Clearly, this approach sets it apart from existing theories of L2 communication ability. As reviewed previously, theories of L2 communicative competence, interactional competence, and speaking proficiency have been based on either reaction to, or elaboration of, previous theories (e.g., Bachman & Palmer, 1996, 2010; Canale & Swain, 1980; Hymes, 1972; Kramsch, 1986) or empirical analysis of L2 speakers’ performance on language tests (e.g., Bachman & Palmer, 1982; De Jong et al., 2012; Hulstijn, 2011). In other words, they have not incorporated the perspectives of linguistic layperson who is most likely to be the ultimate arbiter of communication ability in real-life contexts. This is arguably a limitation because theories, no matter how sophisticated, might not be able to explain and predict actual judgment.

This study has explored the role of ability for use in the laypersons’ judgment of communication ability. Ability for use includes non-language-exclusive cognitive and affective factors. As Hymes (1972) states, “The specification of ability for use as part of competence allows for the role of noncognitive factors, such as motivation, as partly determining competence” (p. 283). McNamara (1996) argues that theories of L2
communication ability have not satisfactorily explained *ability for use*—although it considerably affects performance of communicative tasks—since the factors underlying the ability are not exclusively related to the domain of language proficiency. Instead, existing theories typically emphasize language-related knowledge, skills, and abilities (Luoma, 2004). Nevertheless, the layperson raters in this study were clearly oriented to the speakers’ *ability for use*, and expressed its role in their judgment of communication ability. More specifically, they mentioned the importance of confidence, anxiety, and willingness to communicate. Furthermore, non-language-exclusive cognitive factors, such as general reasoning powers were found to potentially affect the impression of an L2 speaker, since the quality of content (e.g., elaboration and coherence) is a crucial criterion of competent communication as defined by the linguistic layperson. Although *ability for use* is complex and further research is needed, this study has explored some potential features of the ability that are likely to affect the linguistic laypersons’ judgment of communication ability. The findings of this study suggest that excluding *ability for use* from L2 communication ability undermines the validity of theories explaining evaluative judgment of communication ability.

In a similar vein, this study has explored speech features and behaviors that are not exclusively related to L2 knowledge, skills, and abilities, and which applied linguists have rarely paid attention to. For example, it was found that the speakers’ non-verbal behavior and vocal paralanguage facilitated or hampered comprehensibility of the message and thus affected judgment of communication ability. These are explained in models of interpersonal competence and communication competence (e.g., Spitzberg & Cupach, 2002), but are not necessarily emphasized in existing theories of L2 communication ability. As non-linguistic features that affect communicative performance and other factors, such as confidence, they can be considered another
aspect of *ability for use*. Inclusion of these features and behaviors broadens the notion of conventional L2 communication ability, which has been restricted mostly to L2 language-related elements.

Since this study incorporated the perspectives of English speakers with a variety of L1 backgrounds (i.e., EFL, ESL, and ENL speakers), the features explored reflect L2 communication ability valued by a wide range of English speakers in ELF settings (see Section 2.3.4). This characteristic of communication ability different from the existing theories of L2 communication that usually reflect idealized NES competence and narrow NES communities (Leung, 2005; Seidlhofer, 2011). For example, Leung (2005) claims that the influential communicative competence model of Canale and Swain (1980) considers NESs a pedagogic reference point. Furthermore, McNamara (2011) points out that the CEFR (Council of Europe, 2001), a globally used framework for L2 teaching and assessment, presumes that L2 speakers only communicate with NESs and emphasizes the responsibility of L2 speakers for communicative success and breakdown. It does not take into account the ability to comprehend NNES interlocutors and to convey an intelligible message to NNESSs. This study has shown that conformity to the NES norm is not necessarily an integral part of L2 communication ability in ELF settings, and deviance from the NES norm is not an error or deficit, as Jenkins (2006b, 2009b) and Seidlhofer (2011) argue. At the same time, this study suggests that the existing theories do not satisfactorily explicate the knowledge, skills, and abilities relevant to ELF communication (Canagarajah, 2006; Jenkins, 2000, 2006b, 2009b). This study's conceptualization of the ability is more suitable for ELF communication than the existing theories that presume exclusive communication with NESs.

Finally, this study has not only explored the components of communication
ability, but also revealed the relationships among the components, described by figures in Chapter 4. More specifically, this study found (a) how an individual component of communication ability can affect judgment of other components (i.e., causal relationships); (b) which components can compensate for a lack of other components; and (c) which components are perceived to be more or less important than other components. Widdowson (2003) has pointed out that the interdependency of components of communication ability has not always been explicitly described in theoretical models, although it is essential in a model of communication. In fact, as reviewed in Section 2.2.3, some researchers have investigated relationships among components (Bachman & Palmer, 1982, 1996, 2010; Douglas, 2000; Hulstijn et al., 2007; Savignon, 1983, 2002). This study has found more evidence for the relationships among a range of components of L2 communication ability, including linguistic and non-linguistic features.

6.3 Implications for Language Testing

This study has revealed potential factors affecting communication ability from the perspective of the linguistic layperson, to whom applied linguists have rarely paid attention. The linguistic laypersons’ conceptualization of L2 communication ability can contribute to language assessment, particularly to the development of general-purpose oral proficiency tests measuring the ability to communicate in spoken English.

With regard to the construct of a test, Bachman and Palmer (2010) contend that “we need to make a conscious and deliberate choice to specify particular components of the ability or abilities to be assessed in a way that is appropriate to a particular testing situation” (p. 212). This study suggests that the construct of general-purpose oral proficiency tests should incorporate the linguistic laypersons’ perspectives, and assess a
set of abilities that affect their impressions of L2 communication ability. General proficiency tests generally aim to measure overall language ability defined in terms of theories of L2 communication ability (Davies et al., 1999), and their scores are normally used as evidence for how well test-takers can communicate in English in various real-life domains. If the construct to be measured is overall L2 communication ability in real-life domains, the tests must assess the abilities that are valued highly or considered important by stakeholders in the domains. Bachman and Palmer (2010) claim that test developers need to ensure that “the criteria and procedures for recording the responses to the assessment tasks correspond closely to those that are typically used by language users in assessing performance in TLU tasks” (p. 236). Since the ultimate arbiters of communication ability in real-life domains are unspecified linguistic laypersons, incorporating their unique perspective into the tests’ construct definition will make test scores more accurately measure real-life communication ability as perceived by real-world interlocutors or audiences. Importantly, it will enhance the validity of test score interpretation and use.

This argument aligns with that of researchers who investigated indigenous assessment criteria, or the assessment criteria derived from domain experts in LSP settings (Douglas & Myers, 2000; Elder, 1993; Jacoby, 1998; Plough et al., 2010). As reviewed in Section 2.3.3, these researchers argue that domain experts’ indigenous assessment criteria would improve the assessment of performance in specific TLU domains. In particular, such criteria are believed to be better than conventional linguistically-oriented criteria that have been widely used in oral proficiency tests because (a) L2 speakers’ performance is judged based on the perspectives of domain experts in the real-life LSP settings and (b) language proficiency is not necessarily crucial in their judgment. With regard to the importance of indigenous assessment
criteria for LSP testing, Jacoby and McNamara (1999) state that

locating indigenous assessment activities in various professional settings and analyzing them in the manner described in this paper could play a role in stimulating critical reflection on appropriate criteria for special-purpose communicative language assessment. This would result in criteria that better match the real-world expectations and demands of those responsible for professional certification. (p. 235)

Accordingly, researchers have developed rating criteria based on indigenous assessment criteria for various LSP tests (e.g., Abdul Raof, 2011; Fulcher, Davidson, & Kemp, 2011).

However, unfortunately, the construct of most current general-purpose oral proficiency tests has been narrowly defined in terms of language ability and may underrepresent L2 communication ability valued in real-life domains. McNamara (1996) argues that most general-purpose proficiency tests are weak performance tests, which are particularly concerned with the quality of the test-takers’ language production. A similar argument was made by Hughes (2002, 2011), who points out that the concept of language proficiency is heavily emphasized in language testing, and “the notion of range, structural complexity and quantity equating to valuable output is deeply embedded in the thinking of language test developers” (Hughes, 2011, p. 86). Nevertheless, these components of language ability do not necessarily align with the elements that linguistic laypersons actually value in authentic oral interactions, such as the outcome of performance, ideas, demeanor, and non-verbal behavior. Therefore, a narrowly linguistically-oriented construct may underrepresent L2 communication ability and undermine the validity of scoring or evaluation of the observed performance. Kane
(2006) argues that assessment criteria that fail to include important and relevant abilities will undermine the validity of scoring and the test overall. This issue of construct-underrepresentation weakens the validity of score interpretation and use of the test (Kane, 2006, 2013; Messick, 1989). Not omitting any important knowledge, skills, and abilities in the construct definition and assessment criteria is important, and integrating linguistic and non-linguistic features that are likely to affect people’s evaluative judgment of L2 communication ability is necessary.

The assessment criteria derived from the present study contribute to the assessment of communication ability required in a range of ELF (English as a lingua franca) settings. Since the participants in this study were from Kachru’s (1988) three circles, the findings include not only the perspective of ENL speakers, but also those of ESL and EFL speakers. Their perspectives can be useful for assessing L2 communication ability valued in ELF contexts where speakers of different L1s—including ENL, ESL, and EFL speakers—communicate in English. The criteria based on these speakers and the assessment of ability valued in ELF are innovative in language testing because current oral proficiency tests typically presume that test-takers communicate only with ENL speakers. Therefore, their assessment criteria do not necessarily take account of the ability required for effective ELF communication (Hughes, 2011; Jenkins, 2006b; McNamara, 2011). Assessing the features considered important by the participants in this study is indispensable for the assessment of communication ability valued by a wide range of English speakers in globalized society.

One of the significant characteristics of the criteria derived from the raters in this study is that conformity to NS standard English (linguistic accuracy and native-like pronunciation) is not necessarily idealized and mutual understanding is heavily emphasized. In this sense, the nature of the assessment seems to align with those
proposed by researchers (Canagarajah, 2006; Elder & Davies, 2006; Jenkins, 2006b) who suggest that speaking tests should emphasize strategic competence—code switching, language awareness, sociolinguistic sensitivity, and negotiation skills—and the use of accommodation strategies that make ELF interaction successful, rather than the accuracy of language based on NS norms. In particular, Jenkins (2006b) claims that exams—and therefore teaching—could turn their attention to rewarding the successful use of accommodation strategies and penalizing their absence, and to focusing for error correction on the use of forms that are not mutually intelligible in EIL [English as an International Language], such as NS idioms. (p. 49)

Although the use of such strategies often results in the violation of NS standard forms, the non-conformity to NS forms can lead to successful communication and does not induce negative impressions from the linguistic layperson. Seidlhofer (2011) in fact argues that “a failure to measure up to approved NS norms does not seem to prevent effective communication in non-conformist ELF taking place on a global scale” (p. 185). Since linguistic features, including native-like accuracy of grammar and pronunciation, were found peripheral in the linguistic laypersons’ judgment of communication ability, conformity to approved NS norms does not need to be part of the construct definition of oral proficiency tests.

However, this finding may not necessarily reflect the perspectives of ENL speakers; it was found that some ENL raters were sensitive to linguistic forms and negatively commented on grammar, vocabulary, and pronunciation that did not conform to NS norms (e.g., pluralization of uncountable nouns, pronunciation of /θ/), although they were intelligible. Accordingly, spoken English that conforms to standard NS
English may have to be included in the construct if the test’s purpose is to assess the L2 communication ability valued by ENL speakers only. Nevertheless, conformity does not need to be weighted as highly as achievement of communicative goals or comprehensibility because the ENL raters, like EFL and ESL raters, at least consciously acknowledged that mutual understanding should be prioritized in communication.

Selecting the components of criteria and deciding relative weights of criteria for oral proficiency tests have more specific practical implications. In particular, results will be useful when determining (a) the components of assessment criteria and (b) their relative weights. The following are implications applicable for both monologic and dialogic tasks. It should be noted that the implications are not applicable to any monologic and dialogic tasks but only to tasks similar to those used in the study: informative speeches and factual interactions (see Sections 3.3.1 and 3.3.2). Moreover, the implications may be restricted to the assessment of the oral performance of intermediate L2 speakers.

First, communication outcomes should be emphasized and assessed. This aligns with the nature of strong performance tests, in which the fulfillment of the task is the main criterion (McNamara, 1996). In the case of tasks with tangible communicative outcomes (e.g., obtaining a professor’s agreement to write a letter of recommendation), directly assessing task outcomes and effectiveness of task achievement would be possible and appropriate (see task-based assessment: Ellis, 2003). In contrast, in cases where there is no tangible outcome of communication (e.g., informative presentations on air pollution), message conveyance or comprehensibility needs to be a basic criterion for successful communication. Accordingly, strategies for message conveyance such as circumlocution or grammatical reduction should be highly valued. Referring to the comprehensive list of lexical, grammatical, and phonological problem-solving
mechanisms provided by Kormos (2006) and Cohen (2011) would be helpful. Furthermore, L2 speakers should be encouraged to use interactional strategies, such as accommodation or negotiation, which are particularly effective in a wide range of settings where no ENL speakers are present (Elder & Davies, 2006; Jenkins, 2000, 2006b; Seidlhofer, 2004).

Second, assessment criteria should include the content of oral performances, such as relevance to the topic and elaboration of topic. Like influential criteria for assessing writing (Jacobs, Zinkgraf, Wormuch, Hartfiel, & Hughey, 1981; Spandel & Stiggins, 1977), the assessment of oral communication ability could focus on the quality of content. The criterion entitled *Topic Development* in the rating scales of the TOEFL iBT (ETS, 2009) can be a useful reference source.

Third, language proficiency should be assessed, but intelligibility should be emphasized more than accuracy. In fact, grammar, vocabulary, and pronunciation were found to have a direct impact on comprehensibility and raters’ attention to speech. For low-proficiency speakers, language proficiency, particularly pronunciation, should be more heavily weighted because comprehensibility and message conveyance would be seriously impeded by lack of proficiency. However, conformity to NES norms or native-like accuracy should not be required as long as linguistic forms do not hamper intelligibility. More specifically, innovative ELF lexicogrammar (see Table 2.11) and L1 accent with the Lingua Franca Core features (Jenkins, 2000) should not be penalized. Assessment of pronunciation might be meaningless for high-proficiency test-takers who can successfully convey messages using accented English. This implication aligns with Pollitt and Murray’s (1996) findings of what attracts raters’ attention in oral performances at different proficiency levels (see Figure 2.5).

Finally, non-verbal behavior, in particular gestures and eye contact, should be
assessed. Appropriate non-verbal behavior (e.g., iconic gestures) not only facilitates comprehensibility but also greatly influences perceived levels of confidence, anxiety, and engagement. Thus, appropriate use of gestures and eye contact should not simply be regarded as a coping strategy to be adopted by low-proficiency speakers. Instead, they should be assessed and emphasized in all levels of test-takers. Currently, few oral proficiency tests explicitly assess gestures and eye contact (Sato, 2013). Some examples of criteria or rating scales including these aspects of communication are a performance decision tree for a service encounter (Fulcher et al., 2011) and Nonverbal Ability (NOVA) Scales (Jungheim, 2001). The former assesses eye contact, facial expressions, and posture by making a dichotomous decision, whereas the latter are for assessing multiple aspects of head nods, gaze, and hand gestures. These criteria are useful as references for the development of criteria that assess test-takers’ non-verbal behavior.

The following is an implication primarily for monologic tasks of the kind used on a general-purpose proficiency test. Fluency should be emphasized more strongly than linguistic accuracy and pronunciation. In monologic tasks, in particular planned speech, fluency is likely to affect comprehensibility, the listener’s attention to content, and impressions of demeanor and preparedness. Fluency is a ubiquitous criterion of oral proficiency tests, and often focuses on the flow of speech and language-related hesitations (Sato, 2013). The weight of the criterion should be higher than that of linguistic resources and pronunciation unless speakers’ language forms seriously impede comprehensibility. This aligns with the relative contributions of linguistic features according to proficiency levels found by Higgs and Clifford (1982). Among the components of fluency, pause phenomena within clauses should be penalized, whereas slow utterance speed and pauses at clause junctures should be less strictly penalized. Self-correction for linguistic errors should not be encouraged even though this is
regarded as a speaking strategy (Cohen, 2011).

Lastly, the following is an implication primarily for dialogic tasks of the kind used on a general-purpose proficiency test. Speakers’ interactional features and behaviors should be assessed and emphasized. In fact, it is not uncommon for oral proficiency tests to assess test-takers’ interaction, focusing on turn-taking, contribution to interaction, and comprehension of partners (Sato, 2013). Since interaction is likely to influence the impression of communication ability more strongly than linguistic accuracy, assessment of the interactional features and behaviors is necessary in dialogic tasks. Furthermore, these behaviors are not necessarily captured by linguistically-oriented assessment criteria. A test should penalize those only displaying their own language proficiency and encourage altercentrism and interaction management related to supportiveness, clarification seeking, topic initiation, questioning, and responding (Spitzberg & Hurt, 1987). In contrast to interaction, fluency may not deserve a heavy weight and dysfluency features should not be severely penalized in spontaneous interactive tasks.

6.4 Summary

This chapter has discussed implications of this study for the theory of L2 communication ability and language testing. The conceptualization of L2 communication ability based on the linguistic laypersons’ perspectives is beneficial for various stakeholders in applied linguistics and language education since the theory can inform them about the knowledge, skills, and abilities needed to be viewed as a competent communicator. This study also suggests that the construct of general-purpose oral proficiency tests should incorporate the linguistic laypersons’ perspectives and assess a set of abilities that strongly affect their impressions of L2 communication
ability to enhance the validity of score interpretation and use of the tests.
Chapter 7 Conclusion

7.1 Introduction

This chapter concludes the present thesis by evaluating the significance and importance of the study. First, it presents a summary of the key findings of the study. Second, the significance of the study is addressed. Finally, the chapter presents limitations of the study and recommendations for further research.

7.2 Summary of the Study and Key Findings

The objective of this study was to explore the features that influence the linguistic laypersons’ judgment of L2 speakers’ oral communication ability. Twenty-three non-native and native English-speaking graduate students from disciplines other than applied linguistics and TESOL (raters) participated in this study. First, they watched video recordings of seven individual presentations on the College English Test-Spoken English Test (CET-SET) and three paired interactions from the Cambridge Main Suite Examinations. Next, the raters indicated their intuitive impressions of each speaker’s communication ability on a scale of 1 (Poor) to 7 (Excellent), and provided justifications for their ratings. Then, they reviewed the same performances and performed stimulated recall, verbalizing features of the performances that influenced their impressions. Finally, they responded to semi-structured interviews to provide supplementary information about their judgments. The ratings provided by the raters were analyzed using FACETS statistical software. Features that had affected their impressions were explored in the verbal protocol data. The interview data were also analyzed qualitatively to examine what features were perceived to most strongly influence their impressions.
The results showed that the raters’ judgments did not necessarily correlate with the assessments obtained from the proficiency tests. The raters’ verbal protocol data showed that the outcome of communicative performances—successful message conveyance and the quality of content—was one of the main criteria used to judge communication ability. Although linguistic features such as grammar, vocabulary, pronunciation, and fluency were recognized as factors influencing the outcome of communication, the raters neither considered them crucial nor penalized errors harshly unless comprehensibility was seriously impeded. Their impressions were also influenced by the test-takers’ non-verbal behaviors and perceived level of confidence, which were found to be closely intertwined. In addition, the raters frequently noted interactional features such as engagement and the extent of an individual’s contribution in the paired interactions.

7.3 Significance of the Study

This study defined L2 communication ability from the perspective of the linguistic layperson. Speech features and behaviors of speakers explored in the study are those affecting the degree to which an L2 speaker is viewed as a competent communicator by audience or interlocutors in similar communicative settings to real life. Furthermore, this study explored the role of both language ability and non-linguistic features, such as Hymes’s (1972) ability for use, in the linguistic laypersons’ judgment of L2 communication ability. This study also revealed the components of L2 communication ability valued by a wide range of English speakers, including EFL and ESL speakers. It was not restricted to ENL speakers; thus, the features explored in the study reflect the knowledge, abilities, and skills that are required to be perceived as a competent communicator in globalized settings where English is used as a lingua franca.
The perspectives of linguistic laypersons in the study broadened the scope of theories of L2 communication ability, which in turn informed the assessment criteria used to measure this ability. In particular, assessment criteria comprising the features that affect the linguistic laypersons’ impressions of communication ability makes oral proficiency tests measuring the ability more valid because the test score more accurately predicts the impressions of the ultimate arbiters in real-life settings. Assessment criteria based on a wide range of English speakers also contribute to the assessment of the ability valued in various ELF settings where ENL speakers are not always present.

7.4 Limitations and Recommendations

The present study used low- to high-intermediate L2 speakers’ oral performances on two types of elicitation task as stimuli: short individual oral presentations and task-oriented dyadic discussions. Since what features are influential are considered to be different according to speakers’ language proficiency and communicative tasks (Chalhoub-Deville, 1995, 1996; Pollitt & Murray, 1996), factors not explored in this study could be salient to the linguistic laypersons’ judgment of communication ability if different speakers’ performances on different tasks were used. Further investigation is recommended by using a variety of elicitation tasks performed by different levels of L2 speakers.

The primary purpose of this study was to explore what affects the linguistic laypersons’ judgment of communication ability. Although the study identified some specific features that positively or negatively affected their judgment (e.g., iconic gestures positively affected linguistic laypersons’ impressions), future research could further examine what features and behaviors within each category might have influenced their judgments or indeed escaped their notice. For example, Rater H
mentioned that she had not noticed any grammatical errors in the presentation of CET A+; it would be possible to reveal what errors would likely be unnoticed by conducting discourse analysis and examining actual grammatical errors made by the speaker. This could be investigated through the *subject-object approach* in which researchers examine the relationship between the raters’ subjective judgments and objective measures of performances (De Jong et al., 2012). Objective measures of oral performances could be obtained through discourse analysis and conversation analysis.

This study collected data from a small number of linguistic laypersons. Moreover, they were those with high educational background, high English proficiency, various L2 learning experiences, and relatively great familiarity with English spoken by NNESs. Since these factors influence the perception of L2 speakers’ ability, further research needs to investigate perspectives of linguistic laypersons with a wider range of experiences, characteristics, and backgrounds. For example, this study speculated that raters’ L1 background would influence the impression of communication ability and found some differences in the perspective on linguistic accuracy among EFL, ESL, and ENL raters. Due to a small sample size, however, it was not possible to systematically investigate the perceptional differences among raters with different L1 backgrounds, as some researchers have done in the past (e.g., Brown, 1995; Fayer & Krasinski, 1987; Kim, 2009a, 2009b; Zhang & Elder, 2011, 2013).

The relative contribution of features to judgments of communication ability was inferred based on the frequency of comments and post-session interview data. However, these data do not necessarily represent the degree to which the feature affected their impression. This limitation is related to the methodological nature of verbal protocol analysis and self-reports (Cohen, 1994; Smagorinsky, 2001). An informant cannot give an account of automated processes and may provide what they
assume that researchers wish to know. Accordingly, this study may not have revealed features that were not consciously attended to but that were nevertheless potentially highly influential. Judgments at the unconscious level could be investigated using different approaches to raters’ judgments: (a) an analysis of the relationship between holistic and analytic ratings and (b) an analysis of the relationship between raters’ judgments and objective measures. The results of the current study can contribute to further research using these quantitative approaches. It would be possible for example to seek statistical confirmation of the relative weights of the factors identified by the raters in this study. Such approaches would also be beneficial in collecting a larger data set from more participants with various backgrounds.

Finally, linguistic laypersons’ perspectives on language and communication have seldom been regarded as a legitimate source of theory construction and construct definitions. However, laypersons should surely contribute to the field of applied linguistics because the field itself is essentially for the benefit of laypersons, as Wilton and Stegu (2011) have argued. As the present study has demonstrated, their perspectives have many commonalities with claims made by language specialists, demonstrating that these so-called “linguistic laypersons” are not completely linguistically naïve. At the same time, there are still many aspects of perspectives that cannot be easily explained by theories and language specialists’ perspectives. Incorporating linguistic laypersons’ perspectives into a theory’s construction is believed to make it more robust and more beneficial for the field of applied linguistics.
References


Barnwell, D. (1989). ‘Naive’ native speakers and judgements of oral proficiency in


Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives:


development of a preliminary analytic framework \textit{TOEFL Monograph Series.}


Jungheim, N. O. (2001). The unspoken element of communicative competence:

Evaluating language learners’ nonverbal behavior. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 1-34). Honolulu: University of Hawaii Second Language Teaching and Curriculum Center.


Lowenberg, P. H. (1993). Issues of validity in tests of English as a world language:


*Melbourne Papers in Language Testing, 1*(2), 19-44.


competence. In G. Rickheit & H. Strohner (Eds.), *Handbook of communication competence* (pp. 15-62). Berlin: De Gruyter Mouton.


Smagorinsky, P. (1994). Think-aloud protocol analysis: Beyond the black box. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research*


Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?

Appendix 1 Plain Language Statement

**Project title:** “Factors Underlying Second Language Communicative Ability: The Perspective of Linguistic Laypersons on Second Language Oral Performance”

**Introduction**
Normally, linguists and language teachers are solely responsible for defining what it means to be able to communicate in a second language. Accordingly, English language tests (e.g., IELTS) are often intended to assess the communication ability defined by these professionals. However, the definition of “good communication” should include the evaluative judgment made by linguistic laypersons because non-native English speakers are more likely to communicate with them than language specialists. Therefore, we are interested in what consists of second language communication ability defined by linguistic laypersons. Thus, this project will investigate the factors underlying people’s evaluative judgments of second language communication ability. More specifically, this project aims to discover the features or behaviors that linguistic laypersons strongly focus on when watching speaking performance of non-native English speakers. These features or behaviors will contribute to the development of better assessment criteria of oral English proficiency tests. This project has been approved by the Human Research Ethics Committee.

**What you will be asked to do**
As a research participant, you will be asked to do the following:
1. Fill in the background information form.
2. Read the instructions on what you are asked to do (Steps 3 to 6 below) and receive basic training.
3. Watch videos of speaking performances of non-native English speakers.
4. Indicate your impression of each speaker’s communication ability by choosing one of the 7 levels from Poor to Excellent.
5. Talk about the reasons for your rating.

6. Watch the performance again and talk about every detail of performance that affected your impressions.

7. Answer some interview questions.

Your talks will be audio recorded. The whole process will take about two hours. We can identify no possible risks to you if you take part in this project.

**How your confidentiality will be protected**

We intend to protect your anonymity and the confidentiality of your responses to the fullest possible extent, within the limits of the law. We will keep your name and contact details in a separate, password-protected computer file from any data that you supply. Only researchers involved in this project will have access to this data. The data will be kept securely in the School of Languages and Linguistics at the University of Melbourne for seven years after the completion of the project. The materials will then be destroyed. In any publication and/or presentation, we will provide information in such a way that you cannot be identified, except with your permission. We will keep confidential any information obtained in connection with this project that can identify you and we will only use it for the purpose of this research project (subject to legal limitations). In accordance with relevant Australian and/or Victorian privacy and other relevant laws, you have the right to access the information collected and stored by us about you. Contact one of the researchers named above if you would like to access your information.

**How you will receive feedback about the completed project**

If you are interested to learn what the results of this project are, you can give us your contact information in the consent form. When the project is completed, we will send you a brief summary of the findings and any information about any publications based on the research.

**How participation or non-participation affects you**

Your participation in this project is voluntary. If you want to withdraw at any stage, or to withdraw any data you have supplied, you are free to do so without prejudice.

**How you can get further information**

If you require any further information, or have any concerns, contact us by email or telephone using the details at the top of the first page. If you have any concerns about the conduct of the project, contact the Executive Officer, Human Research Ethics,
How you agree to participate
If you would like to participate, contact the student researcher. You will need to complete and sign the consent form to show that you have read and understood this information. You may want to keep a copy of the form for yourself.
### Appendix 2 Background Information Questionnaire (NNES)

Before you begin the actual survey, please answer the following background questions. All of this information will be kept confidential.

<table>
<thead>
<tr>
<th>Question</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>20-25</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>Female</td>
</tr>
<tr>
<td><strong>Nationality</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Mother tongue</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Current academic course</strong></td>
<td>MA</td>
</tr>
<tr>
<td><strong>Faculty/Department</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Area of academic expertise</strong></td>
<td></td>
</tr>
<tr>
<td><strong>How long have you been in Australia?</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Your most recent TOEFL/IELTS score</strong> (including listening and speaking scores if you remember)</td>
<td>TOEFL Score:</td>
</tr>
<tr>
<td><strong>How old were you when you started to learn English?</strong></td>
<td>years old</td>
</tr>
<tr>
<td><strong>Where did you learn English mostly? Please check all that apply.</strong></td>
<td>Home</td>
</tr>
<tr>
<td><strong>Who were your main English teachers?</strong></td>
<td>Non-native English speaker</td>
</tr>
<tr>
<td><strong>How did you mostly learn English? Please check all that apply.</strong></td>
<td>Studied grammar and vocabulary</td>
</tr>
<tr>
<td><strong>Have you learned any foreign language(s) other than English?</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>Language:</strong></td>
<td>How long?</td>
</tr>
<tr>
<td><strong>Have you ever taken any courses on applied linguistics?</strong></td>
<td>No</td>
</tr>
</tbody>
</table>

333
<table>
<thead>
<tr>
<th>Have you ever taken any courses on language teaching?</th>
<th>No ☐</th>
<th>Yes ☐</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have you ever taken any courses on communication studies?</td>
<td>No ☐</td>
<td>Yes ☐</td>
</tr>
<tr>
<td>Have you ever taught English to non-native speakers as a tutor or instructor?</td>
<td>No ☐</td>
<td>Yes ☐</td>
</tr>
<tr>
<td>To whom?</td>
<td>How long?</td>
<td></td>
</tr>
<tr>
<td>How often have you had contact with people speaking with the following accents when they speak English?</td>
<td>No Contact</td>
<td>Some Contact</td>
</tr>
<tr>
<td>Chinese</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Korean</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>French</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Italian</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Swiss</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Saudi</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Mexican</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>How familiar are you with English spoken by non-native speakers in general?</td>
<td>Not familiar</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If you are familiar with particular non-native speakers' spoken English, please write the nationality of the persons.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3 Background Information Questionnaire (NES)

Before you begin the actual survey, please answer the following background questions. All of this information will be kept confidential.

<table>
<thead>
<tr>
<th>Age</th>
<th>20-25</th>
<th>26-30</th>
<th>31-36</th>
<th>36-40</th>
<th>40 or over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mother tongue

Current academic course | MA | PhD |

Faculty/Department

Area of academic expertise

How long have you been in Australia?

Have you learned any foreign language(s) other than English? No | Yes
Language: How long?
Language: How long?
Language: How long?

Have you ever taken any courses on applied linguistics? No | Yes

Have you ever taken any courses on language teaching? No | Yes

Have you ever taken any courses on communication studies? No | Yes

Have you ever taught English to non-native speakers as a tutor or instructor? No | Yes
To whom?
How long?

How often have you had contact with people speaking with the following accents when they speak English?

<table>
<thead>
<tr>
<th></th>
<th>No Contact</th>
<th>Some Contact</th>
<th>Regular Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swiss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationality</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>-------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Saudi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexican</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How familiar are you with English spoken by non-native speakers in general?
Not familiar 1 2 3 4 5 Very familiar

If you are familiar with particular non-native speakers’ spoken English, please write the nationality of the persons.
Appendix 4 Written Instructions for Verbal Protocol

INSTRUCTIONS

4 STEPS & PROCEDURES
First, you will be asked to watch speaking performances of non-native English speakers in some English tests. **While watching each performance, please judge the speaker’s COMMUNICATION ABILITY.** Please follow your impression and focus on any aspects of the performances as you like. You can take notes of your impression.

*A sample image*
STEP 2

After watching each performance, please indicate your impression of the speaker’s communication ability on the rating form.

Rating form

Please place a “✓” on one of the 7 positions from POOR to EXCELLENT. For example, if you think that the communication ability of a speaker is slightly below excellent, you can place the “✓” as follows:

Please talk about the reasons for your rating. You can begin to talk immediately after indicating your impression of the speaker’s communication ability on the form. Please talk to the IC recorder in front of you. Please try to say as much as possible.
STEP 4

After talking about the reasons, you will be asked to watch the performance again. This time, I would like you to elaborate on the reasons you have mentioned. Please stop the video every time you find anything that affected your impression, and talk about it in detail. Stop as frequently as you can and comment anything you like. Please comment in as much detail as you can.

Click on this button to stop and replay the video.
Procedures

Before the real experiment, I would like you to practice using a sample performance. Afterwards, you will move on to the real experiment phase and watch 10 performances in total. There are 7 individual presentations and 3 pair conversations.

I am sitting behind you and will not interrupt you if not necessary. If you have any questions, please ask me anytime.
Appendix 5 Rating Form (Sample)

Real Experiment (1/10)

Individual Presentation
Topic: “Causes of Air Pollution”

Note

His communication ability is:

Appendix 6 Post-session Interview Questions

About Judging Communication Ability and Commenting
1. You watched and judged 10 performances. Was it difficult to judge their communication ability?
2. You talked about the things that affected your impression. Do you think the same things affect your impression of people in a daily life too?
3. (To NNES participants) Was it difficult to talk in English? Would it be easier if you were allowed to use your mother tongue?
4. The speakers talked about topics like environmental issues or the places they live. Were these topics interesting?

Questions about the Heeded Features
1. Could you tell me three things that mostly affected your impressions of communication ability?
2. Now, I am going to tell you some aspects of communication. Please tell me how strongly these aspects affected your impression of the speakers’ communication ability. Some of them may be redundant.
   a. Grammatical accuracy
   b. Vocabulary
   c. Pronunciation or accent
   d. Fluency or flow of speech
   e. Non-verbal behaviors (gestures, facial expressions, eye contact)
   f. The content of the speech
   g. The organization of the speech
   h. Confidence

Conclusion
1. Is there anything else you would like to comment on?
Appendix 7 Transcription Notation Symbols

(from Atkinson & Heritage, 1984 pp. ix-xvi; Lazaraton, 2002, pp.60-72)

1. **Numbers within parentheses** – periods of silence, timed in second and inserted within parentheses.

2. **Colon (:)** – a lengthened sound or syllable; more colons prolong the stretch.

3. **A period (.)** – a stopping fall in tone, not necessarily the end of a sentence; **a question mark (?)** – a rising intonation, not necessarily a question; **an exclamation mark (!)** – an animated tone, not necessarily an exclamation.

4. **hhh** – exhalation; **hhh!** – strong exhalation.

5. **.hhh** – an inhalation; **.hhh!** – strong inhalation.

6. **(hhh)** – breathiness within a word.

7. **Percent signs (% %)** – a quiet passage of talk.

8. **tch** – a tongue click.

9. **Underlined syllable** – an emphasized syllable.

10. **Parentheses ( )** – uncertain words or expressions; empty parentheses indicate utterances could not be transcribed.

11. **Double parentheses (( ))** – non-vocal action.

12. **Italics** – English translations of Japanese words or expressions written above.
Appendix 8 NVB and Rater Perceptions

Asad

*the houses [is too big]*\(^7\).

One thing I notice was Asad likes using hand gestures to associate with words. (K-Asad-SR-1575)

Chulsoo

*the room is (.2) bigger than* previous one.

Around here, he incorporates gestures in a good way, for example, to express bigger. (E-Chulsoo-SR-636)

Asad

*so that’s you know make the family [.2] err not always together.*

So the fact that he was using his hand to gesture “the family being together” was quite good. (V-Asad-SR-3012)

\(^7\) Brackets mark gesture phrases.
**Chloe**

*but for example? [this one? I think I prefer do this job than]*

but Chloe expresses herself more in body language because she would point to the pictures and stuff like that.
(M-Chloe-SS-4653)

and then she points another one.
(G-Chloe-SR-951)

---

**Anita**

*I think (.2) for example [in this] first one? the kids [can’t enjoy (.)] the same (mess) in a] sunny day in a summer day?*

So uses her hands. (U-Anita-SR-2833)

Body language is positive. / She’s gesturing with her hands.
(W-Anita-SR-3162, 3163)

---

**CET A**

*I think err the err the [private] cars*

He’s using some hand gestures even though there’re quite slight gestures. / It’s engaging to who is speaking with.
(Q-A-SR-2307, 2308)

The whole time I was a bit distracted by how he was holding his cue card / but he was also waving it around.
(K-A-SS-4413, 4414)
CET A
(She maintains this hand position during her speech)

I like the way she put her hands together. / I mean she interlocked it. / And she inter-raised it. (M-A⁺-SR-1800, 1801, 1802)

Her hands are crossed, which totally hampers any forms of body language, gestures and so on. (F-A⁺-SS-3853)
Appendix 9 Phonological Features Referred by Raters

- So he said *development* and sound like /w/ instead of /v/. (R-B-SR-2417)
- Sometimes all turns l into r, / he says instead of “air pollution,” *air poorrution*. (W-D-SR-3139, 3140)
- For example, there was one spot where he said *voom* instead of “room.”
  (T-Chulsoo-SS-5330)
- Pronunciation *factories*. / She has a little bit problem (D-A+-SR-414, 415)
- *Factors* kind of sound like “factories” to me. (W-B-SR-3075)
- And not later on that she’s gonna say /chri:z/ instead “trees” /tri:z/ you know. (M-C°-SR-1834)
- And *plastic* / and I don’t know how framed the word *plastic*. (M-C°-SR-1836, 1837)
- And then she was trying to say “sociable,” / she said /soככהב/c/. (R-Aida-SR-2481, 2482)
- So she said *pits* instead of “pets.” (R-Anita-SR-2462)
- He could’ve pronounced *emission* slightly differently, (T-A-SR-2595)
- You hear she said /kits/ instead of saying /kidz/, (T-Anita-SR-2648)
- The word “ideas” sounded more like *idears*. / Just a small thing. (V-A-SR-2895, 2896)
- “Dimensions” sounded like *dimenions*. (V-A-SR-2897)
- “Secondly” sounded like *secondary*. (V-A-SR-2902)
- *Human body* /bʌdi/ instead of “human body” /baːdi/, (V-S-SR-2953)
- She says *tropical* /trɒpɪkəl/ instead of “tropical” /trɒpɪkəl/; which is the way maybe I as a New Yorker or American or native English speaker. (W-Marie-SR-3180)
- Some of the words that Aida’s saying *people* /pi:plu:/ stands to me
  (W-Aida-SR-3220)
- So, there are … you can notice some pronunciation problems like *be*. (D-Anita-SS-3698)
- In addition, there are a few mispronunciations when she says strong message as well as the word *awareness*. (W-A°-SS-5564)
- His pronunciation of the words *contaminating materials* was a bit off in my opinion. (W-A-SS-5578)
Appendix 10 Vocabulary and Wording Referred by Raters

- So of course or absolutely is just a bit unnecessary. (V-B⁺-SR-2911)
- How worse is the air pollution. / Worse. / How “bad” it is or how … what are the consequences. (G-C⁺-SR-849, 850, 851)
- Again, the usage of harm to the earth. / I don’t think that’s a better way of expressing herself. (G-C-SR-878, 879)
- Many different furniture. / Probably (understood in) “different signs” or “interior designs” in the same place. (G-Asad-SR-991, 992)
- I think it’s a bit unclear what means by private cars. (K-A-SR-1514)
- So maybe that’s not the best usage of the word illustrate. (N-A-SR-1925)
- I think she means, maybe she means “commentator” or something there, one of the sports journalists. (N-Chloe-SR-2029)
- I think he said that, the government has got a much responsibility when I think it should be “the government has got a lot of responsibility.” (R-A-SS-5088)
- So he’s using gasolines instead of “gas” or “petro.” (U-B-SR-2759)
- Another incorrect usage of phrase when she says look forward to the Tibet sky. (V-B⁺-SR-2921)
- And he says actually. / … / That’s not word that I would expect one to use in a response to a question. (W-A-SR-3040, 3041, 3042)
- So whereas he says take the public transport but “public transportation.” (W-A-SR-3045)
- The word strenuous … I thought it was really well put together in / the word strenuous is not … I don’t think really basic English word. / He does a nice job of using that here in this context. (W-A-SR-3047, 3048, 3049)
- The use of word promotion probably wouldn’t be the word that choice that I would expect to use or other native English speaker. (W-C⁺-SR-3094)
- but … and the other one the factory probably he was thinking of “industrial.” (G-A-SS-3969)
- and at the end she said something of a bad sky. / I mean “bad weather” okay, (G-B⁺-SS-3988, 3989)
- He used words like classic and traditional, which shows that he has good conversation skills and he picks off the right words. Good words rather. (P-Asad-SS-4986)
- But she seemed to have quite a good vocabulary to express herself. / So I think she had the word tsunami or something. (S-Marie-SS-5227, 5228)
- He says among other things, which I … that was a positive for me. (W-B-VP-5608)
Appendix 11 Grammatical Features Referred by Raters

- She adds the words *how it worse.* / And that makes it just … she’s adding that for whatever reason and you … takes away from the meaning of the sentence. (V-C+-SR-2944, 2945)
- She also used the double negatives. / So she said *can’t barely,* which should have been “can barely.” (R-Marie-SS-5146, 5147)
- And *how worse* or *how was air pollution* is … it’s not the right tense. (W-C+-SR-3095)
- So she says *gossips* instead of “gossip.” / She pluralizes the singular word. (W-Aida-SR-3223, 3224)
- and then Chulsoo comes back with *a big family yeah?* / “Yeah” is not actually a question. (W-Chulsoo-SR-3247, 3248)
- *Everything is flood* is well … is not strictly correct to put it this way in English. (F-Anita-SR-722)
- But she used *many many.* / I’m not sure whether it’s a correct grammar in English. (H-C-SR-1064, 1065)
Appendix 12 Other Behaviors/Features Referred by Raters

- I feel like his English has a little more influence from Americans cuz he uses a lot of you know you know kind of thing. (O-Asad-SR-2141)
- But what I’d say is that again I would doubt that as trying to choose an easier topic so that she could speak more on it. (P-Aida-SR-2246)
- So maybe there’s a … there’s a confusion bet … between her mother language and her English to combine them together to convey her thinking. (J-C-SS-4368)
- He actually tries to draw to attention in a successful manner, successfully. (A-A-SS-3292)
- and I think she had always been thinking about what she should say, what she could answer Anita’s answers and continue this conversation. (H-Marie-SS4198)
- And about the examples, I think she was quite good, giving impromptu sentences. (K-B−-SS-4420)
- I mean, even emotionally nobody is trying to impose her idea. (A-Asad/Chulsoo-SR-116)
- Probably she’s looking for appealing. (G-Chloe-SR-941)
- In this part she mentioned what the other girl just answered to this question. (H-A−-SR-1001)
- And he wasn’t looking for different way to say it. (O-D-SR-2110)
- He has a notepad in front of him, which I am sure he wrote down his point. (M-A-SS-4593)
- She comes across as an art student. (U-B−-SS-5358)
- and she tries to give more meaning to her sentences. (C-Chloe-SR-384)
- And some point he realizes he cannot get over it and cannot proceed it. (G-D-SR-895)
- Anita seems to notice that both should be speaking a little bit. (W-Anita-SR-3172)
- and yet at the same time I wonder if she is regurgitating some fact from somewhere she read. (M-C-SR-1860)
- I know that how she speaks in English doesn’t necessarily correspond to what she understands, (T-Anita-SR-2668)
- And she also ended her speech with thank you, which I think is pretty standard for public speaking. (K-B−-SR-1525)
- And he looks like he’s asking for some divine help. ((laughs)) (G-D-SR-904)
- Again, he is talking to himself in some sense. (G-D-SR-903)
- Is the text that in front of him, he do not understand? (I-D-SR-1284)
Author/s: 
SATO, TAKANORI

Title: 
Linguistic laypersons' perspective on second language oral communication

Date: 
2014

Persistent Link: 
http://hdl.handle.net/11343/42017

File Description: 
Linguistic Laypersons’ Perspective on Second Language Oral Communication Ability