

An investigation into variability of tasks and teacher-judges
in second language oral performance assessment

Youn-Hee Kim
Integrated Studies in Education
McGill University, Montreal

August 2005

A thesis submitted to McGill University in partial fulfillment of the requirements
for the degree of Master of Arts

© Youn-Hee Kim, 2005



Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 978-0-494-28652-4

Our file Notre référence

ISBN: 978-0-494-28652-4

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.


Canada

Abstract

This study examined how second language oral performance is assessed by different groups of teacher-judges across different tasks and task types. The substantive focus of the study investigated whether native English-speaking (NS) and non-native English-speaking (NNS) teacher-judges exhibited internal consistency and interchangeable severity, and how they influenced task difficulty and the calibration of rating scales across different tasks and task types. It was also identified what the salient construct elements for evaluation were to the two groups of teacher-judges across different tasks and task types when no evaluation criteria were available for them to consult.

A Many-faceted Rasch Measurement analysis of 1,727 ratings and a grounded theory analysis of 3,295 written comments on students' oral English performance showed little difference between the NS and NNS groups in terms of internal consistency and severity. Additionally, the two groups were neither positively nor negatively biased toward a particular task type. The NS and NNS groups, however, did differ in how they influenced the calibration of rating scales, and in how they drew upon salient construct elements across different tasks and task types. The need for context (task)-specific assessment, the suitability of the NNS teacher-judges, the usefulness of the Many-faceted Rasch Measurement, and the legitimacy of mixed methods research are discussed based on these findings.

Résumé

Cette étude avait pour but d'examiner la manière dont différents groupes d'enseignants/examineurs évaluaient la performance à l'oral en seconde langue à travers différentes tâches et types de tâches. L'objectif principal de l'étude était de déterminer dans quelle mesure les enseignants/examineurs de langue maternelle anglaise (LMA) ou n'étant pas de langue maternelle anglaise (NLMA) exerçaient une influence sur la difficulté de la tâche et le calibrage des échelles de notation pour les différentes tâches et types de tâches et dans quelle mesure ils faisaient preuve d'une homogénéité et d'une sévérité constantes. Cette étude a également permis d'identifier quels étaient les éléments structurels saillants pour les deux groupes d'enseignants/examineurs selon les différentes tâches et types de tâches quand aucun critère d'évaluation n'était disponible à la consultation.

Une analyse multi facettes selon le modèle de Rasch de 1727 évaluations et une analyse en théorie ancrée de 3295 commentaires écrits portant sur la performance à l'oral d'étudiants en anglais ont montré peu de différence entre les groupes LMA et NLMA en termes d'homogénéité et de sévérité. Elles n'ont pas non plus démontré une différence significative dans l'influence exercée sur la difficulté de la tâche dans les différentes tâches et types de tâches. Néanmoins, les groupes LMA et NLMA différaient dans leur manière d'influencer le calibrage des échelles de notation et d'utiliser les éléments structurels saillants pour les différentes tâches et types de tâches. La nécessité d'une évaluation spécifique liée au contexte (tâche), la pertinence des enseignants/examineurs NLMA, l'utilité d'une analyse multi facettes selon le modèle de Rasch et la légitimité d'une recherche menée avec des méthodes variées font actuellement l'objet de discussions sur la base de ces résultats.

Acknowledgements

This research project could not have been accomplished without the help and support of many people. I would like to thank my supervisor, Dr. Carolyn Turner, for providing me with the advice, patience, and enthusiasm that inspired me to accomplish this research project. Her expertise in second language testing and evaluation was an invaluable resource, and she provided a great deal of insight and guidance.

My appreciation goes to Dr. Michael Linacre for taking the time to answer my persistent questions about FACETS. His extensive knowledge of both FACETS and the Many-faceted Rasch Measurement was the basis for many suggestions with regard to data analysis, including the identification of more proficient methods of conducting it. I would also like to thank Dr. Carol Myford, at the University of Illinois at Chicago, for her suggestions of a rater effects analysis, and my extraordinary friends, Talia Isaacs, Kazufumi Miyagi, and Zhidong Zhang, who took the time to offer thoughtful and critical feedback when it counted the most.

I am grateful to Mr. Jaewoon Choi, the former principal of Daegu Foreign Language High School, and Dr. Yae-Sheik Lee at Kyungpook National University. Their intellect and insights were an inspiration, and their encouragement and trust made this study possible. My warmest thanks go out to my family in Korea. It would not have been possible for me to carry out my research without their love, patience, and understanding.

This research project was fully supported by the Internal Social Sciences and Humanities Research Grant, funded by McGill University. I am also deeply grateful for the financial support by the Canadian Chamber of Commerce in Korea and the Quebec Ministry of Education, which enabled me to continue my Master's program in Canada for two years.

Table of Contents

Abstract.	i
Résumé.	ii
Acknowledgements.	iii
Table of Contents.	iv
List of Tables.	viii
List of Figures.	x
 CHAPTER 1: INTRODUCTION.	 1
Research Rationale.	1
Chapter Overview.	3
 CHAPTER 2: REVIEW OF THE LITERATURE.	 4
Overview of Second Language Performance Assessment.	4
Theoretical Models of Second Language Performance Assessment.	7
Systematic Variations of Second Language Performance Assessment.	12
Rating scales.	13
Tasks.	18
Raters.	21
 CHAPTER 3: RESEARCH QUESTIONS AND METHODOLOGY.	 27
Research Purpose and Questions.	27

Participants.	29
Students.	29
Native English-speaking teacher-judges.	31
Non-native English-speaking teacher-judges.	33
Instruments.	35
The Computer-Assisted Test of Oral English (CATOE).	35
The CATOE rating scale.	39
Background questionnaires.	40
Procedure.	41
The CATOE administration.	41
The CATOE scoring procedure.	42
Data Analysis.	44
Many-faceted Rasch Measurement analysis.	44
Grounded theory analysis.	50
 CHAPTER 4: RESULTS AND DISCUSSION.	 55
Calibration of Students, Teacher-Judges, and Tasks.	57
Systematic Variability Findings in the Tasks and Teacher-Judges.	65
 CHAPTER 5: CONCLUSION.	 100
Summary of the Research Findings.	100
Variability of tasks.	100
Variability of teacher-judges.	102
Implications.	104

The need for context (task)-specific assessment.	104
Suitability of the NNS teacher-judges.	106
Usefulness of the Many-faceted Rasch Measurement.	107
Legitimacy of mixed methods research.	108
Limitations of the Study and Suggestions for Further research.	109
References.	113
Appendices.	129
Appendix A: Student Questionnaire.	129
Appendix B: NS Teacher Questionnaire.	133
Appendix C: NNS Teacher Questionnaire.	136
Appendix D: CATOE (Computer-Assisted Test of Oral English)	139
Appendix E: CATOE Rating Scale.	149
Appendix F: Certificate of Ethical Acceptability.	150
Appendix G: Frequently Asked Questions.	152
Appendix H: Coding Protocol of Comments.	154
Appendix I: Tables of CATOE Scale Category Statistics and Figures of CATOE Scale Structures for Tasks 2-8.	158
Appendix J: Tables of CATOE Scale Category Statistics and Figures of CATOE Scale Structures for Situation-Based and Topic-Based Tasks.	162
Appendix K: Tables of CATOE Scale Category Statistics and Figures of CATOE Scale Structures for Tasks 2-8 by NS and NNS Groups.	163
Appendix L: Tables of CATOE Scale Category Statistics and Figures of CATOE	

Scale Structures for Situation-Based and Topic-Based Tasks by NS and NNS Groups.	170
Appendix M: Tables of Number and Percentage of Comments for Tasks 2-8. .	172
Appendix N: Tables of Number and Percentage of Comments for Situation- Based and Topic-Based Tasks.	175
Appendix O: Tables of Number and Percentage of Comments for Tasks 2-8 by NS and NNS Groups.	176
Appendix P: Tables of Number and Percentage of Comments for Situation- Based and Topic-Based Tasks by NS and NNS Groups.	180

List of Tables

Table 1. Coding Scheme of Teacher-Judges' Comments.	52
Table 2. Student Proficiency Measurement Report.	60
Table 3. Teacher-Judge Severity Measurement Report.	62
Table 4. Task Difficulty Measurement Report.	64
Table 5. Distribution of Teacher-Judges According to Consistency.	67
Table 6. Rater Effect Criteria.	69
Table 7. Rater Effect for NS and NNS Teacher-Judges.	69
Table 8. Mean Severity Measures for Teacher-Judge Groups.	70
Table 9. Summary Statistics for Teacher-Judge Groups.	71
Table 10. Task Difficulty Measures.	73
Table 11. Difficulty Measures of Task Type.	73
Table 12. Task Difficulty Measures by NS and the NNS Groups.	75
Table 13. Bias Analysis Report: Interactions between Teacher-Judges and Tasks.	78
Table 14. Difficulty Measures of Task Type by NS and NNS Groups.	79
Table 15. Bias Analysis Report: Interactions between Teacher-Judges and Task Types.	80
Table 16. CATOE Scale Category Statistics for Task 1.	83
Table 17. CATOE Scale Category Statistics for Picture-Based Task.	84
Table 18. CATOE Scale Category Statistics for Overall Tasks by NS and NNS Groups.	85
Table 19. CATOE Scale Category Statistics for Task 1 by NS and NNS Groups.	87
Table 20. CATOE Scale Category Statistics for Picture-Based Task by NS and NNS Groups.	89
Table 21. Number and Percentage of Comments for Overall Tasks.	90

Table 22. Number and Percentage of Comments for Task 1	92
Table 23. Number and Percentage of Comments for Picture-Based Task.	94
Table 24. Number and Percentage of Overall Comments by NS and NNS Groups. . . .	95
Table 25. Number and Percentage of Comments for Task 1 by NS and NNS Groups. .	97
Table 26. Number and Percentage of Comments for Picture-Based Task by NS and NNS Groups.	99

List of Figures

Figure 1. FACETS Variable Map.	59
Figure 2. Task Difficulty Measures by NS and NNS Groups.	76
Figure 3. Bias Analysis between Teacher-Judge Groups and Tasks.	77
Figure 4. Bias Analysis between Teacher-Judges and Tasks.	78
Figure 5. Bias Analysis between Teacher-Judge Groups and Task Types.	80
Figure 6. Bias Analysis between Teacher-Judges and Task Types.	80
Figure 7. CATOE Scale Structure for Task 1.	83
Figure 8. CATOE Scale Structure for Picture-Based Task.	84
Figure 9. CATOE Scale Structure of NS Group for Overall Tasks.	85
Figure 10. CATOE Scale Structure of NNS Group for Overall Tasks.	85
Figure 11. CATOE Scale Structure of NS Group for Task 1.	87
Figure 12. CATOE Scale Structure of NNS Group for Task 1.	87
Figure 13. CATOE Scale Structure of NS Group for Picture-Based Task.	89
Figure 14. CATOE Scale Structure of NNS Group for Picture-Based Task.	89

CHAPTER 1

INTRODUCTION

Research Rationale

The past several decades have seen an increase in language testing literature, from both a theoretical and practical perspective. At the heart of this flourishing growth has been the advent and evolution of communicative language testing, particularly since the 1980s. Much work has been done to develop a test that is able to assess communicative competence in the real world, and this concerted effort has enriched the theoretical and practical grounds for performance assessment.

Concurrently with this increasing interest in performance assessment, language testing researchers have devoted considerable attention to the idea that performance assessment is inexorably interlinked with the potential variability, which may jeopardize the reliability, validity, and fairness of the assessment. In second language performance assessment, task and rater variability have long been recognized as the major factors that threaten the reliability and the validity of the construct being measured, and that consequently prevent an accurate inference about test-takers' language abilities.

Many previous studies on task variability have been pertinent to task type and task difficulty, and it has been argued that task attributes exert an impact on test performance or estimation of constructs (Bachman, 1990; Henning, 1983; Shohamy, 1983; Shohamy, Reves, & Bejarano, 1986). Although studies have

repeatedly reported that task attributes have a systematic effect on the test scores or constructs being measured, the nature and significance of these effects reveal gaps in our knowledge (Fulcher, 2003). In particular, further studies are needed to address how a rating scale is calibrated for a task, or what latent task factors might contribute to that calibration, from a psychometric perspective. The lack of substantive understanding about the complexity and variability of tasks suggests new areas for research and motivates more rigorous research.

As is the case with task variability, rater variability is a potential source of measurement error. Rater-involved assessment has long drawn the attention of language testing researchers, who have raised the concern that it inevitably engages subjective judgments, thus making complete rater consensus close to impossible. A number of studies have explored differences in rater behavior, and these have found that raters tend to differ according to their backgrounds and prior experience (Barnwell, 1989; Brown, 1995; Chalhoub-Deville, 1995a, 1995b; Fayer & Krasinski, 1987; Galloway, 1980; Hadden, 1991; Hill, 1997; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994, 1998). However, the outcomes of these studies have often been contradictory, possibly because they utilized different native languages, a small sample of raters, or different methodologies (Brown, 1995; Chalhoub-Deville, 1995a).

As a continuation of the ongoing discussion about task and rater variability, this study intends to comprehensively examine how second language oral performance is assessed by different groups of teacher-judges across different tasks and task types. The substantive focus of the study investigates how native English-speaking (NS) and non-native English-speaking (NNS) teacher-judges

influence task difficulty and the calibration of rating scales across different tasks and task types, and whether they exhibit internal consistency and severity. It also explores the evaluation criteria or construct elements that are salient to the two different groups of teacher-judges across different tasks and task types. Although there has previously been some analysis of differences between native (NS) and non-native English-speaking (NNS) teachers' judgments of students' oral English performance (e.g., Brown, 1995; Fayer & Krasinski, 1987; Galloway, 1980), there has been little attempt to examine how the two groups influence the calibration of rating scales, or what evaluation criteria or construct elements they draw on to infer language ability when no evaluation criteria are available for them to consult. This study addresses the extent to which task and rater variability impact second language performance assessments, using the Many-faceted Rasch Measurement and grounded theory analysis.

Chapter Overview

Chapter 1 presents an introduction to the thesis. Chapter 2 deals with general discussion of second language performance assessment and specific empirical studies on variability of performance assessment, in three major sections: 1) overview of second language performance assessment, 2) theoretical models of second language performance assessment, and 3) systematic variations of second language performance assessment. Chapter 3 outlines the methodology used in this research. Chapter 4 reports the findings of the study. Chapter 5 addresses the conclusions of the study, citing implications, limitations, and suggestions for further research.

CHAPTER 2

REVIEW OF THE LITERATURE

In this chapter, the theoretical and empirical discussions that inform previous research in second language performance assessment will be addressed in order to develop a better-rounded perspective. The chapter covers topics from general second language performance assessment to specific empirical studies on the variability of second language performance assessment, and consists of three major sections: 1) overview of second language performance assessment, 2) theoretical models of second language performance assessment, and 3) systematic variations of second language performance assessment.

Overview of Second Language Performance Assessment

While the origin of second language performance assessment is not language testing per se, it has always interested second language practitioners and researchers, and has therefore seen consistent advancement over the last four decades (McNamara, 1997). Even in the 1950s, the era of psychometric-structuralists (Spolsky, 1975, 1977, 1981, 1995), when discrete point tests were a dominating trend, a practical call for *authentic* language testing emerged in occupational training and personnel selection (McNamara, 1996). Second language performance assessment has gradually developed since the 1970s, as have underlying theories of communicative competence (McNamara, 1996, 1997).

Interest in second language performance assessment has recently increased, since it promises not only authentic language use within a test context, but also a beneficial washback on classroom teaching and learning.

Performance tests, also known as *authentic* or *direct* tests, are defined as tests “in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed” (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999, p. 144). Haertel (1992; as cited in McNamara, 1997) offers two different definitions of performance measurement. One is defined more narrowly as “the sampling and quantification of some behavior that would occur whether it were being assessed or not,” and the other, defined more broadly, as “any tests in which the stimuli presented or the responses elicited emulate some aspects of nontest settings” (p. 984). According to McNamara (1996), a defining characteristic of performance assessment is that “actual performances of relevant tasks are required of candidates, rather than more abstract demonstration of knowledge, often by means of pencil-and-paper tests” (p. 6).

Slater (1980; as cited in McNamara, 1996) and Jones (1985; as cited in McNamara, 1996)¹ classify performance tests into three types: direct assessment, work sample methods, and simulation techniques. They suggest that a maximum fidelity can be attained through direct observation of performance, because a direct assessment does not manipulate the performance tasks. In work sample

¹ Jones (1985) extends Slater’s (1980) study of performance tests in occupational assessment to second language settings.

methods, on the other hand, performance tasks are controlled to create more reliable and standardized tests. Simulation techniques are distinguished from both of the other test types in that they take certain aspects of reality and typify them as performance task sets. In particular, Jones (1985) emphasizes simulation techniques in second language performance assessment because they are able to strike a balance between test validity and overall accuracy.

Similarly, McNamara (1996) draws a distinction between a *strong* versus a *weak* sense of second language performance tests in terms of evaluation criteria. According to the strong view, a real-life task is replicated as a test task, and the test performance is evaluated using a real-life standard. Thus, language proficiency itself is not crucial in assessing performance, because completion of the task is the primary interest of assessors. Referring to Messick's (1994) distinction between performances and the products of performance assessment, McNamara (1996) goes on to note that "Such a test thus involves a second language as the *medium* of the performance; performance of the *task* itself is the *target* of assessment" (p. 43). On the other hand, the weak view of second language performance assessment attaches more weight to language performance than to fulfillment of the task. Since language proficiency is one of the major components of performance assessment evaluation criteria, most second language performance tests take the weak stance.

Theoretical Models of Second Language Performance Assessment

As McNamara notes (1996), much of the work on second language performance assessment has been done as part of the communicative or authentic language tests of the 1980s. In an attempt to define the components of communicative language ability, researchers have examined what constitutes language competence and performance. In his 1972 theoretical paper “On communicative competence,” Hymes presented an impressive discussion on linguistic competence and linguistic performance. He begins by pointing out that Chomsky’s generative grammar (1965) is too limited to explore the concept of language use, and that a substantial awareness of sociocultural factors is necessary in order to identify the area involved in the underlying competence for use. He then argues that Chomsky’s notion of performance must be clarified, and proposes two different possible interpretations: the first is that actual performance is distinguished from underlying competence (weak version of distinction), and the second is that underlying models or rules of performance is distinguished from underlying grammatical competence (strong version of distinction). As Hymes notes, the weak version of distinction between competence and performance is well understood, whereas the strong version of distinction is not. Hymes also points out that Chomsky’s strong version of interpretation, that competence is interlinked with grammatical competence only, is inappropriate in that contextual or sociolinguistic competence is not taken into account (Hymes, 1972; as cited in Canale & Swain, 1980).

Hymes (1972) then takes a rather different perspective along with this criticism. For him, competence is the general capability of a person, which relies upon tacit knowledge and ability. He proposes that every individual has different ability to use knowledge, and that its specification is inexorably interlinked with such non-cognitive factors as motivation. On the other hand, performance implies “the interaction between competence (knowledge and ability for use), the competence of others, and the cybernetic and emergent properties of events themselves” rather than behavioral evidence, and thus denotes “actual use and actual events” (Hymes, 1972, p. 283).

McNamara (1996) points out that the distinctions suggested by Chomsky (1965) and Hymes (1972) are not clear-cut, and there seems to be a gray area between them. According to him, knowledge of language, rather than underlying competence, was Chomsky’s main interest, while Hymes’ communicative competence includes characteristics of both knowledge and performance. In other words, Hymes’ model of communicative competence takes into account both the sociolinguistic aspects of language knowledge and the psychological aspects of language performance (McNamara, 1996). As linguists will attest, Hymes’ work on communicative competence was and remains influential in the field, and is still credited for its integration of communicative competence and social context.

While Hymes’ (1972) communicative competence was derived from a concern for first language (McNamara, 1996), a seminal work on communicative competence in second language was completed by Canale and Swain in 1980. Contrary to Hymes, Canale and Swain (1980) argue that ability for use cannot be

integrated into the definition of communicative competence because no theory of human behavior can properly explain its definition and application. They also suggest that the inclusion of ability for use presumes that language users have linguistic deficits.

Along with this criticism they propose three primary communicative competences, all of which are involved in language knowledge only: grammatical competence, sociolinguistic competence, and strategic competence. Grammatical competence is the knowledge of lexical features, morph-syntax, semantics and phonology used by second language learners to produce grammatically accurate sentences. Sociolinguistic competence is the rules that are used to interpret speech in a given social context, including sociocultural rules of use² and rules of discourse³. The final communicative competence, strategic competence refers to the ability to compensate for communication failures, and can be either verbal or non-verbal. According to Canale and Swain, even if little is known about these communication strategies, learning how to exploit them can benefit early second language learners. They also assert that their model is pertinent to second language teaching and testing, and can thus equip second language learners with the grammatical rules of the second language through both sociolinguistic and strategic competence. They acknowledge that their theory is based on existing work (i.e., Allen & Widdowson, 1975; Halliday, 1970; Hymes, 1967, 1968; Johnson, 1977; Morrow, 1977; Stern, 1978; Widdowson, 1978; & Wilkins, 1976),

² Sociocultural rules of use are rules by which statements are generated and delivered appropriately within a given context.

³ Rules of discourse are understood in terms of cohesion and coherence.

and is not in fact original.

Three years later, Canale (1983) presents a slightly revised model that introduces a new feature of language knowledge, discourse competence. According to Canale, discourse competence, which was part of sociolinguistic competence in the earlier model, is a way of integrating grammatical forms and meanings, and unified text can be attained through cohesion in form and coherence in meaning. Cohesion is defined as a way of constructing controlled utterances using such devices as pronouns, synonyms, ellipsis, conjunctions and parallel, whereas coherence indicates the relationships that exist among various textual meanings.

In this later paper, Canale (1983) makes two comments on his theoretical framework. The first is that he regards communicative competence as divided rather than universal. The second is that he does not consider how his four competencies interact with one another. With regard to this, Shohamy (1988) also claims that more rigorous research be conducted on the interaction of the four elements of communicative competence.

Drawing on Canale and Swain's earlier model (1980), Bachman (1990) presents a new theoretical framework for communicative language ability, which he defines as "both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, conceptualized communicative language use" (p. 84). It focuses on both competence and performance and consists of language competence, strategic competence, and psychophysiological mechanisms. Unlike Canale and Swain, Bachman includes both language

knowledge and general ability of language use in his model. According to him, language competence⁴ comprises organizational competence and pragmatic competence, each of which subsumes its subordinated features. Grammatical competence and textual competence belong to organizational competence, and refer to the abilities required to create and identify grammatically correct sentences, to figure out their prepositional content, and to arrange them to structure texts. Pragmatic competence, on the other hand, is composed of illocutionary competence, which requires pragmatic practices to carry out acceptable language performance, and sociolinguistic competence, which involves appropriate language use in a given context.

Bachman (1990) suggests that Canale and Swain's (1980) view of strategic competence is incomplete in that they do not include the mechanisms by which strategic competence works. Bachman himself regards strategic competence as one component of communicative language ability, not as a component of language competence. In other words, to Bachman, strategic competence is ability for use, not knowledge (McNamara, 1996). In his model, strategic competence is composed of three components: assessment, planning, and execution. He also includes psychophysiological mechanisms, which engage in the channel and mode of language use by which competence is realized.

⁴ Bachman notes that the construction of language knowledge is based on empirical studies that attempt to confirm the components of Canale and Swain's model (e.g., Bachman & Palmer, 1982). Studies that attempted to examine the validity of Canale and Swain framework include Allen, Cummins, Mougeon, and Swain (1983), Harley, Allen, Cummins, and Swain (1990) and Swain (1985).

More recently, Bachman and Palmer (1996) have proposed a revision of Bachman's earlier model (1990). They propose that the components of language use consist of language ability, topical knowledge and affective schemata, and they interact with each other and with other features of language use setting. According to them, language ability includes language knowledge and strategic competence. Language knowledge refers to the same concept as language competence in Bachman's earlier work; only the term illocutionary competence has been replaced with functional knowledge. Strategic competence is defined as a series of metacognitive strategies that allow for cognitive control in language use and other cognitive performance: goal setting, assessment, and planning. Topical knowledge indicates knowledge schemata or real-world knowledge that assists language use. More importantly, affective schemata mean emotional association with topical knowledge, and can either facilitate or restrict flexibility of language use. McNamara (1996) regards the inclusion of affective factors in language use as a notable development, in that Hymes' idea of ability for use is explicitly represented.

Systematic Variations of Second Language Performance Assessment

One important factor that distinguishes performance assessment from traditional assessment is scoring procedure. Contrary to traditional fixed response assessment,⁵ performance assessment involves several construct relevant or

⁵ According to McNamara (1996), a traditional fixed response assessment elicits scores from the instrument only, with no interaction from other variables. This type of assessment usually takes the form of a true/false or multiple-choice test, in which candidates' responses are limited by the instrument itself and scored based on a pre-determined answer key. This traditional assessment

irrelevant variables, which add the issues of complexity and variability. According to Upshur and Turner's (1999) refined model of performance assessment, test-takers are asked to produce a spoken or written performance instead of simply marking an answer among choices. The performance is then judged by a human rater with the aid of a rating scale or scoring guide. The involvement of a human rater introduces a new dimension of interaction: test-takers must interact with the task in order to generate discourse, and the discourse, the rating scale, the raters and the task itself must interact with each other in order to produce the final scores, by which the inference about test-takers' ability is possible.⁶

Rating scales.

As has been stated, rating scales can have a systematic effect on performance assessment scores. In order to judge the performances of test takers, raters must use a rating scale as a yardstick, but the final scores may be affected by the inherent variables of the scale. A rating scale is usually expressed in numerical values or descriptive statements, and conveys how well the individual being tested has performed a certain task. In order for such scores to be meaningful, each scale should relate to both the language constructs to be measured and the purposes of the test within a specific context (Alderson, 1991).

A rating scale can be classified in a variety of ways. Alderson (1991) divides such scales into three categories according to their purpose: a user-

procedure has proven restrictive in recent language assessment, however, making an advanced assessment scheme necessary.

⁶ McNamara (1997) notes that ratings are remarkably influenced by the interaction between raters and rating scales, no matter what is the quality of the performance is.

oriented scale, an assessor-oriented scale, and a constructor-oriented scale. A user-oriented scale informs those being tested about the meaning of the ratings, while an assessor-oriented scale is created to facilitate rating practices. In the same way, a constructor-oriented scale assists the creators of the test. Luoma (2004) also classifies rating scales, as rater-oriented, examinee-oriented, and administrator-oriented. A rater-oriented scale helps raters to make their decisions, while an examinee-oriented scale provides performance information relating to test-takers' strengths and weaknesses. An administrator-oriented scale provides the most detailed rating guidelines. In a slightly different view, Brindley (1998) distinguishes between behavior-based and theory-derived rating scales: a behavior-based scale describes features of language use within a specific context, whereas a theory-derived scale describes language ability as it relates to a specific situation.

One of the conventional distinctions of a rating scale is its scoring method, which further classifies the scale as holistic or analytic. A holistic rating scale, also known as a global or impressionistic rating scale, assumes that language ability is a single unitary ability, and therefore assigns a single score to test performance (Bachman & Palmer, 1996). In other words, a rater will simultaneously note various traits in an examinee's performance, and will assign a single score to reflect his or her general impression of it. Typical holistic rating scales are the American Council of the Teaching of Foreign Languages (ACTFL) speaking scale (ACTFL, 1999), the Test of Spoken English (TSE) rating scale (ETS, 2001) and the Interagency Language Roundtable (ILR) rating scale (ILR,

1991), all of which are variations of the Foreign Service Institute (FSI) rating scale (Clark & Clifford, 1988).

The perceived advantages of the holistic rating scale are speed and high reliability (Cooper, 1977; Davies et al., 1999; Luoma, 2004; & White, 1985). For the past several decades, holistic rating scales have been held up as a means of economical and practical scoring, but they have also been heavily criticized. A major weakness of holistic rating is its lack of diagnostic information beyond relative rank ordering (Charney, 1984; Davies et al., 1999; Hamp-Lyons, 1991; Luoma, 2004; & White, 1985). As Hamp-Lyons (1995) notes, “A holistic scoring system is a closed system, offering no windows through which teachers can look in and no access points through which researchers can enter” (p. 760-761).

Bachman and Palmer (1996) have suggested that holistic scales promote inference, since it is unclear what a single score says about language knowledge and language use in a specific situation. They further claim that holistic scales are problematic when raters have difficulty determining the level at which the test-taker’s performance should be matched. When all the criteria of a holistic scale are not met concurrently, which is often the case, a rater must (whether consciously or unconsciously) prioritize some criteria over others (Bachman & Palmer, 1996). Holistic rating scales also generally denote successful performances by using quantifiers (e.g., some, many, a few, few) and quality indicators (e.g., satisfactorily, effectively, well), so that one level of a scale cannot be interpreted without dependence on the adjacent levels (Luoma, 2004; North, 1996; & Underhill, 1987).

An analytic scale, on the other hand, assumes that an examinee's performance score can be expressed as the sum of the separate scores awarded to each criterion. The ratings assigned to each component of language ability provide examinees with detailed information about relative strengths and weaknesses in their performance. However, a common criticism of the analytic scale is that rating each task feature distracts raters from test-takers' overall performance (Davies et al., 1999). Further flaws lie in the fact that the criteria chosen for analytic scoring can be arbitrary, lacking in consistency, or can even overlap with other criteria (Matthews, 1990). More seriously, there is not a great deal of theoretical underpinning that suggests that language ability can be explained by "the accumulation of a series of subskills" (White, 1985, p.123).

In response to such criticisms (e.g., Chalhoub-Deville, 1997; Fulcher, 1987, 1988; Lantolf & Frawley, 1985, 1988; Pienemann, Johnston & Brindley, 1988; Pollitt & Murray, 1996; Shohamy, 1990), Turner and Upshur (1996) and Upshur and Turner (1995) introduced a groundbreaking approach to the development of rating scales. Known as empirically-derived, binary-choice, boundary-definition (EBB) scales, they are constructed, using performance samples in a particular task, by asking raters to make a sequence of yes/no choices about characteristics of test performance that distinguish boundaries between score levels (Upshur & Turner, 1995). In other words, an EBB scale is composed of a set of hierarchical binary questions about small samples of the particular task being rated. EBB scales are different from traditional scales in that they depict the boundaries between categories instead of illustrating the midpoint of a band

(Upshur & Turner, 1995).

Upshur and Turner (1995) argue that the simplicity and clarity with which EBB scales distinguish boundaries eliminates the problems inherent in scales with co-occurring characteristics, thus minimizing the chance that raters will have different interpretations to scale descriptors and enhancing rater reliability. The floor or ceiling effect is also reduced, since raters do not make assumptions about the development of ability and features in a given performance, and use empirical data as a starting point for scale development (Upshur & Turner, 1995).

While EBB scales certainly have advantages over other scales in that they are simple and easy to use within a specific test context and provide pedagogical information about student progress to both teachers and students, the inability to generalize across contexts has been recognized as their main weakness (Fulcher, 2003; Shohamy, 1996). Indeed, Brindley (1998) suggests that EBB scales should be complemented by further research containing more theoretical grounding in task generalization and text complexity (as cited in Turner & Upshur, 2002).⁷ Despite these criticisms, there is no doubt that the clarity and practicality of EBB scales can prove valuable in certain contexts.

In a similar vein, North (1995, 1996, 1997; North & Schneider, 1998) introduces a new approach to the development of scaling descriptors. Strong criticism of intuitively-developed scales led to the construction of scaling descriptors from large pools of different descriptors. By consulting with teachers

⁷ However, Chalhoub-Deville (1997) argues that context-specific assessment models, such as those of Hinofotis, Bailey, and Stern (1981) and Chalhoub-Deville (1995a, 1995b), reflect subcomponents of universal theoretical construct (as cited in Turner & Upshur, 2002).

in workshops and by administering questionnaires, these descriptors could be advanced into valid, stand-alone criteria. While Fulcher acknowledges the use of the Rasch analysis, through which descriptors are scaled (Fulcher, 2003), the absence of a theoretical model was held up as a shortcoming of this approach by North and Schneider (1998) and Fulcher (2003).

Fulcher (1987, 1988, 1993, 1996b, 1997) proposes a *data-based* or *data-driven* fluency rating scale, citing the necessity that scales be empirically-based. He argues that observed test performance should be quantifiable, and that the development procedures of rating scales should reflect real linguistic performance. Unlike an a priori method of rating scale development,⁸ Fulcher's data-based fluency scale sets out a large database of speech samples, which are then used to collect fluency rating descriptors (Fulcher, 1993, 1996b, 2003). According to him, the use of discourse analysis during development procedures makes such descriptors much more detailed, thereby distinguishing itself from other traditional rating scales.

Tasks.

Although it has been widely agreed that tasks have a systematic effect, both on test scores and on estimating the abilities of test-takers, researchers have not reached a consensus about its nature and significance, citing the need for more rigorous study (Fulcher, 1997). While studies of task types in second language

⁸ An a priori method means developing rating scales based on experts' (e.g., experienced teachers, language testers, or language testing specialists in examination board) intuitive judgments concerning the development of language proficiency, a teaching syllabus, or a needs analysis (Fulcher, 1996b, 2003).

acquisition are associated with task difficulty, studies of task types in second language testing appear to be associated with a more profound discussion of the dimensionality of a test and the generalization of its score. In light of this, two different views have been elucidated: that different tasks assess differences in language ability, and that general language ability can be assessed regardless of task types.

For example, in a study that investigated multiple variables in an oral performance test, Shohamy (1983) examined whether differences in speech styles and topics had a significant effect on scores on an oral Hebrew proficiency test. When the tasks were presented in an interview format, test scores were considerably higher than when they were presented on different topics in a reporting format. As Shohamy notes, even though speech styles and topics certainly influenced the scores, it was not clear whether it was speech styles, topics, or the interaction between the two that resulted in significant scoring differences.

In a similar vein, Shohamy, Reves, and Bejarano (1986) experimented with four different types of tests in a project to develop a new oral English proficiency test. The four tests were in oral interview, role-play, reporting task, and group discussion formats, and were crafted to represent the different speech styles that generally emerge in oral communicative situations. It was found that the mutual variance among the tests was relatively low, which they took as evidence of the necessity of employing multiple tests with a view to assessing general oral proficiency.

A more evolved view can be found in Chalhoub-Deville's study (1995a), which compares oral interview, narration, and read-aloud task types. Three different dimensions were identified that underlay test-takers' scores across tasks: grammar-pronunciation, creativity in presenting information, and amount of detail provided. Chalhoub-Deville noted that language constructs can be presented differently according to given tasks, and suggested that empirically-developed, context-specific rating scales be employed.

Contrary to the previous view, Fulcher (1993, 1996a) found that while tasks significantly affect test scores, the effect was not critical enough to reduce the generalization of scores across different tasks. Using G-theory and the Many-faceted Rasch Measurement, Fulcher compared the test score variances of three different tasks: a picture description, an interview, and a group discussion. The results showed that the variance generated by test-takers was much greater than the variance generated by raters or tasks, suggesting that task effect is restricted unless a rating scale delineates task-specific performance in its descriptors.

Similar results can be found in Bachman, Lynch, and Mason's study (1995). When the sizes of the variances associated with test-takers, raters, and tasks were examined, the variance attributed to test-takers was the largest, but the variance attributed to raters and test-takers was very small. In addition, a relatively small interaction effect between rater and task was found, indicating that differences in rater behavior across different tasks are small.

Raters.

McNamara (1996) discusses different cases in which raters differ from each other. Firstly, there is a basic difference in their overall severity. In some cases, raters exhibit different patterns of severity or leniency towards particular candidates or tasks; for example, when a rater interacts with a particular item, he or she may be consistently lenient on assessing fluency, but consistently severe on assessing accuracy in a speaking test. Severity or lenience may also be tied to candidates who have particularly high or low language ability.

In other cases, as McNamara explains, each rater may interpret the rating scale differently. If a candidate's performance falls approximately at the junction 2 or 3 of discrete rating categories, for example, rater A might give the candidate a rating of 2, while rater B might give the same candidate a rating of 3. Thus, the equal intervals of the rating scale may be differently interpreted by rater A and B. Raters might also differ in the range of scores they use: while some might assign the occasional extreme scores to candidates, others might be more likely to assign middle scores across the board. Finally, raters may not even be self-consistent, exhibiting inconsistency from one performance setting to another and leading to random error.

Much research has been done in rater variability, and earlier work tended to focus on how raters differed according to their background (Barnwell, 1989; Brown, 1995; Chalhoub-Deville, 1995a; Fayer & Krasinski, 1987; Galloway, 1980; Hadden, 1991; Hill, 1997; Shohamy, Gordon, & Kraemer, 1992). In a study of raters' linguistic backgrounds, for example, Fayer and Krasinski (1987)

examined how the English-speaking performance of Puerto Rican students was perceived by native English-speaking raters and native Spanish-speaking raters. A one and a half to two minute-long spoken English performance of seven Puerto Rican ESL students was assessed by 40 native English speakers and 88 native Spanish speakers. In their analysis of such linguistic factors as grammar, pronunciation, intonation, lexical and discourse errors, and of such non-linguistic factors as distraction and irritation, Fayer and Krasinski found that the Spanish raters tended to be more severe in general and to express more annoyance when rating linguistic forms, and that pronunciation and hesitation were the most distracting factors for both sets of raters.

In a study that investigated rater variability in terms of professional background, Hadden (1991) compared the perceptions of teachers and non-teachers rating Chinese students' competence in spoken English. Eight Chinese ESL students were asked to discuss a given topic for a maximum of three and a half minutes, and their performance was videotaped. Two rater groups, consisting of 25 English teachers and 32 non-teachers, all of whom were native speakers of American English, were asked to make ratings on student performance using a 24-item questionnaire. Through a factor analysis and a MANOVA, it was found that as far as linguistic ability is concerned, teachers tend to be more severe than non-teachers. However, there were no significant differences in other factors such as comprehensibility, social acceptability, personality, and body language.

In a similar vein, Chalhoub-Deville (1995a) compared three different rater groups – native Arabic-speaking teachers living in the U.S., non-teaching native

Arabic speakers living in the U.S., and non-teaching native Arabic speakers living in Lebanon – to see how their perceptions might be reflected in their oral assessments on three tasks: a modified oral proficiency, narrating a story from pictures, and reading a text aloud. Three dimensions were found to emerge from the rating criteria: the first was grammar and pronunciation, the second was creativity and adequacy of information, and the third was the amount of detailed information provided. By employing multidimensional scaling (MDS) and individual differences scaling (INDSCAL), Chalhoub-Deville found that native Arabic teachers in the U.S. tended to put more emphasis on the second dimension; non-teaching Arabic native speakers in the U.S. tended to focus on all three dimensions; and non-teaching Arabic native speakers in Lebanon tended to rely on the first dimension. The results of this study are not comparable to Hadden's (1991) in that the teachers in this study weighed more on the creativity and adequacy of information in the narration than on linguistic features. Chalhoub-Deville notes that this may be because her study was conducted using modern standard Arabic (MSA), whereas Hadden's study was conducted in English.

Combining two different rater features, Galloway (1980) investigated how raters with different linguistic and professional backgrounds perceived non-native speakers' communicative competence differently. Four different groups consisting of eight raters each were involved in this study: non-native Spanish teachers, native Spanish teachers, non-teaching native Spanish speakers living in the target language country, and non-teaching native Spanish speakers not living in the target language country. Ten students were asked to speak in Spanish for a

maximum of three and a half minutes, and their videotaped performance was rated by each of the four rater groups. Galloway found that non-native teachers tended to focus on grammatical forms, and reacted more negatively to non-verbal behavior and slow speech, while non-teaching native speakers seemed to put more emphasis on content and build up an instant rapport with the students who endeavored to express themselves despite experiencing difficulty.

Brown (1995) also investigated how raters' linguistic and work-related backgrounds affected their assessment of test-takers' performance on an industry-specific Japanese spoken-language test. The performances of 51 examinees, some of whom had an industry background, were assessed by 33 native or non-native Japanese-speaking raters with backgrounds as either teachers or travel guides. The results demonstrated that native speakers and raters with industry backgrounds tended to be more severe than non-native speakers and raters with teaching backgrounds, but the difference was not significant. However, raters with teaching backgrounds tended to be more severe in such areas as grammar, vocabulary and fluency, while raters with industry backgrounds tended to be more severe on pronunciation. Raters with teaching backgrounds were also less willing to give excessively low marks to less competent test-takers than were raters with industry backgrounds. She concludes that that these differences seemed to persist in spite of rater training and explicit evaluation criteria, and consequently "there is little evidence that native speakers are more suitable than non-native speakers ... However, the way in which they perceive the items (assessment criteria) and the way in which they apply the scale *do* differ" (Brown, 1995, p. 13).

Along with the refinement of approaches to addressing rater variability, researchers' interest began turning to rater training. With regard to this, Barnwell (1989) investigated whether a difference exists between ACTFL-trained raters and native speakers who have not been given rater training. Four American students of Spanish were interviewed, and their oral proficiency was rated by both ACTFL-trained raters and untrained native raters. Untrained native speakers were found to be more severe than ACTFL-trained raters, even though they pursued similar patterns in comparing each test-taker. This result conflicts with that of Galloway (1980), in which naïve native speakers are more lenient than teachers. Barnwell suggests that both studies are small in terms of their research scope, and that it is therefore premature to make conclusions about native speakers' responses to non-native speaking performance. Furthermore, Hill (1997) points out that the use of two different versions of rating scales, one of which is presented in English and the other is in Spanish, remains questionable.

Shohamy, Gordon, and Kraemer (1992) also examined reliability in the assessment of written essays by raters with different backgrounds of teaching experience and training. Fifty writing samples were rated by four groups of five raters: English teachers who received rater training, English teachers who did not receive rater training, native English speakers who received rater training, and native English speakers who did not receive rater training. The researchers found that there was no difference in inter-rater reliability between raters who had teaching experience and raters who did not, while trained raters showed higher inter-rater reliability than untrained raters.

More recently, in a study investigating the suitability of non-native speakers as raters, Hill (1997) found that native English-speaking raters tended to be significantly more severe in rating student writings when the reference standard was established as a non-native English speaker for the specific purposes opposed to an ideal native speaker. Hill suggests that there is no solid rationale for believing that non-native speakers are less suitable to rate an English test for specific purposes than native speakers.

In summary, studies of rater variability report that in general, teachers and non-native speakers tend to be more severe than non-teachers and native speakers. However, the outcomes of different studies do contradict one another in some cases; this may be because the studies used different native languages, a small sample of raters, and different methodologies (Brown, 1995; Chalhoub-Deville, 1995a).

This chapter has reviewed the theoretical and empirical discussions that form the background for the study. It has outlined general accounts of second language performance assessment and their underlying conceptual frameworks. Studies pertinent to the systematic variation of second language performance assessment have also been addressed, as have their associated empirical studies. The next chapter will address the research questions the study will attempt to answer. It will also present the research design of the study, along with a full description of instrument development and the data analysis procedure.

CHAPTER 3

RESEARCH QUESTIONS

AND

METHODOLOGY

Research Purpose and Questions

While performance assessment has broadened and enriched the practice of language testing, there have been ongoing questions as to whether issues of complexity and variability in performance assessment might influence the usefulness of a test. That testing tools and human factors must be involved in test-taking and rating procedures is inevitable, but these factors have long been recognized as potential sources of variance that is irrelevant to a test's construct.

This study continues the ongoing discussion about task and rater variability by comprehensively examining how second language oral performance is assessed by different groups of teacher-judges across different tasks and task types. The substantive focus of the study investigates how native English-speaking (NS)⁹ and non-native English-speaking (NNS) teacher-judges influence task difficulty and the calibration of rating scales across different tasks and task types, and whether they exhibit internal consistency and severity. It also explores the evaluation criteria or construct elements¹⁰ that are salient to the two different

⁹ In the language literature, NS and NNS are widely used as abbreviations for Native Speakers and Non-Native Speakers, respectively.

¹⁰ The terms *evaluation criteria* and *construct elements* are equivalent in this study. For purposes

groups of teacher-judges across different tasks and task types. Using the Many-faceted Rasch Measurement and grounded theory analysis, this study seeks to answer the following research questions:

- 1) Does the behavior of NS and NNS teacher-judges differ in terms of internal consistency and severity?
 - 1-1) Do some teacher-judges rate student performance inconsistently?
 - 1-2) Is one group of teacher-judges more severe or lenient?
 - 1-3) Is one group of teacher-judges more homogeneously severe than the other?

- 2) How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks and task types?
 - 2-1) How are task difficulty measures influenced across different tasks and task types?
 - 2-2) How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks?
 - 2-3) How are task difficulty measures influenced by NS and NNS teacher-judges across different task types?

- 3) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks and task types?
 - 3-1) How is the calibration of rating scales influenced across different

of expedience, the two are jointly described as *construct elements*.

tasks and task types?

3-2) How is the calibration of rating scales influenced by NS and NNS teacher-judges as a whole?

3-3) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks?

3-4) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different task types?

4) What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks and task types?

4-1) What are the salient construct elements across different tasks and task types?

4-2) What are the salient construct elements drawn on by NS and NNS teacher-judges as a whole?

4-3) What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks?

4-4) What are the salient construct elements drawn on by NS and NNS teacher-judges across different task types?

Participants

Students.

Korean students enrolled in ESL courses in a college-level language institute were chosen as the population for two reasons. First, because this study

includes how NNS teachers, who are native Korean speakers, judge their students' oral English performance in an EFL context, it was necessary to include Korean-speaking students as participants. Second, because of the developmental stage of the students' English proficiency and the difficulty of the Computer-Assisted Test of Oral English (CATOE), an oral English test used in the study (see *Instruments* section in this Chapter), college students were considered more appropriate than secondary school students.¹¹

Ten Korean students made up the test-taker sample. A stratified random sample procedure was used so that the student sample would represent the whole population in terms of language proficiency. In order to ensure this, students were selected using the same yardstick (class level), and all of them were recruited from one college-level language institute in Montreal. This English Language Program places students into one of five different levels according to their placement test results. In other words, students with low English proficiency are placed in Level I, while students with high English proficiency are placed in Level V. Although the intention was to include two students from each level, this turned out to be impossible, because so few students were enrolled in Levels I and II. One student from Level I participated in the study; one student from Level II, three students from Level III, three students from Level IV, and two students from Level V.

Background information about the students was obtained via a questionnaire after the oral English test was administered (see Appendix A and

¹¹ The proficiency threshold for the CATOE was set as low-intermediate for adult learners.

Instruments section in this Chapter). Six of the students were male and four were female, and they ranged in age from early 20 to early 30. Six students had already completed an undergraduate degree, while four were in the progress of obtaining an undergraduate degree. The academic majors of the students varied widely, with the following departments represented: one from Arms Control, two from Business, one from Korean Language and Literature, one from Economics, one from Sociology, one from Biology, one from Law, one from Chemistry, and one from Psychology.

All of the students had been educated in Korea until they came to Canada. Their study of English in Korea ranged from six to nine years, with their English classes focusing primarily on reading and grammar. Their total time in Canada at the time of this study ranged from 1 to 24 months.

Four students reported that they were studying English for business purposes, three for academic purposes, and three in order to improve their personal communication skills. In terms of English language self-assessment, two students described themselves as beginners, five as intermediate, two as upper-intermediate, and one as advanced. Few students (three out of ten) had had prior experience of taking an oral English exam.

Native English-speaking teacher-judges.

The native English-speaking teacher-judges were Canadian teachers of English in a college-level language institute in Montreal, Canada. In order to ensure that the NS teachers were sufficiently qualified as teacher-judges to rate

the students' oral English performance, certain criteria were followed; 1) they were native English speakers; 2) they had at least one year of prior experience teaching an English conversation course to non-native English speakers in a college-level language institute; and 3) they had at least one graduate degree in a field related to linguistics or language education.

Twelve Canadian English teachers were selected for the NS teacher-judge sample. Their background information was obtained via a questionnaire after their student evaluations were completed (see Appendix B and *Instruments* section in this Chapter). Of the 12 NS teacher-judges, four were male and eight were female, and they ranged in age from 30 to 50. Ten NS teacher-judges were native English speakers, and two were perfect bilinguals in German and Romanian.¹² In terms of educational background, 11 of the NS teacher-judges had a Master's degree in Linguistics or Language Education, and one had a Master's degree in Psychology. All of the NS teacher-judges had experience teaching English language courses (ESL conversation, ESL Academic writing, EAP [English for Academic Purposes], ESP [English for Specific Purposes], Business English, etc). The number of years they had taught such courses varied widely; one had taught English for less than 3 years, two for 3 to 6 years, six for 7 to 10 years, and three for more than 11 years. All of the teacher-judges reported that they were very familiar with the spoken English of non-native English speakers.

¹² In the multilingual context of Montreal where the study was conducted, there are many English speakers whose first language is not English. In this study, although two NS teacher-judges' first language was not English, they reported that they speak English most of the time, and that their English is in fact much better than their native language.

With regard to their ability to evaluate spoken English, nine NS teacher-judges had taken courses specifically in Second Language Testing and Evaluation, while three had not. Of the nine who had taken such courses, four had been trained as raters of spoken English. All of the NS teacher-judges were familiar with rating the spoken English of non-native English speakers; six said they were familiar with rating the spoken English of non-native English speakers “to some extent,” two “a lot,” and four “very familiar.” All NS teacher-judges reported that they used anecdotal notes, checklists, marks and scores, in addition to rating scales, as evaluation tools in their daily teaching practices.

Non-native English-speaking teacher-judges.

The non-native English-speaking teacher-judges were Korean teachers of English in a college-level language institute in Daegu, Korea. In order to ensure that the NNS teachers were sufficiently qualified as teacher-judges to rate their students’ oral performance, teachers who were selected for the study met the following criteria: 1) they were native Korean speakers with high proficiency in spoken English; 2) they had at least one year prior experience teaching an English conversation course to non-native English speakers in a college-level language institute; and 3) they had at least one graduate degree in a field related to linguistics or language education.

Twelve Korean English teachers made up the NNS teacher-judge sample. Their background information was obtained via a questionnaire after the evaluations of their students’ tests had been completed (see Appendix C and

Instruments section in this Chapter). Of the 12 NNS teacher-judges in the study, two were male and 10 were female, and they ranged in age from 20 to 40. All were native speakers of Korean, and had been educated in Korea until they had graduated from secondary school. In terms of their educational background, 10 NNS teacher-judges had earned a Master's degree in Linguistics or Language Education from an English-speaking country (e.g., Australia, the U.K., or the U.S). One had earned a Bachelor's degree in English Language and Literature from Korea, and the other had earned a Master's degree in English Translation from Korea. All of the NNS teacher-judges had lived in English-speaking countries for one to seven years for academic purposes, and reported their English proficiency levels as advanced (six teacher-judges) or near-native (six teacher-judges).

All of the NNS teacher-judges had experience teaching English language courses (English Grammar, English Reading and Listening Comprehension, TOEFL [Test of English as a Foreign Language], TOEIC [Test of English for International Communication], ESL Academic writing, ESL conversation, Business English, etc). The number of years they had taught such courses varied; one NNS teacher-judge had taught English for less than 3 years, eight for 3 to 6 years, and three for 7 to 10 years. All reported that they were very familiar with the spoken English of non-native English speakers.

With regard to their ability to evaluate spoken English, eight NNS teacher-judges had taken courses specifically in Second Language Testing and Evaluation, while four had not taken such courses. Of the eight who had, one had been trained as a rater of spoken English. In terms of their familiarity with rating the spoken

English of non-native English speakers, one teacher-judge reported that he or she was “a little” familiar with rating the spoken English of non-native English speakers, eight reported that they were familiar with such ratings “to some extent,” two reported that they were familiar “a lot,” and one reported that he or she was “very familiar.” All of the NNS teacher-judges reported that they used anecdotal notes, checklists, marks and scores, in addition to rating scales, as evaluation tools in their daily teaching practices.

To summarize, both NS and NNS teacher-judges shared common educational and professional backgrounds, and differed only in terms of their first languages – either Korean or English.

Instruments

The Computer-Assisted Test of Oral English (CATOE).

An oral English test, called the Computer-Assisted Test of Oral English (CATOE), was developed specifically for the study (See Appendix D for the final version of the CATOE). The purpose of the CATOE was to assess the overall oral communicative language ability of non-native English speakers within an academic context. Throughout the test, communicative language ability is evidenced by the effective use of language knowledge and strategic competence (Bachman & Palmer, 1996).

As has been reported, different task types elicit differences in oral language output, systematically affecting test scores (Chalhoub-Deville, 1995a, 1995b; Henning, 1983; Shohamy, 1983, Shohamy, Reves, & Bejerano, 1986;

Upshur & Turner, 1999). In order to assess the diverse oral language output of test-takers, the test was designed to consist of three different task types: picture-based, situation-based, and topic-based. The picture-based task asks test-takers to describe or narrate visual information, such as describing the layout of a library (T1)¹³, sharing information with someone else about library use (T2), telling a story from pictures (T4), and describing a graph of human life expectancy (T7). The situation-based task requires test-takers to perform the appropriate pragmatic function in a hypothetical situation, such as congratulating a friend on being admitted to school (T3). Finally, the topic-based task asks test-takers to offer their opinions on a given topic, such as explaining their personal preferences for either individual or group work (T5), discussing the harmful effects of the Internet use (T6), and suggesting reasons for an increase in human life expectancy (T8). Before these eight questions are given, two warm-up questions are presented to give the test-takers the opportunity to practice. These warm-up questions are not scored.

The test is to be administered in a computer-mediated indirect interview format. The indirect method was selected for this study because the intervention of interlocutors in a direct speaking test can affect reliability (Stansfield, 1991; Stansfield & Kenyon, 1992a, 1992b). Although the lexical density produced in direct speaking tests and indirect speaking tests have been found to be different (O'Loughlin, 1995), it has consistently been reported that scores from indirect speaking tests have a high correlation with those from direct speaking tests (Clark

¹³ Hereafter, T1, T2, etc denote Task 1, Task 2, etc.

& Swinton, 1980a, 1980b; Clifford, 1978; Lowe & Clifford, 1980; O'Loughlin, 1995; Stansfield, Kenyon, Paiva, Doyle, Ulsh, & Antonia, 1990; Shohamy, Shmueli, & Gordon, 1991).

The test questions will be presented in English using audio prompts. This method is preferable for two reasons: because communicative competence includes not only the ability to use a language appropriately but also the ability to understand a message that has been delivered, listening skills should also be assessed. In addition, if instructions are given in the test-takers' mother tongues, it is more likely that they will unnaturally translate their responses from the mother tongue into the target language (Luoma, 2004).

The test lasts approximately 25 minutes, of which 8 minutes are allotted for responses. The length of the response time for each task varies depending on task difficulty and the amount of information to be delivered. To ensure that test-takers are ready to respond to each question, 20 to 60 seconds of preparation time is provided before they must begin their answers. A timer, showing the number of elapsed seconds, is presented at the bottom right side of the computer screen. As soon as test-takers finish one task, a "next" button appears on the computer screen, asking whether they are ready to perform the next task. Test-takers can proceed to the next task by clicking the button. This was done to make the test more user-friendly. In order to effectively and economically facilitate an understanding of the task without providing test-takers with a lot of vocabulary (Underhill, 1987), each task is accompanied by visual stimuli.

Before the CATOE was developed, a formal needs analysis could not be conducted due to practical constraints, but the test was constructed based on the personal and professional experience of the researcher who was herself a non-native English-speaking test-taker and an English teacher in Korea. In developing the CATOE, the guiding principles of the Oral Proficiency Interview ([OPI], Weinstein, 1979) and the Simulated Oral Proficiency Interview ([SOPI], Malone, 2000) were referenced; more specifically, the Test of Spoken English ([TSE], ETS, 2001) and the Multimedia Assisted Test of English ([MATE], MATE, 2000) were consulted as references. The initial test development began with the identification of target language use domain, target language tasks, and task characteristics (Bachman & Palmer, 1996). The test tasks were selected and revised to reflect potential test-takers' language proficiency and topical knowledge, as well as task difficulty and interest.

Before the CATOE was finalized, four rounds of drafts, tryouts, and revisions took place. To ensure that each task was well written and functioned as intended, item format analysis (recommended by Brown [1996] as one of the procedures of developing norm-referenced tests) was carried out. Each embryonic test draft was tried out by potential test-takers covering a wide range of language abilities in each revision. These early trials, plus feedback from a second language testing expert, helped to ensure that the tasks were clear and worked as intended. The speed of the audio prompts and the preparation and response times were fine-tuned, as well. Finally, the test-development procedures ensured that the test had content validity.

The CATOE rating scale.

After development of the CATOE was completed, a rating scale was constructed (See Appendix E for the final version of the CATOE rating scale). The CATOE rating scale scores responses holistically focusing on the successfulness of communication.

The CATOE rating scale has four levels, labeled 1, 2, 3 and 4. It does not clarify any rating criteria except successfulness of communication. In other words, the band descriptors are intended not to provide teacher-judges with any information about language features or construct to draw on. Because this study investigates how teacher-judges rate oral communication ability and define the construct to be measured, no specific evaluation criteria are given. Raters are asked to assign an individual holistic score from one of the four levels to each task.

In developing the CATOE rating scale, the researcher referred to the American Council for the Teaching of Foreign Languages (ACTFL) Speaking Scale (ACTFL, 1999), the Test of Spoken English (TSE) Scale (ETS, 2001), and the Multimedia Assisted Test of English (MATE) Speaking Level Guidelines (MATE, 2000). To deal with cases in which teacher-judges “sit on the fence,” an even number of levels were sought in the rating scale. Moreover, in order not to cause a cognitive and psychological load on the teacher-judges, six levels were set as the upper limit during the initial stage of CATOE rating scale development. Throughout the trials, however, the six levels describing the degree of successfulness of communication proved to be indistinguishable without dependence on the adjacent levels. More importantly, it was unlikely that teacher-

judges would use all six levels of the rating scale in their evaluations. For these reasons, the rating scale was trimmed to four levels, which enabled the teacher-judges to distinguish consistently among them.

Before the CATOE rating scale was finalized, three rounds of drafts, tryouts, and revisions took place. To ensure that the rating scale was well articulated and worked as intended, it was tested on potential teacher-judges who were either NS or NNS teachers. These pilot trials, plus feedback from an expert in the field of second language testing, helped to ensure that the rating scale was fine-tuned and that it had content validity.

Background Questionnaires.

Background questionnaires were developed in order to elicit information about the students and teacher-judges taking part in the study (see Appendices A, B, & C). Three different questionnaires were developed: one for students, a second for NS teachers, and a third for NNS teachers. The student questionnaire requested age, gender, education background, duration of English studies, etc. It was written in Korean because participating students were drawn from all class levels, including Level I. The NS and NNS teacher questionnaires, on the other hand, were written in English and consisted of two main parts: background information and evaluation skills of spoken English. The background section asked teachers for their age, gender, first language, and educational background, and the evaluation section asked about previous rater training experience and use of evaluation tools and rating scales. The NS and NNS teacher questionnaires

were identical, except for two questions. The NS questionnaire contained questions asking teachers about their previous teaching experience in foreign countries and their degree of familiarity with the spoken English of non-native English speakers, while the NNS questionnaire contained questions about teachers' previous experiences studying in English-speaking countries, and their English proficiency levels.

Procedure

The CATOE administration.

All appropriate ethical procedures for data collection were followed (see Appendix F for a copy of the Ethical Certificate from the Faculty of Education at McGill University). Student participants were recruited through research announcements that were posted on the bulletin boards at the language institute. Those who agreed to participate in the study were informed about the research project both orally and in writing, and signed informed consent forms.

The CATOE was administered individually to each of 10 Korean students. The test was run by the Macromedia Flash Player 7, and the responses were simultaneously recorded to a digital sound file via the Sound Forge Audio Studio 7.0. By using the professional sound recording software, high-quality audio reproduction was obtained. In order to familiarize students with the test format and to show them how to interact with the computers using a microphone and headset, a practice session was held before the test day. The test was administered in a quiet language laboratory to control environmental issues, one of many

extraneous variables that can dramatically affect the validity and reliability of a test (Bachman & Palmer, 1996; Brown, 1996). Although taking an oral language test on a computer in a language laboratory is not an authentic communicative situation, it was an acceptable tradeoff for greater control of the test administration. Upon completion of the CATOE, student participants were asked to fill out the background information questionnaire.

The CATOE scoring procedure.

The 10 students' responses were distributed to both NS and NNS teacher-judges. In order to allow the teacher-judges to work more efficiently, the students' responses were stored as individual digital sound files on a CD, itemized by task. This meant that there was no need for the teacher-judges to rewind or forward audio tapes while they listened to the students' responses, saving time and minimizing fatigue – another factor that affects scoring reliability. In order to minimize a potential ordering effect, the order of the 10 students' test response sets was randomized. Of possible test response sets, 12 were passed out to both groups of teacher-judges.

A meeting was held with each teacher-judge in order to explain the research project and to go over the scoring procedure. The scoring procedure had two phases: 1) rating the students' test responses according to the CATOE rating scale and 2) justifying those ratings by providing comments about them in writing. The rationale for requiring teacher-judge comments was that they would supply not only the evaluation criteria that they draw on to infer student language ability,

but that this criteria would also help to identify the construct being measured. Before the actual ratings were carried out, teacher-judges were asked to familiarize themselves with the CATOE and the CATOE rating scale. They were then asked to listen to sample responses representing the various levels of oral English that students might exhibit during the test. Since the purpose of these sample responses was purely to make the teacher-judges more familiar with the levels of English proficiency they might encounter during their evaluations, they were not scored. Upon completion of the familiarization process, teacher-judges were allowed to listen to test responses and rate them according to the CATOE rating scale. After rating a single task response by one student, they were asked to justify that rating in writing. They then moved on to the next task response of that student. Thus, each teacher-judge was asked to score and make comments upon 80 speech samples.¹⁴

To decrease the subject expectancy effect, information about the students was not provided, regardless of teacher-judges' requests. One group of teacher-judges was not aware of the existence of the other group of teacher-judges. In an attempt to minimize the researcher expectancy effect, I, the researcher, took every effort to ensure that my knowledge of the Korean context would not influence the Korean teacher-judges' perceptions.

After meeting with the teacher-judges, a supplementary document containing frequently asked questions was distributed (see Appendix G). Meetings with the NS teacher-judges were held in Montreal, Canada and meetings with the

¹⁴ Each student responded to eight tasks, and ten students participated in the study.

NNS teacher-judges followed in Daegu, Korea. Each meeting lasted approximately 30 minutes.

Data Analysis

Two different approaches were taken in analyzing the data. A Many-faceted Rasch Measurement was used to analyze the CATOE scores and calibrations of CATOE rating scale, while grounded theory was used to analyze the teacher-judges' written comments on the students' oral English performance. Each of these approaches is explained in more detail in the following sections.

Many-faceted Rasch Measurement analysis.

The Many-faceted Rasch Measurement is one of the most promising recent measurement tools for controlling and analyzing complex performance assessment schemes. This model enables the analysis of measurement error inherent in a performance-based test, and adjusts examinees' scores for systematic variations of each facet (e.g., item difficulty, rater severity, occasion stringency, etc).

The most striking part of the Rasch theory is that the model is *probabilistic*, or *stochastic* (Rasch, 1980). Unlike classical test theory (or true score theory), which *determines* the ability of an examinee in a particular test condition, the Rasch theory *estimates* the latent ability of the examinee while taking the entity of such test conditions into consideration (Linacre, 1989). In other words, the measure of examinees' latent ability is freed from the severity of

a particular rater, as well as from the difficulty of the item and from the arbitrary nature of the rating scale categories (Linacre, 1989). By virtue of the very nature of probability, then, it is possible to equate scores that have been obtained from different sets of items intended to measure the same trait, and to make general inferences about them (Linacre, 1989; Smith, 2004a).

The rating probability for a certain item from a particular rater for a particular examinee can be predicted mathematically from given facets, such as the ability of the examinee, the difficulty of the item, and the severity of the rater. All facets are placed simultaneously on a single common logit scale (Perline, Wright, & Wainer, 1979; Rasch, 1980), with the measurement units expressed as logits. The logit units have an advantage over raw scores in that they lie on a linear interval scale and enable mathematic operations within and across facets (Smith, 2004b).

The Rasch model requires the following data specifications: 1) parameter separation, 2) unidimensionality, and 3) local independence. The usefulness of the data as measurement can be evaluated by analyzing the fit of the data (Wright, 1991). Parameter separation means that each parameter of the test condition (i.e., examinees, items, raters, rating scales, etc.) should be independent of other parameters. That is, the estimated ability of examinees is freed from the distribution of the item parameters, and the estimated difficulty of the items is freed from the distribution of the examinee parameters (Smith, 2004a). Unidimensionality means that the items should measure a single ability or trait (Hambleton & Cook, 1977), enabling the addition of scores across different items

and different subsets of the test (Smith, 2004c).¹⁵ Multidimensionality causes problems when the data present two or more distinct dimensions and fail to identify which dimension the model measures (Smith, 2004c). It should be noted that if unidimensionality is not satisfied, combining scores from the different items or subsets of the test will be meaningless. Finally, the principle of local independence states that there should be independence among the residual differences between the observed data and the expected data (Linacre, 1997). For example, each item should make an independent contribution to the measurement, providing new information that the other items do not provide (McNamara, 1996).

The form of the data to be analyzed determines the model of Rasch families. The earliest and simplest is the Basic Rasch Model, developed by Rasch (1980), which deals with such dichotomous data as responses on true/false tests, multiple-choice questions, or short answer questions without partial credit. Polytomous data without judges or other facets, such as responses on short answer questions with partial credit, or a Likert or a semantic differential scale are analyzed using the Rating Scale Model (Andrich, 1978a, 1978b), an extended form of the Basic Rasch Model, which assumes the same step difficulty across all items. Where it is necessary to examine the step difficulty of each item, the Partial Credit Model (Wright & Masters, 1982) is used. The most recent and advanced form of the Rasch family is the Many-faceted Rasch Measurement, which enables

¹⁵ Lumsden (1976; as cited in Baker, 1997) indicates “a confusion between unidimensionality and theoretical singularity” (p. 267), and argues that a test holds unidimensionality even if it is compounded with different theoretical constructs. A similar view is taken by McNamara (1996), who argues that unidimensionality should be interpreted in two different ways. In a psychological sense, unidimensionality indicates a single underlying construct. In a psychometric sense, it means a single underlying measurement dimension.

researchers to extend the model to include as many facets as needed (Linacre, 1989). In this study, the Many-faceted Rasch Measurement is implemented using the FACETS computer program (Linacre, 2005).

The output of the Rasch model provides three major statistics for each facet element: 1) a measure, 2) a standard error, and 3) fit statistics. A measure is a logit estimate of each element, with the measure average conveniently set at zero on the logit scale. Values greater than zero indicate more able examinees, more severe raters, or more difficult items than average. A standard error is related to a measure of interest, its size depending on the sufficiency of a data matrix. For example, if examinees are rated by many raters on many items, the associated standard error will be small.

The degree to which the data fit the model is expressed as fit statistics. Elements which show greater or less variation than the model expects are flagged as misfit or overfit, respectively. There are no straightforward rules for interpreting fit statistics or for setting upper and lower limits. As Myford & Wolfe (2004a) note, it is more or less context related, and depends on the targeted use of the test results. In the case of high-stakes tests, tight quality control limits (such as mean squares of 0.8 to 1.2) would be set; however, if the stakes are low, looser limits would be allowed.¹⁶ Wright and Linacre (1994, as cited in Myford & Wolfe, 2004a) propose the mean square values of 0.6 to 1.4 as reasonable values for data in which a rating scale is involved, with the caveat that the ranges are likely to

¹⁶ It must be noted that the mean square values of 0.8 to 1.2 are based on “well-behaved data from multiple-choice tests” (Linacre & Williams, 1998, p.653).

vary depending on the particulars of the test situation. A more evolved view of the interpretation of fit statistics can be found in Linacre:

From the measurement perspective, the crucial aspect is not ‘significance’ but ‘distortion.’ My work since the 1994 reference suggests that mean squares less than 1.5 are productive of measurement. Between 1.5 and 2.0 are not productive but not deleterious. Above 2.0 are distorting. Even though 2.0, if only produced by a few unexpected observations, the distortion is so local as to have no overall impact (as cited in Myford & Wolfe, 2004a, p. 508).

In this study, the data were analyzed using the Many-faceted Rasch Measurement and the FACETS computer program, Version 3.57.0. Five facets were specified: student, teacher-judge, teacher-judge group, task, and task group. The teacher-judge group and task group facets were entered as dummy facets and anchored at zero.¹⁷ A hybrid Many-faceted Rasch Measurement Model (Myford & Wolfe, 2004a)¹⁸ was used to differentially apply the Rating Scale Model to teacher-judges and tasks, and the Partial Credit Model to teacher-judge groups and task groups. The model equation of the analysis is as follows:

¹⁷ Setting up a dummy facet with anchoring all elements at zero does not change main analysis (J. M. Linacre, personal communication, May 08, 2005).

¹⁸ As the name implies, a “hybrid” Many-faceted Rasch Measurement Model combines the Rating Scale Model with the Partial Credit Model.

$$\log\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_{ijk}^{19}$$

P_{nijk} = the probability of examinee n being awarded a rating of k when rated by judge j on item i

$P_{nij(k-1)}$ = the probability of examinee n being awarded a rating of $k-1$ when rated by judge j on item i

B_n = the ability of examinee n

D_i = the difficulty of item i

C_j = the severity of judge j

F_{ijk} = the difficulty of achieving a score within a particular score category (k) modeled separately for each item and judge

In addition to the primary analysis described above, additional analyses were conducted with numbers of facets that varied according to the specific research questions. For example, in order to investigate the overall differences between the two groups of teacher-judges regardless of task types, another hybrid Many-faceted Rasch Measurement Model was used: the Rating Scale Model was applied to teacher-judges and tasks, and the Partial Credit Model was applied to teacher-judge groups. When an analysis focuses on the variability of teacher-judges, the student and task facets were centered by anchoring logit measure means at zero, while the teacher-judges were allowed to float.

¹⁹ If all the elements of a facet are anchored at zero, then it does not enter into the estimation equation (J. M. Linacre, personal communication, May 08, 2005). Thus, the dummy facets of teacher-judge group and task group were excluded from the equation.

FACETS analyzed 1,727 valid ratings, awarded by 24 teacher-judges to 72 sample responses by 10 students on eight tasks.²⁰ Every teacher-judge rated every student's performance on every task, so that the data matrix was fully crossed. The upper and lower quality control limits were set at 0.5 and 1.5, respectively (Lunz & Stahl, 1990), given the test's rating scale and the fact that it investigates the perceptions of teacher-judges in a classroom setting rather than those of trained raters in a high-stakes test setting.

Grounded theory analysis.

According to Strauss and Corbin (1998), grounded theory is defined as “theory that was derived from data, systematically gathered and analyzed through the research process” (p. 12). In other words, rather than being predetermined, theory actually emerges from the data. Through a process called *theoretical sampling*, the core concepts of the data are identified, developed and related, and each sample is analyzed and compared in order to determine the categories that will represent all of the collected samples with their varied properties and dimensions (Strauss & Corbin, 1998). A convincing theory is built when 1) no new categories emerge, 2) the categories vary in terms of their properties and dimensions, and 3) validity among categories is obtained.

In this study, the teachers' written comments were open-coded. Utilizing the open coding approach meant that concepts could be grouped under the

²⁰ A rating of NR (Not Ratable) was treated as missing data, and, of 80 speech samples, there were eight such cases. In addition, the teacher-judge, NS 5, failed to make one rating.

categories that emerged from the data. Ambiguous or hard to interpret comments, and comments that were rarely repeated, were excluded from the analysis (about three percent of the total).²¹ Of available comments, the NS teacher-judges made a total of 2,123 comments and the NNS teacher-judges a total of 1,172 comments on the students' oral English performance. Nineteen recurring features were identified in the teacher-judges' written comments, clustered under the five major categories shown in Table 1 (for specific examples of the coding scheme, see Appendix H).

²¹ Teacher-judges sometimes provided only evaluative adjectives, which did not offer evaluative substance (e.g., "accurate," "clear," and so on). So that the evaluative intent would not be misjudged, such comments were not included in the analyses. In addition, comments that occurred fewer than 20 times were excluded as categories (e.g., "low volume," "soft voice," "little confidence," "poor time management," and so on).

Table 1. Coding Scheme of Teacher-Judges' Comments

Major Categories & Definitions	Sub-Categories
1. General Task Fulfillment: the degree to which the response fulfills the general demands of the task	<ul style="list-style-type: none"> • Understanding the task • Overall task accomplishment
2. Content Effectiveness: the degree to which the content of the response is of good quality and effectiveness in conveying an intended message	<ul style="list-style-type: none"> • Strength/soundness of argument • Accuracy of transferred information • Topic relevance
3. Language Use: the degree to which language features of the response are of good quality and effectiveness in conveying an intended message	<ul style="list-style-type: none"> • Overall language use • Vocabulary • Pronunciation • Fluency • Intelligibility • Sentence structure • General grammar use • Specific grammar use

Table 1 (continued). Coding Scheme of Teacher-Judges' Comments

Major Categories & Definitions	Sub-Categories
4. Socio-Contextual	
Appropriateness: the degree to which the response is appropriate and relevant to the intended communicative goals of a given situation	<ul style="list-style-type: none"> • Socio-cultural appropriateness • Contextual appropriateness
5. Organizational Development: the degree to which the response is developed and organized in a coherent and effective manner	
	<ul style="list-style-type: none"> • Coherence • Supplement of details • Completeness of discourse • Elaboration of argument

Once the data were coded and analyzed by the principal researcher, the original uncoded comments of 10 teacher-judges (five NS and five NNS teacher-judges) were examined independently by a second researcher,²² and reached approximately 95 percent agreement. When areas of disagreement were revisited, it was found that the two researchers had different perspectives on two categories: strength/soundness of argument and elaboration of argument. Discussion between the two researchers revealed that one researcher had paid little attention to the

²² The second researcher was a native Korean-speaking graduate student in second language education. Because the NNS teacher-judges' comments were written in Korean, a native Korean speaker who has sufficient background knowledge in second language education was needed.

distinct features of the two categories on the grounds that strength/soundness of argument automatically holds by its elaboration, while the other researcher had considered the two categories to be distinct features, in that strength/soundness of argument indicates the quality of argument, while elaboration of argument indicates how effectively speakers connect their ideas. Since what mattered at that point was identifying the evaluation criteria or construct elements on which teacher-judges drew (rather than identifying the logical relationships among evaluation criteria or construct elements), the two categories were distinguished as two distinct features.

A chi-square test was used to investigate what construct elements are salient across different tasks and task types, and whether there is a difference in attending salient construct elements between the NS and NNS teacher-judges. The analysis was primarily based on the frequency of teacher-judges' comments in the major categories. Since this study seeks to find construct elements across different tasks and teacher-judges rather than specific language features, the analysis of major categories provides more appropriate answers than the analysis of sub-categories.

This chapter has discussed the purpose of the study and has stated the research questions guiding it. It has also described the research design and the instruments used in the study, including the way the data were analyzed. In the next chapter, the results of the study will be reported.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter begins by discussing facet calibrations, then moves on to systematic variability findings in the tasks and teacher-judges. The results are presented according to the following research questions:

- 1) Does the behavior of NS and NNS teacher-judges differ in terms of internal consistency and severity?
 - 1-1) Do some teacher-judges rate student performance inconsistently?
 - 1-2) Is one group of teacher-judges more severe or lenient?
 - 1-3) Is one group of teacher-judges more homogeneously severe than the other?

- 2) How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks and task types?
 - 2-1) How are task difficulty measures influenced across different tasks and task types?
 - 2-2) How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks?
 - 2-3) How are task difficulty measures influenced by NS and NNS teacher-judges across different task types?

- 3) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks and task types?
 - 3-1) How is the calibration of rating scales influenced across different tasks and task types?
 - 3-2) How is the calibration of rating scales influenced by NS and NNS teacher-judges as a whole?
 - 3-3) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks?
 - 3-4) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different task types?

- 4) What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks and task types?
 - 4-1) What are the salient construct elements across different tasks and task types?
 - 4-2) What are the salient construct elements drawn on by NS and NNS teacher-judges as a whole?
 - 4-3) What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks?
 - 4-4) What are the salient construct elements drawn on by NS and NNS teacher-judges across different task types?

Calibration of Students, Teacher-Judges, and Tasks

Analysis of the data revealed that the data fit the model. According to Linacre (2005), in order for the data to fit the model, about 5% of the total standard residuals can lie outside the range of -2 to +2, and about 1% can lie outside the range of -3 to +3. Of a total of 1,727 valid responses, 89 responses (about 5%) had standard residuals above +2 or below -2, and 17 responses (about 1%) had standard residuals above +3 or below -3. The chi-square fit statistics for all facets also showed that the data specifications of parameter separation, unidimensionality and local independence were satisfied,²³ and thus measurement validity (Wright, 1991) was obtained (see infit mean squares in Tables 2 – 4; for more information about data specifications of the Many-faceted Rasch Measurement, refer to Chapter 3).

Before turning to the specific research questions that FACETS may answer, it is worthwhile to present a brief introduction to the FACETS variable map, because it presents analyses of all facets in one reference figure. Three facets were specified in this analysis: student, teacher-judge and task. The Rating Scale Model was used for teacher-judges, and the Partial Credit Model for tasks. Figure 1 displays all the facets graphically on a common logit scale. The first column in the map displays a logit scale, which is applied equally across the facets. The second column displays student proficiency measures for the CATOE Test. In this map, more proficient students are positioned at the top of the column, and less proficient students are positioned at the bottom (i.e., the S4 is the most proficient

²³ Fit statistics can be used as a tool to examine data specifications. If fit values are within an acceptable range (in this case, 0.5 to 1.5 [Lunz & Stahl, 1990]), the specifications are considered met. In this data set, the infit mean square values for tasks ranged from 0.61 to 1.35.

student, whereas the S7 is the least proficient). The third column displays the severity measures of teacher-judges, with more severe teacher-judges positioned at the top, and more lenient teacher-judges at the bottom (i.e., the NNS8 is the most severe, while the NS10 is the most lenient). The fourth column displays the difficulty measures of the tasks. T6 (discussing the harmful effects of Internet use) is the most difficult task, while T2 (sharing information with someone else about library use) is the easiest. Columns five through twelve display the four-point CATOE rating scale used to score student response to each of the eight tasks. They represent the most likely scale structure on each task. The map shows that each category of each rating scale is interpreted differently across different tasks. For example, a score of 3 on T5 is more difficult for students to attain than a score of 3 on T6.

STUDENT	TEACHER-JUDGE	TASK	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8
54			(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
56										
52	NS1									
51	NS2 NS3 NS4 NS5 NS6 NS7 NS8 NS9 NS10 NS11 NS12 NS13 NS14 NS15 NS16 NS17 NS18 NS19 NS20 NS21 NS22 NS23 NS24 NS25 NS26 NS27 NS28 NS29 NS30 NS31 NS32 NS33 NS34 NS35 NS36 NS37 NS38 NS39 NS40 NS41 NS42 NS43 NS44 NS45 NS46 NS47 NS48 NS49 NS50 NS51 NS52 NS53 NS54 NS55 NS56 NS57 NS58 NS59 NS60 NS61 NS62 NS63 NS64 NS65 NS66 NS67 NS68 NS69 NS70 NS71 NS72 NS73 NS74 NS75 NS76 NS77 NS78 NS79 NS80 NS81 NS82 NS83 NS84 NS85 NS86 NS87 NS88 NS89 NS90 NS91 NS92 NS93 NS94 NS95 NS96 NS97 NS98 NS99 NS100	T5 T6 T7 T8 T9 T10 T11 T12 T13 T14 T15 T16 T17 T18 T19 T20 T21 T22 T23 T24 T25 T26 T27 T28 T29 T30 T31 T32 T33 T34 T35 T36 T37 T38 T39 T40 T41 T42 T43 T44 T45 T46 T47 T48 T49 T50 T51 T52 T53 T54 T55 T56 T57 T58 T59 T60 T61 T62 T63 T64 T65 T66 T67 T68 T69 T70 T71 T72 T73 T74 T75 T76 T77 T78 T79 T80 T81 T82 T83 T84 T85 T86 T87 T88 T89 T90 T91 T92 T93 T94 T95 T96 T97 T98 T99 T100								
57			(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)

Figure 1. FACETS Variable Map

More detailed information about student proficiency measures is reported in Table 2. Students are listed in ascending order of proficiency, and summary statistics are presented at the bottom of the Table. Students are identified in the first column, and observed average scores and the model expected average scores (the fair average based on the Many-faceted Rasch Model) are presented in the second and third columns. The fourth and fifth columns show measures or estimates of student proficiency, and the associated errors. Fit statistics are found in the last column.

Table 2. Student Proficiency Measurement Report

Student	Obsvd Average	Fair-M Average	Measure (logits)	Model S.E.	Infit MnSq
S7	1.1	1.03	-4.94	0.42	0.91
S3	1.1	1.11	-3.65	0.28	1.21
S6	1.8	1.85	-0.83	0.12	1.05
S5	2.4	2.38	0.34	0.10	0.79
S1	2.5	2.42	0.42	0.11	0.98
S9	2.5	2.5	0.58	0.10	0.62
S10	2.5	2.55	0.69	0.10	1.11
S2	2.9	2.95	1.48	0.11	1.34
S8	3.1	3.11	1.83	0.11	1.09
S4	3.8	3.82	4.08	0.18	0.95
Mean	2.4	2.37	0.00	0.16	1.00
S.D.	0.8	0.82	2.48	0.10	0.20
RMSE (Model) = 0.19			Adj. S.D. = 2.48		
Separation = 12.88			Separation (not inter-rater) Reliability = 0.99		
Fixed (all same) chi-square = 1126.9			d.f. = 9		
Significance (probability) = 0.00					

The proficiency measures of the 10 students range from -4.94 logits to 4.08 logits, with a 9.02 logit spread. The associated errors for each student are very small and show little variation, except for S7 and S3. This is because S7 and S3 were not proficient enough to take the CATOE, and both students were awarded three NR (Not Ratable) ratings out of their total eight responses, resulting in missing data. As shown in Figure 1 and Table 2, a proficiency measure of about

-1.0 logits seemed to be the threshold level to take the CATOE.²⁴ None of the students were found to misfit, with a spread of infit mean squares from 0.62 to 1.34.²⁵ This indicates that the proficiency of individual students was appropriately measured, and that inferences based on the test scores can be validated (McNamara, 1996).

The separation ratio (G) is a measure of the spread of estimates relative to their precision (Linacre, 2005). The separation reliability ratio indicates how reliably the analysis separates the elements within a facet. The fixed (all the same) chi-square tests the null hypothesis that all elements of the facet share the same measure after taking measurement error into consideration. Student separation ratio (G) and reliability were 12.88 and 0.99, respectively. The fixed chi-square was significant, $\chi^2(9, N = 10) = 1126.9, p = 0.00$, rejecting the null hypothesis of fixed effect. It implies that students can be separated into distinct proficiency strata. Based on the information provided by the separation ratio (G), the student separation index (H) was calculated.²⁶ The student separation index (H) determines how many measurably different levels exist among students (Wright & Masters, 1982). When the student separation index (H) was computed using equation (1), student proficiency could be separated into 18 distinct strata.

²⁴ When the threshold proficiency measures of about -1.0 logits are interpreted as the class levels from which the students were drawn, it would be class level III – S7 is from class level I and S3 is from class level II.

²⁵ It must be noted that determining an acceptable range of infit mean squares for students is not clear-cut because scores are determined by teacher-judges' severity as well as students' proficiency (Myford & Wolfe, 2004a). In this study, however, the lower and upper quality control limits are set at 0.5 and 1.5, respectively.

²⁶ Separation ratio, index and separation reliability are calculated based on the same information, so the inferences made from these indices should be equal. However, it should be pointed out that while separation reliability suffers from a ceiling effect, the separation ratio and index do not (Myford & Wolfe, 2004a).

$$H = \frac{(4G + 1)}{3} \quad (1)$$

Detailed information about teacher-judge severity may be seen in Table 3. The severity measures of 24 teacher-judges range from -0.6 logits to 1.64 logits, with a 2.24 logit spread. The associated errors for each teacher-judge are small, with extremely little variation. The infit mean-squares range from 0.52 to 1.61 and three teacher-judges lie outside the acceptable range of infit mean squares;²⁷ however, none of the teacher-judges show the infit mean-squares below 0.5, indicating that all of them are independent judges. Rater separation ratio and reliability are 2.87 and 0.89, respectively. The fixed chi-square is significant, $\chi^2(23, N = 24) = 214.7, p = 0.00$, so that the null hypothesis of fixed effect should be rejected. Individual teacher-judges are therefore not interchangeable, and there is a significant difference among individual teacher-judges in severity.

Table 3. Teacher-Judge Severity Measurement Report

Teacher-Judge	Obsvd Average	Fair-M Average	Measure (logits)	Model S.E.	Infit MnSq
NS10	2.9	2.78	-0.60	0.20	1.51
NNS10	2.9	2.74	-0.52	0.20	1.26
NNS11	2.8	2.63	-0.29	0.19	1.09
NNS1	2.7	2.52	-0.07	0.19	0.85
NS9	2.7	2.43	0.11	0.19	1.34
NS5	2.6	2.37	0.23	0.19	1.07
NNS9	2.6	2.35	0.26	0.19	1.29

²⁷ The misfitting teacher-judges were not removed in the subsequent analyses because their infit mean squares were slightly greater than an acceptable limit, which were not deleterious.

Table 3 (continued). Teacher-Judge Severity Measurement Report

Teacher-Judge	Obsvd Average	Fair-M Average	Measure (logits)	Model S.E.	Infit MnSq
NS12	2.6	2.32	0.33	0.19	0.96
NNS7	2.6	2.32	0.33	0.19	1.54
NNS5	2.5	2.29	0.40	0.19	0.81
NS7	2.5	2.27	0.44	0.19	1.11
NS11	2.5	2.25	0.47	0.19	1.00
NS4	2.5	2.22	0.54	0.19	0.52
NNS4	2.5	2.22	0.54	0.19	0.52
NNS12	2.4	2.17	0.65	0.19	0.83
NNS2	2.4	2.13	0.72	0.19	0.69
NS3	2.4	2.08	0.83	0.19	0.77
NNS3	2.4	2.08	0.83	0.19	0.85
NS2	2.3	2.02	0.97	0.19	0.67
NS8	2.3	1.99	1.05	0.19	0.78
NS6	2.2	1.91	1.23	0.19	1.30
NNS6	2.2	1.84	1.38	0.19	1.61
NS1	2.1	1.75	1.6	0.20	0.68
NNS8	2.1	1.73	1.64	0.20	0.85
Mean	2.5	2.22	0.54	0.19	1.00
S.D.	0.2	0.27	0.58	0.00	0.31

RMSE (Model) = 0.19

Adj. S.D. = 0.55

Separation = 2.87

Separation (not inter-rater) Reliability = 0.89

Fixed (all same) chi-square = 214.7

d.f. = 23

Significance (probability) = 0.00

Table 4 shows that the difficulty measures of eight tasks range from -0.55 logits to 0.88 logits, with a 1.43 logit spread. The associated errors for each task are very small, with extremely little variation. The infit mean squares range from 0.61 to 1.35, satisfying the data specifications – that is, the CATOE does not hold psychometric multidimensionality and none of the tasks are redundant. Task separation ratio and reliability are 4.28 and 0.95, respectively. The fixed chi-square is significant, $\chi^2(7, N = 8) = 154.3, p = 0.00$, indicating the null hypothesis of fixed effect should be rejected. Thus, there is a significant difference in difficulty among the eight tasks.

Table 4. Task Difficulty Measurement Report

Task	Obsvd Average	Fair-M Average	Measure (logits)	Model S.E.	Infit MnSq
T2	2.6	2.49	-0.55	0.11	0.61
T5	2.7	2.39	-0.44	0.11	0.93
T4	2.5	2.39	-0.43	0.11	1.09
T3	3.0	2.32	-0.39	0.11	1.17
T1	2.3	2.04	0.28	0.10	0.87
T8	2.4	2.22	0.30	0.12	1.12
T7	2.4	2.07	0.34	0.11	0.87
T6	2.4	1.72	0.88	0.11	1.35
Mean	2.5	2.21	0.00	0.11	1.00
S.D.	0.2	0.23	0.49	0.00	0.21
RMSE (Model) = 0.11 Adj. S.D. = 0.48					
Separation = 4.28 Separation (not inter-rater) Reliability = 0.95					
Fixed (all same) chi-square = 154.3 d.f. = 7					
Significance (probability) = 0.00					

Systematic Variability Findings in the Tasks and Teacher-Judges

1. Does the behavior of NS and NNS teacher-judges differ in terms of internal consistency and severity?

1-1) Do some teacher-judges rate student performance inconsistently?

To answer this question, infit indices of each teacher-judge were examined. Teacher-judge fit indicates the degree to which each of teacher-judge is internally consistent in his or her ratings. Infit mean square values greater than 1.5 indicate significant misfit, or a high degree of inconsistency in the ratings. On the other hand, infit mean square values less than 0.5 indicate overfit, or a lack of variability in their scoring. For example, those who use only the middle categories of the rating scale, without utilizing all of the rating categories, are flagged as overfitting (McNamara, 1996). Likewise, teacher-judges who employ “play-it-safe” strategies usually show “flat-line” scoring patterns that create similar, or even identical ratings across tasks (Lee, 2003; Myford & Mislevy, 1995; Wolfe, Chiu, & Myford, 1999). Generally, misfitting teacher-judges are more problematic than overfitting teacher-judges (Linacre, 2005; McNamara, 1996). In Table 3, the fit statistics show that three teacher-judges, NS10, NNS6, NNS7, have misfit values. None of the teacher-judges show overfitting rating patterns.

In order to more precisely identify the teacher-judges whose rating patterns differed greatly from the model expectations, another analysis was carried out. According to Myford and Wolfe (2000), investigating the proportion that each teacher-judge is involved with the large standard residuals between observed scores and expected scores provides useful information about teacher-judge behavior. If teacher-judges are interchangeable, it could be expected that all

teacher-judges would be assigned the same proportion of large standard residuals, according to the proportion of total ratings that they make. Based on the number of large standard residuals and ratings that all teacher-judges make and each teacher-judge makes, the null proportion of large standard residuals for each teacher-judge (π) and the observed proportion of large standard residuals for each teacher-judge (P_r) can be computed using equations (2) and (3):

$$\pi = \frac{N_u}{N_t} \quad (2)$$

Where, N_u = the total number of large standard residuals

N_t = the total number of ratings

$$P_r = \frac{N_{ur}}{N_{tr}} \quad (3)$$

Where, N_{ur} = the number of large standard residuals made by teacher-judge r

N_{tr} = the number of ratings made by teacher-judge r

An inconsistent rating will occur when the observed proportion exceeds the null proportion beyond the acceptable deviation. Thus, the frequency of unexpected ratings (Z_p) can be calculated using equation (4). If a Z_p value for a teacher-judge is below +2, it indicates that the unexpected ratings that he or she made are random error. However, if the value is above +2, the teacher-judge is considered to be exercising an inconsistent rating pattern.

$$Z_p = \frac{P_r - \pi}{\sqrt{\frac{\pi - \pi^2}{N_r}}} \quad (4)$$

In this study, an unexpected observation was reported if the standardized residual is greater than +2, and this was the case in 89 of a total of 1727 responses. Table 5 shows the distribution of teacher-judges according to their rating consistency. One NS teacher-judge and two NNS teacher-judges were found to exhibit inconsistent rating patterns, a result that is similar to what was found in the fit analysis. The two NNS teacher-judges whose observed Z_p values were greater than +2 are NNS6 and NNS7, and these are the ones who were flagged as misfitting teacher-judges by their infit indices. Interestingly, the analysis of NS teacher-judges showed that it was NS9, not NS10, who had Z_p values greater than 2. This may be because NS10 produced only a small number of unexpected ratings, which did not produce large residuals. These small Z_p values indicate that while the teacher-judge gave a few ratings that were not unexpectedly higher (or lower) than the model would expect, those ratings were not highly unexpected (C. Myford, personal communication, May, 31, 2005).

Table 5. Distribution of Teacher-Judges According to Consistency

Z_p	Number of NS Teacher- judges	Percentage of NS Teacher- judges	Number of NNS Teacher- judges	Percentage of NNS Teacher- judges
$Z_p < 2$	11	92%	10	84%
$2 \leq Z_p < 4$	1	8%	2	16%
Total	12	100%	12	100%

Wolfe, Chiu, and Myford (1999, as cited in Myford & Wolfe, 2000) offer more explicit ways in which rater effects²⁸ may be identified, based on fit statistics and the proportion of unexpected ratings for each rater (Z_p). In their analysis, tight quality control indices (i.e., mean squares of 0.7 for the lower limit, and mean squares of 1.3 for the upper limit) were adopted. Table 6 shows the relationship between the rating effects that raters may exhibit, and the values of the fit and Z_p indices. According to Myford and Wolfe's (2004b) definition of rater effects, the randomness effect is "a rater's tendency to apply one or more trait scales in a manner inconsistent with the way in which the other raters apply the same scales" (p. 543), flagged by large fit indices and large proportions of unexpected ratings. If raters display smaller fit indices than expected, along with large proportions of unexpected ratings, they are exercising the halo or centrality effect. The halo effect is displayed when raters assign similar ratings on a distinctive trait, while the centrality effect emerges from the overuse of the middle categories of a rating scale (Myford & Wolfe, 2004b). The extreme effect is flagged by an acceptable range of infit indices, larger outfit indices than expected, and large proportions of unexpected ratings. Raters who assign ratings at the high or low ends of the scale will be pointed out as exercising extreme rating patterns.

As Table 7 shows, the two groups of teacher-judges share similar rating patterns: none of the NS and NNS teacher-judges exhibit halo or centrality effects, and extreme scoring patterns do not appear in either group. One teacher-judge

²⁸ As Myford and Wolfe (2004a) note, the differences among *rater effects*, *rater biases* and *rater errors* are not apparent, and the terms are commonly used interchangeably. Following the definition by Scullen, Mount, and Goff (2000, as cited in Myford & Wolfe, 2004a), rater effects are defined as "systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the rate" (p. 957).

from each group (NS9 and NNS6) shows randomness in the ratings. NNS7 was not included in the randomness group in this analysis due to a slightly small outfit mean square value of 1.29. On the whole, a slightly higher percentage of NNS teacher-judges (58%) showed accurate rating patterns when compared with NS teacher-judges (50%). This result implies that NNS teacher-judges are as accurate as NS teacher-judges in their assessments of students' second language oral performance. The claim that only NS teachers can fairly, reliably and validly assess students' performance is therefore groundless. Obviously, NNS teacher-judges with sufficient teaching experience and educational backgrounds are qualified assessors.

Table 6. Rater Effect Criteria

Rater Effect	Infit MnSq	Outfit MnSq	Z_p
Accurate	$0.7 \leq \text{infit} \leq 1.3$	$0.7 \leq \text{outfit} \leq 1.3$	$Z_p \leq 2.00$
Random	$\text{infit} > 1.3$	$\text{outfit} > 1.3$	$Z_p > 2.00$
Halo/Central	$\text{infit} < 0.7$	$\text{outfit} < 0.7$	$Z_p > 2.00$
Extreme	$0.7 \leq \text{infit} \leq 1.3$	$\text{outfit} > 1.3$	$Z_p > 2.00$

Table 7. Rater Effect for NS and NNS Teacher-Judges

Rater Effect	Number of NS Teacher- judges	Percentage of NS Teacher- judges	Number of NNS Teacher- judges	Percentage of NNS Teacher- judges
Accurate	6	50%	7	58%
Random	1	8%	1	8%
Halo/Central	0	0%	0	0%
Extreme	0	0%	0	0%
Other	5	42%	4	34%
Total	12	100%	12	100%

Note. "Other" indicates rating patterns that do not fit the above-described patterns.

1-2) Is one group of teacher-judges more severe or lenient?

In order to compare the severity measures of the two groups, the teacher-judge group facet was entered as a dummy facet into the data matrix and anchored at zero. Table 8 shows that the NS teacher-judge group has a severity measure of 0.58 logits, whereas the NNS teacher-judge group has a severity measure of 0.51 logits. Since it turned out that the NS teacher-judge group appeared to be a bit more severe than the NNS group, a *t-test* was carried out to determine whether the difference in severity between the two groups was significant. The results showed that although the NS teacher-judge group appeared more severe, the difference was not significant, $t(22) = 0.45, p = .66$.

Table 8. Mean Severity Measures for Teacher-Judge Groups

Teacher-Judge Group	Obsvd Average	Fair-M Average	Measure (logits)	Model S.E.	Infit MnSq
NS Mean	2.5	2.21	0.58	0.19	0.99
NS S.D.	0.2	0.26	0.57	0.00	0.30
NNS Mean	2.5	2.24	0.51	0.19	1.00
NNS S.D.	0.2	0.29	0.59	0.00	0.32

1-3) Is one group of teacher-judges more homogeneously severe than the other?

In order to identify the extent to which the NS and NNS teacher-judge groups are homogeneous or varied in their severity, standard deviations and separation ratios were examined from the same analysis. As Table 9 indicates, the NNS group was a bit more varied in its severity, with a standard deviation of 0.56 logits compared with 0.54 logits for the NS group. This result was confirmed by

the separation ratio of the two groups: the NS teacher-judge group showed a separation ratio of 2.77, whereas the NNS group showed a separation ratio of 2.95. This means that the NNS teacher-judge group can be distinguished by a somewhat more varied strata of severity than the NS group, but the difference between the two groups is extremely small (0.02 logits), so can be considered negligible.

Table 9. Summary Statistics for Teacher-Judge Groups

	NS Group	NNS Group
RMSE (Model)	0.19	0.19
Adj. S.D.	0.54	0.56
Separation	2.77	2.95
Separation (not inter-rater) Reliability	0.88	0.90
Significance (probability)	0.00	0.00

2. How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks and task types?

2-1) How are task difficulty measures influenced across different tasks and task types?

Table 10 shows that T2 (sharing information with someone else about library use) is the easiest task (-0.55 logits), while T6 (discussing the harmful effects of Internet use) is the most difficult (0.88 logits) across different tasks. A closer investigation into difficulty measures brings T5 and T6 to attention. Although these two tasks belong to the same task type (topic-based), their difficulty measures are positioned at the opposite extremes. Although somewhat mitigated, the same pattern was identified on T2 and T7. In order to evaluate

whether task difficulty is significantly different in each task type, all task types were entered as dummy facets and the fixed (same) chi-squares of the picture-based and topic-based tasks were examined. All of the fixed chi-squares were significant: $\chi^2(3, N = 4) = 19.0, p = 0.00$ for the picture-based task; and $\chi^2(2, N = 3) = 39.7, p = 0.00$ for the topic-based task. In other words, task difficulty was not systematically sustained within the task type, and task types failed to function as solid predictors that determined the difficulty of a certain task.

When the difficulty measures of task types were compared, the situation-based task had the lowest difficulty rating (-0.35 logits), while the topic-based task had the highest (0.35 logits). This result partially confirms the findings of Brown, Anderson, Shillcock, and Yule (1984) that static tasks (e.g. diagramming or giving instructions about laying out a pegboard) are the easiest, and abstract tasks (e.g. giving opinions or justification) are the most difficult. According to their research on second language acquisition, the task that provided all the content that a subject is supposed to transmit is easier than a task where the subject bases the information in the task on his own knowledge. However, when their criteria of difficulty were applied to individual tasks, it was found that this was not always the case. As shown in Table 10, T7 (describing a graph) and T1 (describing a location), which belong to static tasks, were ranked second and fourth in terms of difficulty, and T5 (explaining personal preferences), which belongs to abstract tasks, was the second easiest task. The failure of the task difficulty model suggested by second language acquisition research may be because the definition and operationalization of task difficulty employed in

second language acquisition is different from that of second language testing, and thus the definition and operationalization of task difficulty cannot be used interchangeably in the two cases (Iwashita, McNamara, & Elder, 2001).

Table 10. Task Difficulty Measures

Task	Difficulty Measure (logits)	S.E.
Describing a location (T1)	0.28	0.10
Sharing given information (T2)	-0.55	0.11
Congratulating (T3)	-0.39	0.11
Telling a story from pictures (T4)	-0.43	0.11
Explaining personal preference (T5)	-0.44	0.11
Discussing an issue (T6)	0.88	0.11
Describing a graph (T7)	0.34	0.11
Suggesting reasons (T8)	0.30	0.12
Mean	0.00	0.11
S.D.	0.49	0.00

Table 11. Difficulty Measures of Task Type

Task Type	Difficulty Measure (logits)	S.E.
Situation-Based Task	-0.35	0.15
Picture-Based Task	0.00	0.16
Topic-Based Task	0.35	0.16
Mean	0.00	0.16
S.D.	0.30	0.00

2-2) How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks?

Unlike the previous analysis, which examined how task difficulty measures are influenced across different tasks and task types, this analysis was carried out using a different perspective in order to identify how the two groups of teacher-judges responded across different tasks. That is, the two groups of teacher-judges were compared to see whether they consistently influenced difficulty measures across different tasks. Two approaches were employed: first, task difficulty measures derived from the two groups of teacher-judges were compared. Given that task difficulty is determined to some extent by raters' severity in a performance assessment setting, comparison of the task difficulty measures is considered to be a legitimate approach. Second, a FACETS bias analysis was performed to investigate whether the systematic sub-patterns of a particular group of teacher-judges display significantly severe or lenient rating patterns toward particular tasks. According to McNamara (1996), a bias analysis yields more extensive and accurate interaction effects than unsophisticated averages because it is produced by considering all relevant information about the facets of interest.

In order to compare the difficulty derived from the two groups across different tasks, the task group facet was entered into the data matrix as a dummy facet, with the NS teacher-judges' task ratings coded 1-8, and the NNS teacher-judges' task ratings coded 9-16. All elements in the task facets and task groups were anchored at zero. Table 12 shows the task difficulty derived from the NS and the NNS groups of teacher-judges. As can be seen, the ratings of the NS group are

a bit more diverse across tasks, with task difficulty measures ranging from -0.53 logits to 0.97 logits with a 1.50 logit spread. In the NNS group's ratings, however, while the range of task difficulty measures is a bit narrower, it is still very similar to that of the NS group: from -0.59 logits to 0.82 logits, with a 1.41 logit spread. This result is also confirmed by a standard deviation of 0.52 logits for the NS group, compared with 0.49 logits for the NNS group. Both groups exhibited generally similar patterns of task difficulty measures (see Figure 2). T6 was given the highest difficulty measure by both groups of teacher-judges, and T3 and the T2 were given the lowest difficulty measure by the NS and the NNS teacher-judge groups, respectively.

Table 12. Task Difficulty Measures by NS and the NNS Groups

Task	NS Group		NNS Group	
	Difficulty Measures (logits)	S.E.	Difficulty Measures (logits)	S.E.
Describing a location (T1)	0.35	0.14	0.28	0.14
Transferring given information (T2)	-0.51	0.15	-0.59	0.15
Congratulating (T3)	-0.53	0.16	-0.31	0.16
Telling a story from pictures (T4)	-0.32	0.16	-0.47	0.15
Telling personal preference (T5)	-0.52	0.15	-0.48	0.16
Discussing an issue (T6)	0.97	0.16	0.82	0.16
Describing a graph (T7)	0.33	0.16	0.38	0.16
Supporting an opinion (T8)	0.23	0.17	0.38	0.17
Mean	0.00	0.16	0.00	0.16
S.D.	0.52	0.01	0.49	0.01

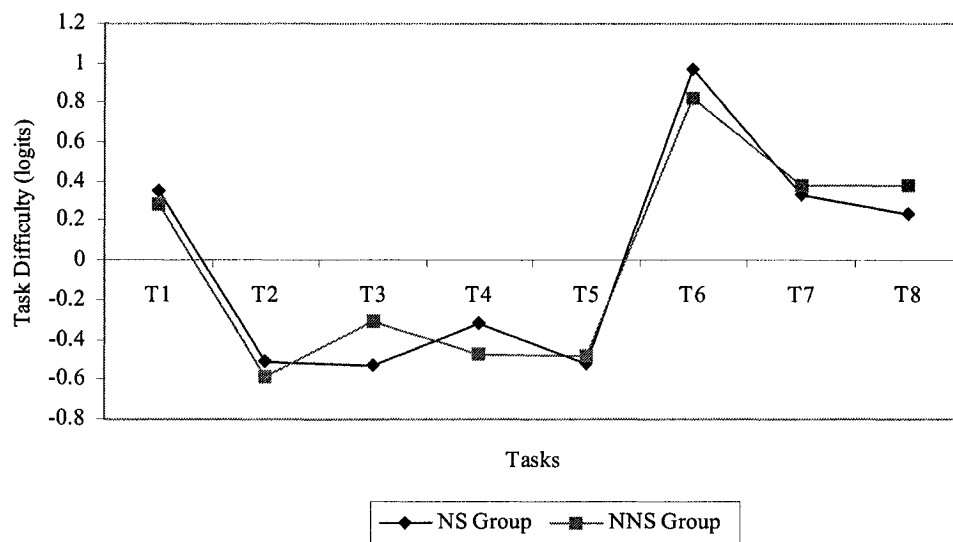


Figure 2. Task Difficulty Measures by NS and NNS Groups

A bias analysis was carried out in order to further explore the potential interaction between teacher-judge groups and tasks. In a bias analysis, an estimate of the extent to which a teacher-judge group was biased on a particular task is standardized to a Z-score. When the Z-score values in a bias analysis fall between -2.0 and +2.0, that group of teacher-judges is considered to be scoring a task without significant bias. Where the values fall below -2.0, that group of teacher-judges is scoring a task leniently compared with the way they have assessed other tasks, suggesting a significant interaction between the group and the task. By the same token, where the values are above +2.0, that group of teacher-judges is considered to be rating that task more severely than the other tasks. As the bias slopes of Figure 3 illustrate, neither of the two groups of teacher-judges was positively or negatively biased toward any particular tasks; thus, the NS and the NNS teacher-judge groups do not have any significant interactions with particular tasks.

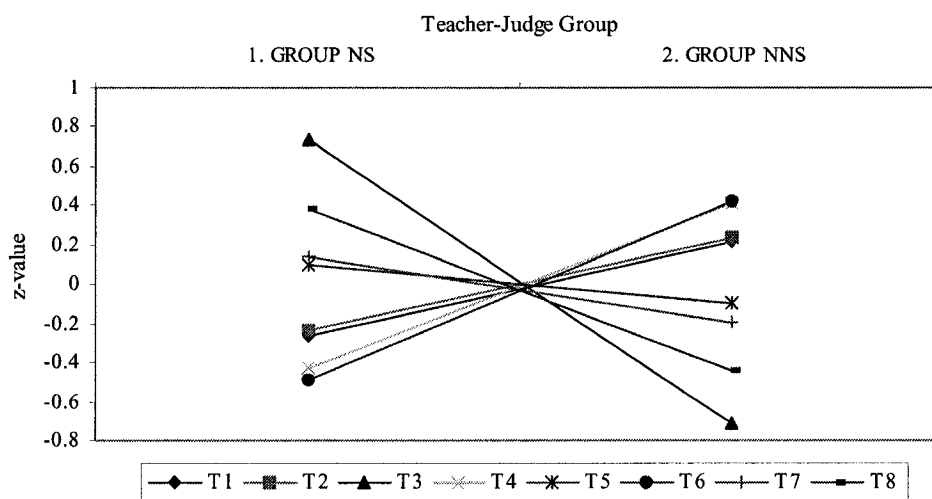
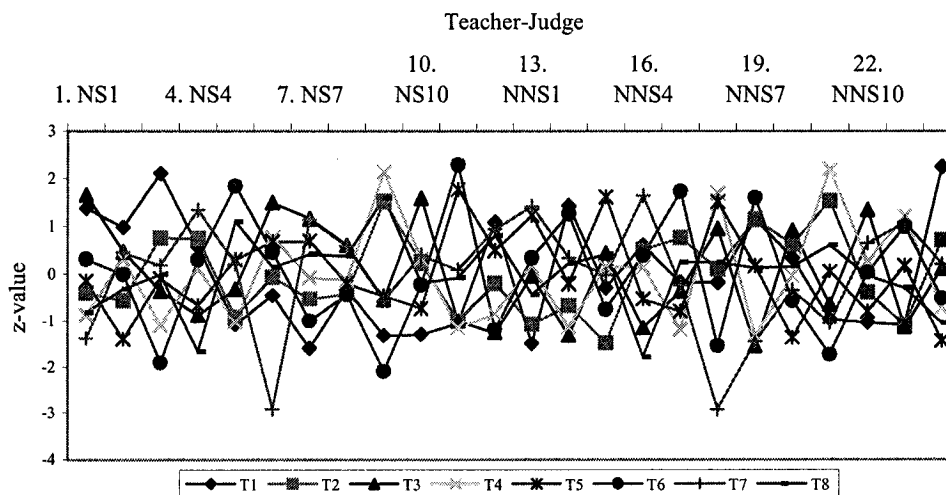


Figure 3. Bias Analysis between Teacher-Judge Groups and Tasks

A bias analysis between individual teacher-judges and tasks confirmed the picture presented in previous analysis: even though an interaction was found between individual teacher-judges and tasks, no bias toward a particular task emerge from a particular group of teacher-judges. Strikingly, certain teacher-judges from each group showed exactly the same bias patterns on particular tasks. As shown in Table 13 and Figure 4, one teacher-judge from each group had significantly lenient rating patterns on T1 and T4, and significantly severe patterns on T7. Two NS teacher-judges exhibited conflicting rating patterns on T6; NS11 showed a significantly more lenient pattern of ratings, while NS9 showed the exact reverse pattern – that is, NS9 rated T6 significantly more severely. It is very interesting that one teacher-judge from each group showed the same bias patterns on T1, T4, and T7. This implies that the two teacher-judges may be interchangeable, in that they display the same bias patterns.

Table 13. Bias Analysis Report: Interactions between Teacher-Judges and Tasks

Teacher-Judge	Task	Obs-Exp Average	Bias Measure (logits)	Model S.E.	Z-Score	Infit MnSq
NS11	T6	0.54	-1.26	0.55	-2.29	0.9
NS9	T4	0.38	-1.23	0.58	-2.13	1.5
NNS9	T4	0.43	-1.22	0.55	-2.19	1.5
NNS12	T1	0.47	-1.18	0.53	-2.24	0.7
NS3	T1	0.44	-1.06	0.50	-2.11	0.8
NS5	T6	0.43	-1.01	0.55	-1.84	1.3
NNS6	T6	-0.34	1.06	0.69	1.54	3.0
NS9	T6	-0.49	1.21	0.58	2.09	2.1
NS3	T6	-0.44	1.21	0.64	1.90	0.7
NS6	T7	-0.60	1.90	0.65	2.92	1.1
NNS6	T7	-0.60	2.02	0.69	2.93	1.1

Figure 4. Bias Analysis between Teacher-Judges and Tasks²⁹

²⁹ There is a mismatch between Table 13 and Figure 4, because the “bias direction” command is not active for Excel plots. This will be corrected in the next update (J. M. Linacre, personal communication, June 08, 2005).

2-3) How are task difficulty measures influenced by NS and NNS teacher-judges across different task types?

Since little difference was found between the NS and NNS teacher-judge groups across different tasks, it is worthwhile to further explore how the two groups respond across different task types. As noted in Chapter 3, the CATOE consists of three types of tasks: picture-based (T1, T2, T4, & T7), situation-based (T3), and topic-based (T5, T6, & T8). The analysis was carried out by entering the facet of the task types as a dummy facet and anchoring all elements at zero.

Table 14 shows that the two groups share the same rating patterns across different task types, giving the lowest difficulty measure to the situation-based task and the highest difficulty measure to the topic-based task. A bias analysis of teacher-judge groups and task types confirmed these results: neither group was biased toward a particular task type (see Figure 5). When a bias analysis was carried out on individual teacher-judges, only NS11 was indicated as significantly lenient on the topic-based task (see Table15 & Figure 6).

Table 14. Difficulty Measures of Task Type by NS and NNS Groups

Task	NS Group		NNS Group	
	Difficulty Measures (logits)	S.E.	Difficulty Measures (logits)	S.E.
Picture-Based Task	0.08	0.08	-0.09	0.08
Situation-Based Task	-0.43	0.16	-0.26	0.15
Topic-Based Task	0.35	0.09	0.36	0.09
Mean	0.00	0.11	0.00	0.11
S.D.	0.33	0.04	0.26	0.03

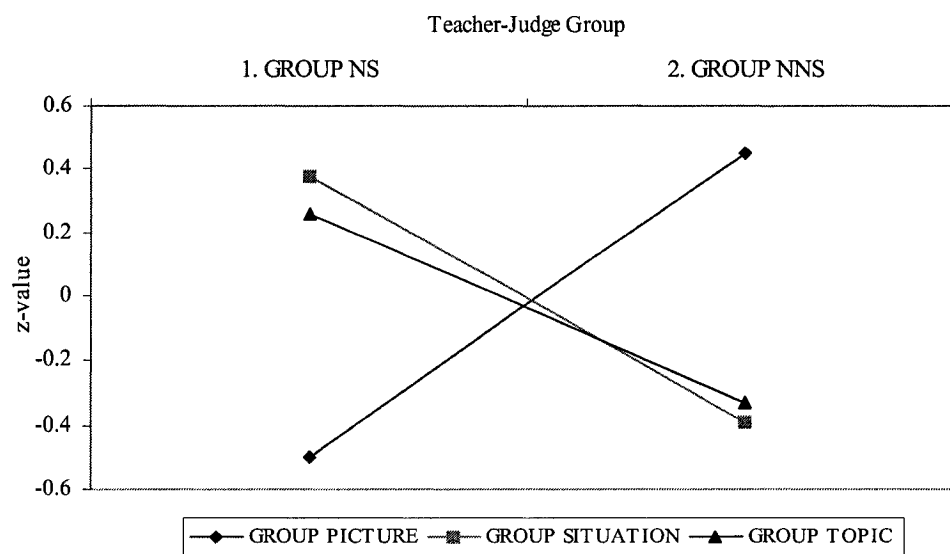


Figure 5. Bias Analysis between Teacher-Judge Groups and Task Types

Table 15. Bias Analysis Report: Interactions between Teacher-Judges and Task Types

Teacher-Judge	Task	Obs-Exp Average	Bias Measure (logits)	Model S.E.	Z-Score	Infit MnSq
NS10	Situation	0.42	-1.66	1.04	-1.59	0.9
NNS10	Situation	0.35	-1.39	1.02	-1.36	0.8
NS11	Topic	0.31	-0.79	0.32	-2.48	0.9

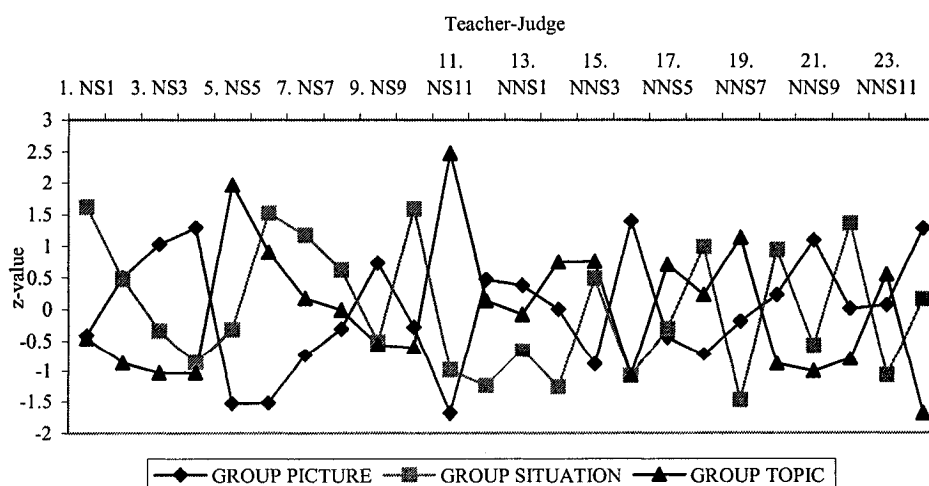


Figure 6. Bias Analysis between Teacher-Judges and Task Types

3. How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks and task types?

3-1) How is the calibration of rating scales influenced across different tasks and task types?

In order to compare the rating scale calibration across different tasks and task types, a hybrid Many-faceted Rasch Measurement Model (Myford & Wolfe, 2004a) was used to apply the Rating Scale Model to teacher-judges, and the Partial Credit Model to tasks and task types. The rating scale calibrations and internal scale structures offered by FACETS were examined for the analysis. Table 16 and Figure 7 show the CATOE scale category statistics and scale structure for T1 (see Appendix I for Tasks 2 – 8). In Table 16, the step calibrations indicate the starting point of the student proficiency measure, at which each step or category begins to be used. Because category 1 starts from infinity, it cannot be determined by numbers. In Figure 7, the measure indicates the student proficiency at which the category begins to be most probable. Table 16 and Figure 7 illustrate all four rating categories function well in all the tasks. Two approaches were used to explore the rating scale calibrations across different tasks and task types. First, the interval of the middle categories 2 and 3 was investigated. It was also noted when category 3 was set as a cut-off or passing line, where the cut-off line is calibrated on the scale.

When the interval of middle categories was examined across different tasks, T8 had the broadest interval (a 5.00 logit spread), while T1 had the narrowest interval (a 3.07 logit spread). This suggests that test-takers are most likely to be awarded middle scores on T8, and extreme scores on T1. In terms of

the cut-off line on the rating scale, the highest proficiency measure was established on T3 (0.51 logits) and the lowest on T8 (-0.34 logits). This means that a test-taker who would have been considered to have passed on T8 might not have passed the other tasks.

A close investigation into rating scale calibration of each task brings T6 and T8 in particular to attention. Although they belong to the same task type (topic-based), their scale calibration ratings display extreme opposite patterns. T8 has the broadest interval of middle categories (a 5.00 logit spread), whereas T6 has the second narrowest interval (a 3.16 logit spread). Similarly, the cut-off line of T8 is much lower (-0.34 logits) than that of T6 (0.00 logits). A similar pattern is also identified in the picture-based tasks: T1 and T4 in particular; the interval of middle categories for T4 is much broader (a 4.79 logit spread) than that for T1 (a 3.07 logit spread) and the cut-off line of T4 is designated much higher (0.28 logits) than that of T1 (-0.23 logits).

The fact that tasks that belong to the same task type are not consonant with each other, and even show extreme opposite patterns in some cases, therefore, points out that task types may not be a principal factor in constructing the rating scale of each task, as they are with difficulty measures. If task types cannot explain the calibration of each rating scale, questions arise as to what the operating principle is. Otherwise, it remains an open question as to whether the variability of rating scale calibrations is due to systematic, as yet undiscovered operations, or to unsystematic random measurement errors caused by human raters.

Table 16. CATOE Scale Category Statistics for Task 1

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-1.42	0.22
3	-0.23	0.18
4	1.65	0.24



Figure 7. CATOE Scale Structure for Task 1

The same analysis was done across different task types. Table 17 and Figure 8 display the CATOE scale category statistics and scale structure for the picture-based task (see Appendix J for the situation-based and topic-based tasks). With regard to the interval of middle categories, the picture-based task had the broadest interval, with a 4.13 logit spread, while the topic-based task had the narrowest interval with a 3.86 logit spread. The difference was only 0.27 logits, which was not as large as in the pervious analysis. Therefore, test-takers may not be significantly affected by task types in being assigned middle or extreme scores. A noticeable difference was found with regard to the cut-off line. While the highest cut-off line was established on the situation-based task with 0.59 logits,³⁰ the difference between the picture-based and topic-based tasks was negligible, with -0.01 logits and 0.02 logits, respectively. Considering that the situation-

³⁰ However, the step calibration of rating category 2 shows fairly large standard errors (0.73). Careful interpretation is needed.

based task was the easiest of the three task types, it is very interesting that the highest cut-off line was established in this task type.

Table 17. CATOE Scale Category Statistics for Picture-Based Task

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-2.06	0.17
3	-0.01	0.13
4	2.07	0.17

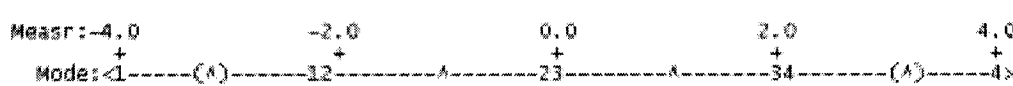


Figure 8. CATOE Scale Structure for Picture-Based Task

3-2) How is the calibration of rating scales influenced by NS and NNS teacher-judges as a whole?

A hybrid Many-faceted Rasch Measurement Model (Myford & Wolfe, 2004a) was used to differentially apply the Rating Scale Model to teacher-judges and tasks, and a Partial Credit Model to teacher-judge groups, thereby providing both groups of teacher-judges with one general rating scale across all tasks. As with the previous analysis, the rating scale calibrations and structures offered by FACETS were examined.

When the interval of middle categories that were designated by the NS and NNS groups were examined (see Table 18 & Figures 9 – 10), it was found that the

NS group used the middle categories a bit more broadly, with a 3.98 logit spread, compared to a 3.71 logit spread for the NNS group. When the scale's cut-off line was examined, there was little difference between the two groups (only 0.04 logits). In other words, the NS group perceived the cut-off line at 0.09 logits, and the NNS group at 0.05 logits. This implies that when the two groups of teacher-judges are in a situation to make decisions about passing or failing students, they establish almost the same cut-off line. Interpreted another way, the two groups have similar ideas about the level students must achieve in order to pass a given test.

Table 18. CATOE Scale Category Statistics for Overall Tasks by NS and NNS Groups

CATOE Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.04	0.13	-1.88	0.13
3	0.09	0.09	0.05	0.09
4	1.94	0.12	1.83	0.11

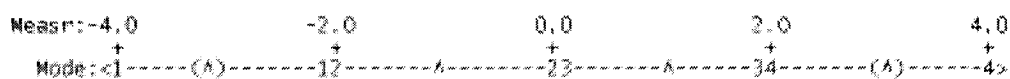


Figure 9. CATOE Scale Structure of NS Group for Overall Tasks

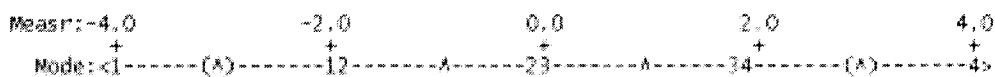


Figure 10. CATOE Scale Structure of NNS Group for Overall Tasks

3-3) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks?

This analysis was completed because it was probable that the two groups would not use a rating scale in the same way across different tasks. In order to examine how the two groups calibrated the rating scale across different tasks, the Rating Scale Model was applied to teacher-judges and the Partial Credit Model to teacher-judge groups and tasks, so that the model provided both groups of teacher-judges with a rating scale across eight tasks. Table 19 and Figures 11 – 12 illustrate the step calibrations and internal scale structures of both groups for T1 (see Appendix K for T2 – T8).

When how broadly the two groups perceive the middle categories of the scale was compared, it was found that the NS group determined middle categories more broadly on T2, T3, T4, T6 and T7. That is, students were more likely to be awarded middle scores on these tasks, but extreme scores on T5 and T8 from the NS group. There was little difference between the two groups on T1 by a less than 0.1 logits. Of these tasks, the T2, T3, T4 and T7 showed that the NS group determined the middle categories of the scale much more broadly than did the NNS group, by a difference of at least 0.70 logits. Interestingly, T2, T4 and T7 were picture-based.

When category 3 was set as a cut-off line, the NS group established higher proficiency measures on T1, T2, T3 and T8. On the other hand, they showed the opposite pattern on T4 and T7, establishing lower proficiency measures. On T5 and T6, the two groups seemed to establish almost the same cut-off line, by a difference of less than 0.1 logits.

Table 19. CATOE Scale Category Statistics for Task 1 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-1.47	0.31	-1.38	0.32
3	-0.11	0.25	-0.35	0.26
4	1.58	0.34	1.73	0.33



Figure 11. CATOE Scale Structure of NS Group for Task 1

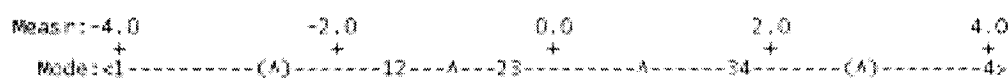


Figure 12. CATOE Scale Structure of NNS Group for Task 1

3-4) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different task types?

Since the two groups of teacher-judges were found to perceive the rating scale categories differently across different tasks, it is worthwhile to explore how the groups interact with the rating scale categories across different task types. This analysis will result in generalized rating scale calibration patterns for the two groups across different task types. The data set was five facets, including two dummy facets (teacher-judge groups and task groups). The Rating Scale Model was employed to the teacher-judge and task facets, and the Partial Credit Model was used to the teacher-judge group and task group facets. The model therefore

provided both groups of teacher-judges with rating scales across three different types of tasks. Table 20 and Figures 13 – 14 show how the two groups interact with the rating scale categories for the picture-based tasks (see Appendix L for the situation-based and topic-based tasks).

When the interval of the middle categories designated by the NS and NNS groups was compared across different task types, the NS group calibrated the middle categories far more broadly on the situation-based task, with a 4.02 logit spread, compared to a 3.01 logits spread for the NNS group. The NS group also designated broader middle categories on the picture-based task than the NNS group, but the difference was not as large as the picture-based task, by 0.51 logits. On the other hand, the NNS group determined broader middle categories than the NS group on the topic-based task, but the difference was small by 0.13 logits. Generally, the two groups of teacher-judges differed critically in how they determined the middle categories for the picture-based task and situation-based task, as opposed to the topic-based task.

When the cut-off line was examined, the NS group established higher proficiency measures on the situation-based task and topic-based task than the NNS group by differences of 0.15 logits and 0.26 logits, respectively. However, when proficiency measures for the picture-based task were compared, the two groups rarely differed exhibiting a difference of less than 0.1 logits. This means that a student whom the NNS group would have felt passed the situation-based or topic-based tasks would not have passed according to the NS group.

Given that the NS and NNS groups did not hold a bias toward particular tasks and task types in terms of severity, it is very interesting that they do in fact

demonstrate different patterns in calibrating rating scales across different tasks and task types. The two groups might award the same scores to test-takers using two different rating scales, and their underlying differences could be easily masked by their overall severity measures. Taking into account that how raters internalize rating scales is an important matter in scoring procedure for the sake of the fairness of the tests, a question arises as to by what operating principle teacher-judges in the two groups show differences in influencing the calibration of rating scales across different tasks and task types.

Table 20. CATOE Scale Category Statistics for Picture-Based Task by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.14	0.18	-1.93	0.18
3	0.05	0.13	0.13	0.13
4	2.10	0.17	1.80	0.15



Figure 13. CATOE Scale Structure of NS Group for Picture-Based Task



Figure 14. CATOE Scale Structure of NNS Group for Picture-Based Task

4. What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks and task types?

4-1) What are the salient construct elements across different tasks and task types?

In order to examine the salient construct elements drawn on by teacher-judges, their comments were analyzed. Table 21 shows the overall frequency and percentage of comments made by the teacher-judges. Of a total of 3295 comments, 2117 were about language use (64.2%), and 517 comments were related to organizational development (15.7%); the two most salient construct elements. Teacher-judges commented 384 times on content effectiveness (11.7%), and paid the least attention to general task fulfillment (4.3%), and socio-contextual appropriateness (4.1%). Overall, teacher-judges were more attentive to language use than socio-contextual appropriateness and content effectiveness, which suggests that although teacher-judges underscore the importance of socio-linguistic competence and topic knowledge to some extent, successful language use is their principal interest when language assessment is concerned.

Table 21. Number and Percentage of Comments for Overall Tasks

	Number of Comments	Percentage of Comments
General Task Fulfillment	143	4.3%
Content Effectiveness	384	11.7%
Language Use	2117	64.2%
Socio-contextual Appropriateness	134	4.1%
Organizational Development	517	15.7%
Total	3295	100.0%

When construct elements were compared across different tasks (see Table 22 for T1 and Appendix M for T2 – T8), language use and organizational development were, in general, the two most salient construct elements on T1, T2, T4 and T7, and language use and socio-contextual appropriateness on T3. On T5, T6 and T8, the teacher-judges drew on language use and content effectiveness most frequently. It is very interesting that the teacher-judges did not pay as much attention to the construct elements on T3, T5, T6, and T8 that are directly related to language itself (i.e., language use or organizational development) as they did on T1, T2, T4 and T7. On those tasks, initially high interest in language use and organizational development seemed drastically reduced, while interest in socio-contextual appropriateness and content effectiveness increased. It appears that when a task included a non-language-related construct element, it was easy for teacher-judges' ideas of what constituted critical evaluation criteria to shift.

Another thing that should be noted is that the number of construct elements that the teacher-judges attended to varied across different tasks. Teacher-judges drew on five different construct elements on T1 and T5, but four on T2, T3, T4, T6, T7 and T8, depending on the task demands and their needs. This implies that the teacher-judges came up with different evaluation criteria, depending on the nature of specific tasks and test-takers performance, and that the comments they made were therefore context-bound. These findings consequently point to the weakness of the theory-based or a priori general language rating scales by providing some support for context-specific rating scales. As this analysis proves, if a general language rating scale is employed, regardless of task demands or test-takers' performance, teacher-judges may not be able to make accurate assessments

and may miss important features of test-takers' performance.

Table 22. Number and Percentage of Comments for Task 1

	Number of Comments	Percentage of Comments
General Task Fulfillment	24	5.7%
Content Effectiveness	36	8.6%
Language Use	274	65.6%
Socio-contextual Appropriateness	13	3.1%
Organizational Development	71	17.0%
Total	418	100.0%

Table 23 shows the frequency and percentage of comments for the picture-based task (see Appendix N for situation-based and topic-based tasks). A prominent feature that stands out is that the number of construct elements also varied across different task types: teacher-judges drew on five different construct elements on the picture-based and topic-based tasks, and four on the situation-based task. As discussed in the previous analysis, this result implies that the teacher-judges exhibited different patterns in attending to salient construct elements, depending on the task demands and the test-takers' performance.

Language use was the most predominant construct element on the picture-based task, followed by the topic-based and situation-based tasks. Considering that the picture-based task required test-takers to describe or narrate visual information, teacher-judges may have been able to predict what test-takers would say on a given task. When teacher-judges are able to easily predict the content information and there are no other construct elements that attract their attention

other than language use, they appear to focus primarily on language use. However, when teacher-judges cannot predict what test-takers will say, as on the topic-based task, and when factors exist that distract their attention from language use, their interest in language use seems to be reduced.

When each task type was investigated, most of the teacher-judges' attention was directed to language use and organizational development on the picture-based task. As was the case with the situation-based task, however, when test-takers were asked to demonstrate socio-linguistic competence as well as other language-related competence, the teacher-judges were very attentive to socio-contextual appropriateness. Therefore, language use and socio-contextual appropriateness were perceived as the two most salient construct elements on the situation-based task by the teacher-judges. As such, when test-takers employed both topical and language knowledge to perform a given task, the interest of the teacher-judges drastically shifted to content effectiveness. On the topic-based task, then, language use and content effectiveness were perceived as the two most salient construct elements by the teacher-judges. In summary, as discussed in the previous analysis, the shifting of the teacher-judges' attention to salient construct elements depended on task demands and test-takers' performance, which supports the necessity for context-specific rating scales.

Table 23. Number and Percentage of Comments for Picture-Based Task

	Number of Comments	Percentage of Comments
General Task Fulfillment	69	3.8%
Content Effectiveness	149	8.2%
Language Use	1260	69.5%
Socio-contextual Appropriateness	13	0.7%
Organizational Development	323	17.8%
Total	1814	100.0%

4-2) What are the salient construct elements drawn on by NS and NNS teacher-judges as a whole?

A comparison of the frequency and percentage of comments made by the NS and NNS groups is presented in Table 24. Strikingly, the NS group made about twice as many comments as the NNS group (2123 comments compared with 1172). The relatively small number of NNS teacher-judges' comments may be because the NNS teachers are not accustomed to making subjective comments on students' performances in an EFL context, rather than awarding them a single score. It could be that the NNS teachers lacked confidence in assessing student spoken English, or because performance assessment is not used as often in an EFL context as pencil and paper tests.

In general, language use was the most frequently tapped construct element for both groups (66.5% and 60.2%, respectively), and organizational development was second (13.8% and 19.0%, respectively). When a chi-square test was conducted to examine potential differences between the two groups, it was found

that the NS and NNS groups significantly differed in how they attended to construct elements they found salient, $\chi^2(4, N = 3295) = 21.19, p = 0.000 < 0.001$. The NS teacher-judges were much more attentive to language use than the NNS teacher-judges by a difference of 5.3%, while the NNS teacher-judges were much more attentive to organizational development than the NS teacher-judges by a difference of 5.2%. Considering that the two groups had not differed significantly in terms of severity measures in the previous analysis, they might have awarded the same scores based on different evaluation criteria.

Table 24. Number and Percentage of Overall Comments by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	95	4.5%	48	4.1%
Content Effectiveness	247	11.6%	137	11.7%
Language Use	1411	66.5%	706	60.2%
Socio-contextual Appropriateness	76	3.6%	58	4.9%
Organizational Development	294	13.8%	223	19.0%
Total	2123	100.0%	1172	100.0%
$\chi^2(4, N = 3295) = 21.19, p = 0.000 < 0.001$				

4-3) What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks?

Generally, as shown in Table 25 and Appendix O, the attention of teacher-judges in both groups was primarily directed towards language use and organizational development on T1, T2, T4 and T7, and language use and socio-

contextual appropriateness on T3. On T5, T6 and T8, the two groups showed somewhat different patterns: while the attention of teacher-judges in both groups was primarily directed towards language use, content effectiveness was the second most salient construct element for the NS group, while organizational development was second for the NNS group.

In addition, a chi-square analysis showed that the NS and NNS groups were not significantly different on T1, T7, and T8, but that they exhibited significant differences on T2, T3, T4, T5 and T6. This indicates that while the two groups may exhibit agreement on salient construct element in some cases, they may exhibit differences in others. An examination of the context in which the same salient construct elements occur is an area for further research.

When the tasks on which the NS and NNS groups showed significant differences were examined, the NS group was far more attentive to language use than the NNS group by a difference of 18.8% on T2, while the NNS group was more attentive to other construct elements. On T3, the NNS group paid more attention to socio-contextual appropriateness than the NS group by a difference of 12.9%, while the NS group paid more attention to language use than the NNS group by a difference of 10.5%. On T4, the NS group focused more on language use than the NNS group by a difference of 5.1%, while the NNS group focused more on content effectiveness than the NS group by a difference of 6.9%. Both groups shared similar patterns on T5 and T6: their interest in language use was somewhat reduced, with a corresponding increase of interest in other construct elements. When the differences between the two groups were examined, the NS group drew on content effectiveness more frequently than did the NNS group,

while the NNS group drew more on organizational development than the NS group.

Table 25. Number and Percentage of Comments for Task 1 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	16	6.0%	8	5.2%
Content Effectiveness	19	7.2%	17	11.1%
Language Use	185	69.8%	89	58.2%
Socio-contextual Appropriateness	5	1.9%	8	5.2%
Organizational Development	40	15.1%	31	20.3%
Total	265	100.0%	153	100.0%
$\chi^2(4, N = 418) = 8.87, p = 0.064 > 0.05$				

4-4) What are the salient construct elements drawn on by NS and NNS teacher-judges across different task types?

Since the NS and NNS groups were found to be significantly different in how they attended to salient construct elements in some cases, it might be worthwhile to examine the dissimilarity patterns they exhibited across different task types. As Table 26 and Appendix P show, the NS and NNS groups differed significantly in attending to construct elements across all the three different task types. On the picture-based task, the NS group was more attentive to language use than the NNS group by a difference of 11.2%, while the NNS group was more attentive to organizational development and content effectiveness by differences of 5.2% and 4.2%, respectively. On the situation-based task, the NS group was

more focused on language use than the NNS group by a difference of 10.5%, while the NNS group was more focused on socio-contextual appropriateness than the NS group by a difference of 12.9%. On the topic-based task, the NS group underscored content effectiveness more than the NNS group by a difference of 6.5%, while the NNS group underscored organizational development more than the NS group by a difference of 7.5%. These results suggest that the NS group consistently drew much more attention to language use on the picture-based and situation-based tasks, while the NNS group was more sensitive to other construct elements (e.g., organizational development or socio-contextual appropriateness). On the topic-based task, the two groups showed a similar amount of primary attention to language use (57.8% and 60.0%), but their secondary interests differed: the NS group was more attentive to content effectiveness, while the NNS group was more attentive to organizational development. In summary, it appears that when a task demanded socio-linguistic competence, the NNS group became more sensitive to socio-contextual appropriateness, and when a task demanded topical knowledge, the NS groups became more sensitive to content effectiveness. When a task did not demand any competence other than language-specific knowledge, the NS group appeared to show more interest in language use, while the NNS group paid attention to other construct elements as well. Speculation on why they exhibit such different underlying perceptions of salient construct elements may be premature at this point; more in-depth, qualitative research is recommended.

Table 26. Number and Percentage of Comments for Picture-Based Task by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	40	3.4%	29	4.5%
Content Effectiveness	79	6.7%	70	10.9%
Language Use	860	73.4%	400	62.2%
Socio-contextual Appropriateness	5	0.4%	8	1.2%
Organizational Development	187	16.0%	136	21.2%
Total	1171	100.0%	643	100.0%
$\chi^2 (4, N = 1814) = 27.64, p = 0.000 < 0.001$				

This chapter has discussed the results of the research questions. The implications regarding these findings and others will be discussed in the next chapter, after the presentation of the summary of the results. Limitations of the study and suggestions for further research are also cited.

CHAPTER 5

CONCLUSION

This study has addressed variability of tasks and teacher-judges in second language oral performance assessment from comprehensive perspectives. The general research questions under investigation were:

- 1) Does the behavior of NS and NNS teacher-judges differ in terms of internal consistency and severity?
- 2) How are task difficulty measures influenced by NS and NNS teacher-judges across different tasks and task types?
- 3) How is the calibration of rating scales influenced by NS and NNS teacher-judges across different tasks and task types?
- 4) What are the salient construct elements drawn on by NS and NNS teacher-judges across different tasks and task types?

The summary of the research findings, their implications and limitations are discussed, along with suggestions for further study.

Summary of the Research Findings

Variability of tasks

This study has examined the complexity and variability of tasks in terms of task difficulty, rating scale calibration, and construct elements. When the difficulty measures of the task types were examined, the situation-based task was

the easiest, while the topic-based task was the most difficult. Although tasks certainly had an influence in determining test scores, the difficulty measures of individual tasks could not be systematically controlled by task types; in other words, task type failed to predict the difficulty measure of a task. Further research exploring the underlying variables that designate task difficulty is necessary in order to enable test developers to select and sequence tasks for test purposes, and curriculum developers to design syllabuses for pedagogical purposes.

When the task difficulty criteria suggested by second language acquisition researchers (i.e., Brown et al, 1984) were applied, their difficulty model was only partially confirmed by this study. A vague hierarchy of task difficulty was captured across different tasks and it was congruent to other studies in second language testing (Brown, Hudson, & Norris, 1999; Elder, Iwashita, & McNamara, 2002; Iwashita, McNamara, & Elder, 2001). The lack of correspondence between the result of this study and those of second language acquisition may be because the definition of task difficulty was not commonly understood (Iwashita et al, 2001). In other words, the notion of task difficulty measures was defined as accuracy, fluency, and complexity in the studies in second language acquisition, while in second language testing research, they were recently introduced as a facet, which is to some extent operationalized by test-takers' ability and raters' severity in the Many-faceted Rasch Measurement (Iwashita et al, 2001). An agreement on the definition and operationalization is warranted for an accurate understanding of principled task difficulty.

Non-systematicity of the performance assessment was also identified in the rating scale calibration. The fact that task types failed to provide an absolute

yardstick for calibrating rating scales of individual tasks not only requires re-conceptualization of task types but also raises questions as to what the latent factors are that exert a crucial influence on rating scale calibration. Moreover, it remains an open question as to whether the variability of rating scale calibrations is due to unsystematic measurement errors caused by human raters, and cannot be resolved by systematic, but as of yet undiscovered, operations.

Nonetheless, tasks conformed to task types when the construct elements were analyzed. Language use and organizational development were the two most salient construction elements on the picture-based task; and language use and language appropriateness were the most salient on the situation-based task. On the topic-based task, language use and content effectiveness were the most frequently tapped construct elements. The variation of construct elements across task types provides evidence that teacher-judges draw on evaluation criteria depending on task demands and test-takers performance. This also provides some support for empirically constructed rating scales.

Variability of teacher-judges.

The NS and NNS teacher-judges rarely differed in the final scores awarded. Teacher-judges in both groups exhibited internally consistent rating patterns: none showed halo, centrality, and extreme effects, and only one from each group showed randomness in his or her rating. Similarly, when overall severity and homogeneity in severity within each group were compared, they were not different. Where a difference was found, a slightly higher percentage of the NNS teacher-judges showed accurate rating patterns compared to the NS teacher-judges.

Even when task effects were taken into consideration, the same picture was obtained: neither of the groups was positively or negatively biased toward a particular task or a particular task type. When the severity measures of the two groups were compared across individual tasks, both groups were most severe on T6 (discussing the harmful effects of Internet use). When the same analysis was done across task types, both groups were most lenient on the situation-based task, and most severe on the topic-based task. More interestingly, a bias analysis carried out between individual teacher-judges and individual tasks showed that one teacher-judge from each group exhibited exactly the same bias patterns on certain tasks.

Substantial dissimilarity emerged in the calibration of the rating scales. Despite little difference between the two groups in calibrating rating scales as a whole, they were obviously different when compared across different tasks. In addition, they were apparently not alike across task types: the NS group was far more likely to award middle scores on the picture-based and situation-based tasks than was the NNS group, with little difference on the topic-based task. Additionally, the NS group established higher proficiency measures on the situation-based and topic-based tasks, with little difference on the picture-based task.

More compelling results were found in the analysis of construct elements salient to the NS and NNS groups. In general, the NS group was much more attentive to language use, and the NNS group was much more attentive to organizational development. Furthermore, task types clearly showed that the NS and NNS teacher-judges had different perceptions of what constructs should be

measured.

Taken together, the NS and NNS teacher-judges appeared to share common ideas as to what score a test-taker should be given, but by different scales and for different reasons. The underlying differences about scale construction and the construct of interest raise questions as to whether this discrepancy is caused by the innate perceptions of the NS and NNS teacher-judges in different contexts, or whether it can be removed if specific a priori evaluation criteria and rigorous rater training are provided. It is also questionable how the validity of the ratings can be defined or even justified, if the difference is persistent.

Implications

The need for context (task)-specific assessment.

The results of this study provide some support for the claim that multiple tasks should be employed for the test to assess the diverse oral language output of test-takers (Chalhoub-Deville, 1995a, 1995b; Henning, 1983; Shohamy, 1983, Shohamy, Reves, & Bejerano, 1986; Upshur & Turner, 1999). As shown in the study, not only did test scores tend to fluctuate across different tasks, but teacher-judges had different perceptions that influenced the calibration of rating scales and the underscoring of the underlying construct elements that were to be measured. It therefore justifies the argument that a test should embrace as many different types of tasks as possible in order to tap overall language ability, as well as for the sake of the fairness of the test.

These findings also validate the necessity of an empirically-derived rating scale suggested by Chalhoub-Deville (1995a, 1995b), Turner and Upshur (1996), and Upshur and Turner (1995, 1999). Since teacher-judges drew attention to different evaluation criteria or construction elements according to task demands, as well as to test-takers' performance, a priori general language proficiency rating scales may not provide meaningful information concerning what is being measured as much as context- or task-specific rating scales do. Moreover, the discrepancy that arises between a test and the test rating scale will certainly threaten the validity of the test when the rating scales are not bound in context. Although Fulcher (2003) speculates that task-specific variance is due to the use of task-specific rating scales in those studies (i.e., Chalhoub-Deville, 1995a, 1995b; Turner & Upshur, 1996; Upshur & Turner, 1995, 1999), this study shows that this is not the case. The teacher-judges who participated in this study were provided with only a general rating scale which did not delineate any task or language-specific features, but they did come up with task or context-specific evaluation features. Even though tasks accounted for extremely small variances in determining test-takers' scores, with most being accounted for by the test-takers' ability,³¹ as with the case of other studies (e.g., Bachman, Lynch, & Mason, 1995; Bonk & Ockey, 2003; Fulcher, 1993, 1996a; Lynch & McNamara, 1998), different task or language features were indeed embedded across different tasks. A reasonable interpretation is that context-specific tasks or rating scales do not necessarily lead to significant score differences, but still provide meaningful

³¹ In order to compare variances among test-takers, tasks and teacher-judges, examine the standard deviation of each facet in the FACETS measurement report.

information about test-takers' performance in a given context.

As Alderson (1991) notes, in order for the test scores to be meaningful, the scale should be related to both the language constructs to be measured and the purposes of the test in a specific context. Likewise, the results of this study have proved that empirically-constructed rating scales need to be employed, depending on the task types and contexts in which they will be used; the decision should depend on the dynamics of the test situation and voices from test-takers, test-constructors, curriculum makers, and language policy makers should certainly be mingled.

Suitability of the NNS teacher-judges.

This study suggests that there is little evidence that the NNS teacher-judges are unsuited to assess students' second oral language performance. Within an EFL context, there has been a persistent folk belief that only NS teachers are able to assess students' performance fairly, reliably and validly, and that NNS teachers are unsuited to judge the language skills of others due to their own lack of mastery of the language. This groundless belief has bestowed too much authority upon NS teachers while taking power and authority away from NNS teachers. However, this study has proved that when the NNS teacher-judges had sufficient teaching experience and educational background, they were able to work as qualified assessors.

Where a difference was found between the NS and NNS teacher-judges, the issue was how they considered students' performance in order to reach a certain score. In other words, the high reliability between the NS and NNS groups

should not be assumed to be evidence that the constructs focused on by the NS and NNS groups share a common nature or are equally valid. Of course, the latent differences that exist between them do not necessarily imply a judgment of what is right or wrong. At issue is the approach by which students are assessed more validly and meaningfully within a given context (either an ESL or an EFL context), and this question opens a new area for further research.

Usefulness of the Many-faceted Rasch Measurement.

This study also provides strong support for the validity of the Many-faceted Rasch Measurement in analyzing language performance data. The Rasch model has been criticized because the data specification of unidimensionality cannot properly satisfy the complexity of the constructs underlying language performance (Buck, 1994; Hamp-Lyons, 1989). However, in the “debate over the constructs and dimensionality” (McNamara, 1996, p. 268), McNamara (1996) and others (Henning, 1992; Lumsden, 1976) pointed out the confusion between the psychometric dimension and the psychological dimension. In the same vein, this study showed that teacher-judges measured not only the language-relevant constructs but also language-irrelevant constructs (i.e., topical knowledge) when they assess test-takers performance on a certain task, but that such tasks were still within a good fit range demonstrates that language performance data can hold psychometric unidimensionality, even when they are compounded with psychologically different constructs.

Contrary to criticism about the Rasch Measurement, the usefulness of this measurement tool has increasingly been reported in the performance assessment

literature (Bonk & Ockey, 2003; Brown, 1995; Hill, 1997; Kondo-Brown, 2002; Lumley & McNamara, 1995; Lynch & McNamara, 1998; Weigle, 1998). Unlike classical test theory (or true score theory), which determines the ability of an examinee in a particular test condition, the Rasch theory provides test-takers with more reliable and generalized information estimating their latent abilities with freeing them from the particulars of test conditions. In addition, the fit-statistics offered by the Rasch Measurement allow rater behaviors to be monitored, and raters to be provided with individual feedback about the internal consistency and bias of their rating patterns during the rater training process. Likewise, the item fit analysis can guide test developers with regard to which items should be selected, revised, or thrown out during the test development process. It is thus apparent that the Rasch Measurement is not only appropriate, but also recommended for language performance data.

Legitimacy of mixed methods research.

By combining the Many-faceted Rasch Measurement with grounded theory, this study proves the legitimacy of mixed methods research. Mixed methods, known as the third wave of research movement, incorporating quantitative and qualitative research techniques and methods in a single study, has been expanded in the social and human sciences (Johnson & Onwuegbuzie, 2004). Along with the development of quantitative and qualitative research methods, the use of multiple methods has the potential to reduce the problems embedded in singular methods while maximizing the strengths (Sechrest & Sidana, 1995, as cited in Johnson & Onwuegbuzie, 2004).

The qualitative analysis (i.e., grounded theory analysis) conducted in this study has supplied meaningful interpretations of how NS and NNS teacher-judges differ in assessing student oral language performance, which would otherwise have been masked by the sole reporting of quantitative results. It is evident that mixed methods research is a more comprehensive tool that offers greater insight and understanding into the questions posed by the research. It appears time for researchers to develop a new research paradigm and delve into its philosophical concepts and bases rather than blindly advocating the traditional purists' incompatibility thesis (Howe, 1988).

Limitations of the Study and Suggestions for Further Research

Despite the attempts to minimize possible drawbacks, the study includes some limitations. First of all, the reliability and validity of the CATOE could not be precisely examined before it was chosen as an instrument in this study. The CATOE was specifically developed to conduct this study, and due to constraints it was not possible to have a parallel form administered to experiment the instrument. Nevertheless, the results of the study confirmed the high validity of the test and the reliability of the teacher-judges. The test placed test-takers into various ability groups, corresponding with the class levels to which they are assigned at the language institute, and also most teacher-judges exhibited an acceptable range of variation in scoring the test-takers' performance. These results reflect the high criterion-related validity of the test and the high reliability of the teacher-judges.

However, one thing should be pointed out in terms of test design. The test did not include three types of tasks in the same ratio: while four were picture-based tasks, there was only one situation-based task. It is certainly questionable whether the task asking test-takers to congratulate a friend on being admitted to school was sufficient to assume an overall pragmatic ability. Having just one situation-based task may not allow an adequate variety of contexts for the meaningful interpretation of research outcomes. Thus, care must be taken when interpreting the nature of the situation-based task in this study. A replication of the study with more diverse situation-based tasks may show more decisive results.

There is also concern about the extent to which the results of the study can be generalized from the teacher-judges who participated in the study to an entire population of teacher-judges. Having had only Canadian or Korean teachers of English in the sample, most of whom were well-qualified, experienced teachers with at least one graduate degree related to linguistics or language education, it might be unwise to apply the results of this study to other contexts and populations. By limiting the research outcomes to the specific context in which this study was carried out, the population validity of the study will be maintained.

Further study is also recommended to determine the steps that the NS and NNS teacher-judges took in the decision-making process, even when they agreed on a score. In this study, the only available data from which emergent constructs were drawn were written comments, which failed to offer a full account of the teacher-judges' in-depth perception. If that perception was derived by means of verbal protocols or in-depth interviews, a clearer picture about what they think

makes for good language performance, and what should consequently be measured would be gained.

Equally importantly, there are questions as to how much the teacher-judges who participated in the study were motivated to judge test-takers' performance. Unfortunately, the study did not contain a procedure to measure teacher-judge motivation or attitudes toward scoring test-takers' performance. Teacher-judges' de-motivation or fatigue effect is probably the largest unwanted variable that has the potential to mislead the outcome of the study. Further qualitative research will add some insight into their motivation and their attitudes towards the rating process.

This study has addressed the complexity and variability of performance assessment across different tasks and teacher-judge groups. Teacher-judges' perceptions of task difficulty, rating scale calibration, and construct elements will of necessity be reflected in their feedback on student second language performance, and will have the potential to affect future teaching objectives, course content and curriculum. The impact will be even more significant when scores that are biased due to measurement error prevent stakeholders from accurately inferring test-takers' capabilities. This is something that is particularly relevant with regard to the results of classroom tests that contribute to class final marks and externally developed high-stakes tests that involve raters to make important decisions that affect the futures of those who take them. By clarifying these issues, it is hoped that the findings of this study will contribute to

communicative language testing research, and provide educators with a better understanding of second language performance testing.

References

- Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Ed.), *Language testing in the 1990s* (pp. 71-86). London: Macmillan.
- Allen, P., Cumming, J., Mougeon, R., & Swain, M. (1983). *Development of bilingual proficiency: Second year report*. Toronto: The Ontario Institute for Studies in Education.
- Allen, J. P. B., & Widdowson, H. G. (1975). Grammar and language teaching. In J. P. B. Allen & S. P. Corder (Ed.), *The Edinburgh course in applied linguistics, Vol. 2*. Oxford: Oxford University Press.
- American Council on the Teaching of Foreign Languages (ACTFL). (1999). *Revised ACTFL Proficiency Guidelines – Speaking*. Yonkers, NY: American Council on the Teaching of Foreign Languages.
- Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, 38, 665-680.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1982). The Construct validation for some components of communicative proficiency. *TESOL Quarterly*, 16, 449-465.

- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12, 238 - 257.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Baker, R. (1997). *Classical test theory and item response theory in test analysis*. Special report No 2: Language Testing Update.
- Barnwell, D. (1989). 'Naïve' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, 6, 152-163.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20, 89-110.
- Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Ed.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge: Cambridge University Press.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12, 1-15.
- Bronw, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Brown, J. D., Hudson, T., & Norris, J. (1999). *Validation of test-dependent and task-independent ratings of performance assessment*. Paper presented at the 21st Language Testing Research Colloquium, Tsukuba, Japan.

- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11, 145-170.
- Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Ed.), *Language and communication* (pp. 2-27). London: Longman.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47.
- Chalhoub-Deville, M. (1995a). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, 12, 16-33.
- Chalhoub-Deville, M. (1995b). A contextualized approach to describing oral language proficiency. *Language Learning*, 45, 251-281.
- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14, 3-22.
- Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, 18, 65-81.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Clark, J. L. D., & Swinton, S. S. (1980a). *An exploration of speaking proficiency measures in the TOEFL context* (TOEFL Report No. 4). Princeton, NJ: Educational Testing Service.
- Clark, J. L. D., & Swinton, S. S. (1980b). *The test of spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Report No. 7). Princeton, NJ: Educational Testing Service.

- Clark, J. L. D., & Clifford, R. T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status, and needed research. *Studies in Second Language Acquisition*, 10, 129-147.
- Clifford, R. T. (1978). Reliability and validity aspects contributing to oral proficiency of prospective teachers of German. In Clark, J. L. D (Ed.), *Direct testing of speaking proficiency: theory and application* (pp.191-209). Princeton, NJ: Educational Testing Service.
- Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Ed.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: National Council of Teachers of English.
- Davidson, F. G. (1988). *An exploratory modeling survey of the trait structures of some existing language test datasets*. Unpublished doctoral dissertation, University of California at Los Angeles, CA.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.
- Educational Testing Service (ETS). (2001). *TSE and SPEAK score user guide: 2001-2002 edition*. Princeton, NJ: Educational Testing Service. Retrieved October 11, 2004, from <http://www.toefl.org/tse/tseindex.html>.
- Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing*, 19, 347-368.
- Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37, 313-326.

- Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal*, 41, 287-291.
- Fulcher, G. (1988). *Lexis and reality in oral evaluation*. Washington DC: ERIC Clearing house for languages and linguistics. (ERIC Document Reproduction Service NO. ED298759)
- Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. Unpublished doctoral dissertation, University of Lancaster, UK.
- Fulcher, G. (1996a). Testing tasks: Issues in task design and the group oral. *Language Testing*, 13, 23-51.
- Fulcher, G. (1996b). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13, 208-238.
- Fulcher, G. (1997). The testing of L2 speaking. In C. Clapham & D. Corson (Ed.), *Encyclopedia of language and education: Volume 7 Language testing and assessment* (pp. 75-85). London: Kluwer.
- Fulcher, G. (2003). *Testing second language speaking*. London: Longman.
- Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, 64, 428-433.
- Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, 41, 1-24.
- Haertel, E. (1992). Performance measurement. In M. C. Alkin (Ed.), *Encyclopedia of educational research*, 6th edition (pp. 984-989). NY: Macmillan.
- Halliday, M. A. K. (1970). Language structure and language function. In J. Lyons

- (Ed.), *New horizons in linguistics*. Harmondsworth: Penguin.
- Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.
- Hamp-Lyons, L. (1989). Applying the partial credit model of Rasch analysis: Language testing and accountability. *Language Testing*, 6, 109-118.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.
- Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, 29, 759-762.
- Harley, B., Allen, P., Cummins, J., & Swain, M. (1990). The nature of language proficiency. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 7-25). Cambridge: Cambridge University Press.
- Henning, G. (1983). Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Language Learning*, 33, 315-332.
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9, 1-11.
- Hill, K. (1997). Who should be the judge?: The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language Assessment: Proceedings of LTRC 96* (pp. 275-

- 290). Jyväskylä: University of Jyväskylä and University of Tampere.
- Hinofotis, F. B., Bailey, K. M., & Stern, S. L. (1981). Assessing oral proficiency of prospective foreign teaching assistants: Instrument development. In A. Palmer, P. J. M. Groot, & G. A. Trosper (Ed.), *Construct validation of tests of communicative competence* (pp. 106-126). Washington, D. C: TESOL.
- Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis, or, Dogmas die hard. *Educational Researcher*, 17, 10-16.
- Hymes, D. H. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, 23, 8-38.
- Hymes, D. H. (1968). The ethnography of speaking. In J. Fishman (Ed.), *Readings in the sociology of language*. The Hague: Mouton.
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Ed.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.
- Interagency Language Roundtable (ILR). (1991). *Interagency Language Roundtable Skill Level Descriptions*. Interagency Language Roundtable.
- Retrieved January 12, 2005, from <http://www.govtilr.org/ILRscale1.htm>
- Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning*, 51, 401-436.
- Johnson, K. (1977). The adoption of functional syllabuses for general language teaching courses. *Canadian Modern Language Review*, 33, 667-680.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A

research paradigm whose time has come. *Educational Researcher*, 33, 14-26.

- Jones, R. L. (1985). Second language performance testing: an overview. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Ed.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19, 3-31.
- Lantolf, J. P., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *Modern Language Journal*, 69, 337-345.
- Lantolf, J. P., & Frawley, W. (1988). Proficiency; Understanding the construct. *Studies in Second Language Acquisition*, 10, 181-195.
- Lee, Y. (2003). *Investigating differential rater functioning for academic writing samples: An MFRM approach*. Paper presented at the National Council on Measurement in Education, Chicago, IL.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (1997). Investigating judge local independence. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 11, 546-547.
- Linacre, J. M. (2005). A user's guide to Facets: Rasch-model computer programs. [Computer software and manual]. Retrieved April 10, 2005, from www.winsteps.com.
- Linacre, J. M., & Williams, J. (1998). How much is enough? *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 12, 653.
- Lowe, P., & Clifford, R. T. (1980). Developing an indirect measure of overall oral

- proficiency. In J. R. Firth (Ed.), *Measuring spoken language proficiency* (pp. 31-39). Washington DC: Georgetown University Press.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54-71.
- Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, 27, 251-280.
- Lunz, M. E., & Stahl, J. A. (1990). Judge severity and consistency across grading periods. *Evaluation and the health professions*, 13, 425-444.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158-180.
- Malone, M. (2000). Simulated Oral Proficiency Interviews: Recent Developments. (2000, December). Retrieved February 23, 2005, from <http://www.cal.org/resources/digest/0014simulated.html>
- Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal*, 44, 117-121.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
- McNamara, T. F. (1997). Performance testing. In C. Clapham & D. Corson (Ed.), *Encyclopedia of language and education: Volume 7 Language testing and assessment* (pp. 131-139). London: Kluwer.
- Messick, S. (1994). The interplay of evidence and consequences in the validation

- of performance assessments. *Educational Researcher*, 23, 12-23.
- Morrow, K. E. (1977). *Techniques of evaluation for a national syllabus*. Reading: Center for Applied Language Studies, University of Reading.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5-12.
- Multimedia Assisted Test of English (MATE). (2000). About MATE. Retrieved January 12, 2005, from <http://www.mate.or.kr/english/index.html>
- Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (Monograph Series No. 94-05). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.
- Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the test of spoken English Assessment System* (Research Report No. 00-06). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.
- Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In Smith, Jr., E. V. & Smith, R. M. (Ed.), *Introduction to Rasch measurement* (pp. 460-517). Maple Grove, MN: JAM Press.
- Myford, C. M., & Wolfe, E. W. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In Smith, Jr., E. V. & Smith, R. M. (Ed.), *Introduction to Rasch measurement* (pp. 518-574). Maple Grove, MN: JAM Press.
- North, B. (1995). The development of a common framework scale of descriptors

- of language proficiency based on a theory of measurement. *System*, 23, 445-465.
- North, B. (1996). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. Unpublished doctoral dissertation, Thames Valley University.
- North, B. (1997). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. In A. Huhta, V. Hohonen, L. Kurki-Suonio, & S. Luoma (Ed.), *Current developments and alternatives in language Assessment: Proceedings of LTRC 96* (pp. 423-447). Jyväskylä: University of Jyväskylä and University of Tampere.
- North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15, 217-263.
- O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, 12, 217-237.
- Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237-255.
- Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition*, 10, 217-243.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.
- Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85,

956-970.

- Sechrest, L., & Sidana, S. (1995). Quantitative and qualitative methods: Is there an alternative? *Evaluation and Program Planning*, 18, 77-87.
- Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning*, 33, 527-540.
- Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition*, 10, 165-179.
- Shohamy, E. (1990). Discourse analysis in language testing. *Annual Review of Applied Linguistics*, 11, 115-128.
- Shohamy, E. (1996). Competence and performance in language testing. In G. Brown, K. Malmkjær, & J. Williams (Ed.), *Performance and competence in second language acquisition* (pp. 138-151). Cambridge: Cambridge University Press.
- Shohamy, E., Reves, T., & Bejerano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal*, 76, 212-220.
- Shohamy, E., Shmueli, D., & Gordon, C. M. (1991). *The validity of concurrent validity of a direct vs. semi-direct test of oral proficiency*. Paper presented at the 13th Language Testing Research Colloquium, Educational Testing Service, Princeton, NJ.
- Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76, 27-33.

- Slater, S. J. (1980). Performance testing: Overview. In J. E. Spirer (Ed.), *Performance testing: Issues facing vocational education* (pp. 3-17). Columbus, OH: National Center for Research in Vocational Education.
- Smith, Jr., E. V. (2004a). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In Smith, Jr., E. V. & Smith, R. M. (Ed.), *Introduction to Rasch measurement* (pp. 93-122). Maple Grove, MN: JAM Press.
- Smith, Jr., E. V. (2004b). Metric development and score reporting in Rasch measurement. In Smith, Jr., E. V. & Smith, R. M. (Ed.), *Introduction to Rasch measurement* (pp. 342-365). Maple Grove, MN: JAM Press.
- Smith, Jr., E. V. (2004c). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In Smith, Jr., E. V. & Smith, R. M. (Ed.), *Introduction to Rasch measurement* (pp. 575-599). Maple Grove, MN: JAM Press.
- Spolsky, B. (1975). Linguistics in practice: The Navajo reading study. *Theory into Practice*, 14, 347-352.
- Spolsky, B. (1977). *Language testing: Art or science*. Paper presented at the 4th International Congress of Applied Linguistics.
- Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Bradley & D. K. Stevenson (Ed.), *Practice and Problems in Language Testing 1: Proceedings of the First International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe held at the*

Bundessprachenamt, Hürth 29-31 July 1979 (pp. 5-21). Frankfurt: Verlag

Peter D. Lang.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (Ed.), *Current developments in language testing* (pp. 199-209). Singapore: SEAMEO Regional Language Center.

Stansfield, C. W., & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, 72, 129-141.

Stansfield, C. W., & Kenyon, D. M. (1992b). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20, 347-364.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I., & Antonia, M. (1990). The development and validation of the Portuguese Speaking Test, *Hispania*, 73, 641-651.

Stern, H. H. (1978). *The formal-functional distinction in language pedagogy: A conceptual clarification*. Paper presented at the 5th AILA Congress, Montreal, August. Mimeo.

Strauss, A., & Corbin, J. M. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. London: Sage.

Swain, M. (1985). Communicative competence: Some roles of comprehensible

- input and comprehensible output in its development. In S. Gass & C. Madden (Ed.), *Input in second language acquisition*. Rowley, MA: Newbury House.
- Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Ed.), *The language testing cycle: From inception to washback* (pp. 55-79). Melbourne: Australian Review of Applied Linguistics.
- Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *Language Testing*, 36, 49-70.
- Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.
- Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Testing Journal*, 49, 3-12.
- Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing*, 16, 82-111.
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15, 263-287.
- Weinstein, A. I. (1979). Steps in a speaking test. In Firth, J. R. (Ed.), *Testing Kit: French and Spanish* (pp. 106-110). Washington DC.

- White, E. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass.
- Widdowson, H. G. (1978). *Teaching language as communication*. London: Oxford University Press.
- Wilkins, D. A. (1976). *Notional syllabuses*. London: Oxford University Press.
- Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (1999). *The manifestation of common rater effects in multi-faceted Rasch analyses* (Monograph Series No. 97-20). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.
- Wright, B. D. (1991). Scores, reliabilities and assumptions. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 5, 157-158.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8, 370.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Appendices

APPENDIX A: STUDENT QUESTIONNAIRE

학생용 설문지

이 설문지는 '영어말하기시험 채점관들의 의식연구' 논문을 위한 자료수집의 한 방법으로 여러분의 영어학습에 대한 일반적인 배경을 알아보고자 합니다. 아래의 질문들은 여러분들에 관한 기본적인 신상질문과 한국 및 캐나다에서의 영어학습실태에 관한 질문들로 이루어져 있습니다. 이에 대한 여러분의 답변은 논문결과해석에 유익한 정보를 제공할 것입니다.

1. 나이:

☐ 20 - 24세

☐ 25 - 29세

☐ 30 - 34세

☐ 35세 이상

2. 성별: ☐ 남

☐ 여

3. 교육수준:

☐ 대학생 (1학년, 2학년, 3학년, 4학년)

☐ 대학 졸업자

4. 전공: _____

5. 한국에서 영어를 공부한 기간은 얼마입니까?

☐ 6년 미만

☐ 6 - 7 년

☐ 8 - 9 년

☐ 10년 이상

6. 영어권 국가에서 체류를 하고 있는 기간은 얼마입니까?

☐ 7개월 미만

☐ 7 - 12 개월

☐ 13 - 18 개월

☐ 19개월 이상

7. 현재 수학하고 있는 어학원에서 어느 단계의 반에 편성이 되어 있습니까?

- ☐ 1단계 ☐ 2 단계 ☐ 3 단계 ☐ 4 단계 ☐ 5 단계

8. 스스로 평가한 자신의 영어말하기 수준은 어느 정도입니까?

- ☐ 초급 ☐ 중하급 ☐ 중급 ☐ 중상급 ☐ 고급

9. 영어를 공부하는 목적은 무엇입니까?

- ☐ 직업 및 실무적 성취를 위해서
☐ 학업적 성취를 위해서
☐ 구어영어의 의사소통능력을 향상시키기 위해서
☐ 기타: _____

10. 현재 수학하고 있는 어학원에서는 영어의 어떤 부분에 중점을 두고 교육을 받습니까?

- ☐ 듣기 ☐ 읽기 ☐ 말하기 ☐ 쓰기

11. 현재 수학하고 있는 어학원에서는 주당 몇 시간을 영어말하기 교육에 할애를 합니까?

- ☐ 6시간 미만 ☐ 6 - 10시간
☐ 11 - 15시간 ☐ 16 - 20시간

12. 한국에서는 영어의 어떤 부분에 중점을 두고 교육을 받았습니까?

- ☐ 듣기 ☐ 읽기 ☐ 말하기 ☐ 쓰기

13. 공식적인 영어말하기시험이나 영어 인터뷰를 받아 본 경험이 있습니까?

- ☐ 네 ☐ 아니오

공식적인 영어말하기시험이나 영어 인터뷰를 받아 본 경험이 있다면, 횟수를 말씀해 주십시오. _____

STUDENT QUESTIONNAIRE (ENGLISH VERSION)

Your answers to the following questions will help me to better understand your teaching and evaluation methods. All information will remain confidential, and will be used for research purposes only. Thank you for your time.

1. Age:

☐ 20 – 24

☐ 25 – 29

☐ 30 – 34

☐ above 35

2. Gender: ☐ Male

☐ Female

3. Educational background:

☐ Undergraduate in progress (U1, U2, U3, U4)

☐ Completed undergraduate

4. Academic major: _____

5. How many years did you study English in Korea?

☐ Fewer than 6 years

☐ 6 – 7 years

☐ 8 – 9 years

☐ 10 years or more

6. How many months have you lived in English-speaking countries?

☐ Fewer than 7 months

☐ 7 – 12 months

☐ 13 – 18 months

☐ 19 months or more

7. What is your class level in the language institute you currently attend?

☐ Level 1

☐ Level 2

☐ Level 3

☐ Level 4

☐ Level 5

8. What is your self-assessed level of spoken English?

- ☐ Beginner
- ☐ Lower-intermediate
- ☐ Intermediate
- ☐ Upper-intermediate
- ☐ Advanced

9. What is your reason for studying English?

- ☐ Business
- ☐ Academic
- ☐ Improved personal communication skills
- ☐ Other, Specify: _____

10. What skills does your language institute English class focus on?

- ☐ Listening
- ☐ Reading
- ☐ Speaking
- ☐ Writing

11. In your class, how many hours per week are devoted to speaking skills?

- ☐ Fewer than 5 hours
- ☐ 5 hours to 10 hours
- ☐ 11 hours to 15 hours
- ☐ 15 hours to 20 hours

12. What skills did your English class in Korea focus on?

- ☐ Listening
- ☐ Reading
- ☐ Speaking
- ☐ Writing

13. Have you ever taken an oral English test or participated in an oral English interview?

- ☐ Yes
- ☐ No

If yes, please specify the number of times you have done so: _____

APPENDIX B: NS TEACHER QUESTIONNAIRE

TEACHER QUESTIONNAIRE

Your answers to the following questions will help me to better understand your teaching and evaluation methods. All information will remain confidential, and will be used for research purposes only. Thank you for your time.

I. Background Information

1. Age:

☐ 20 – 29

☐ 30 – 39

☐ 40 – 49

☐ above 50

2. Gender: ☐ Male

☐ Female

3. First language(s): _____

If you are bilingual in English, please specify the other language(s) you speak:

4. Educational background:

☐ B.A. in _____

☐ M.A. in _____

☐ Ph.D. in _____

5. Do you have specific training in ESL?

☐ Yes

☐ No

6. How many years have you taught English to non-native English speakers?

☐ Less than 3 years

☐ 3 – 6 years

☐ 7 – 10 years

☐ 11 years or more

7. In what type of language institute do (did) you teach?

- ☐ Private language institute
- ☐ College/University-bound language institute
- ☐ College/University
- ☐ Other, Specify: _____

8. How many hours of English do (did) you teach a week?

- ☐ Less than 5 hours
- ☐ 5 – 10 hours
- ☐ 11 – 15 hours
- ☐ 16 hours or more

9. Please specify course titles you have taught in the past or that you currently teach:

10. Have you ever taught English in non-English-speaking countries?

- ☐ Yes ☐ No

If yes, specify the country/countries and the number of years/months:

11. How familiar are you with the spoken English of non-native English speakers?

- ☐ A little ☐ Some ☐ A lot ☐ Very familiar

II. Evaluation of Spoken English

12. Have you taken courses specifically in testing and evaluation?

- ☐ Yes ☐ No

13. Have you ever been trained as a rater of spoken English?

- ☐ Yes ☐ No

If yes, specify the year(s) that you received training and the number of

training hours completed (i.e., dates): _____

14. How familiar are you with rating the spoken English of non-native English speakers?

- ☐ A little ☐ Some ☐ A lot ☐ Very familiar

15. What tools do you use to evaluate spoken English?

- ☐ Anecdotal notes (use word descriptions)
☐ Checklists
☐ Rating scales
☐ Marks, scores (use numbers)
☐ Other, Specify: _____

16. Have you ever used a rating scale to evaluate spoken English in your classroom evaluation?

- ☐ Yes ☐ No

If yes, what kind of rating scale have you used?

- ☐ Holistic rating scales
☐ Analytic rating scales
☐ Empirical rating scales
☐ Other, Specify: _____

17. When you rate speech samples in this study, how many times did you listen to them, on average?

- ☐ Once
☐ Twice
☐ Three times
☐ More than three times

APPENDIX C: NNS TEACHER QUESTIONNAIRE

TEACHER QUESTIONNAIRE

Your answers to the following questions will help me to better understand your teaching and evaluation methods. All information will remain confidential, and will be used for research purposes only. Thank you for your time.

I. Background Information

1. Age:

☐ 20 – 29

☐ 30 – 39

☐ 40 – 49

☐ above 50

2. Gender: ☐ Male

☐ Female

3. First language(s): _____

If you are bilingual in Korean, please specify the other language(s) you speak:

4. Educational background:

☐ B.A. in _____

☐ M.A. in _____

☐ Ph.D. in _____

5. Do you have specific training in ESL?

☐ Yes

☐ No

6. How many years have you taught English to non-native English speakers?

☐ Less than 3 years

☐ 3 – 6 years

☐ 7 – 10 years

☐ 11 years or more

7. In what type of language institute do (did) you teach?

- ☐ Private language institute
- ☐ College/University-bound language institute
- ☐ College/University
- ☐ Other, Specify: _____

8. How many hours of English do (did) you teach a week?

- ☐ Less than 5 hours
- ☐ 5 – 10 hours
- ☐ 11 – 15 hours
- ☐ 16 hours or more

9. Please specify course titles you have taught in the past or that you currently teach:

10. Have you ever studied in English-speaking countries?

- ☐ Yes ☐ No

If yes, specify the number of years/months you studied in such countries: _____

11. Indicate your English proficiency level:

- ☐ Upper-intermediate ☐ Advanced ☐ Near-native

II. Evaluation of Spoken English

12. Have you taken courses specifically in testing and evaluation?

- ☐ Yes ☐ No

13. Have you ever been trained as a rater of spoken English?

- ☐ Yes ☐ No

If yes, specify the year(s) that you received training and the number of training hours completed (i.e., dates): _____

14. How familiar are you with rating the spoken English of non-native English speakers?

- ☐ A little ☐ Some ☐ A lot ☐ Very familiar

15. What tools do you use to evaluate spoken English?

- ☐ Anecdotal notes (use word descriptions)
☐ Checklists
☐ Rating scales
☐ Marks, scores (use numbers)
☐ Other, Specify: _____

16. Have you ever used a rating scale to evaluate spoken English in your classroom evaluation?

- ☐ Yes ☐ No

If yes, what kind of rating scale have you used?

- ☐ Holistic rating scales
☐ Analytic rating scales
☐ Empirical rating scales
☐ Other, Specify: _____

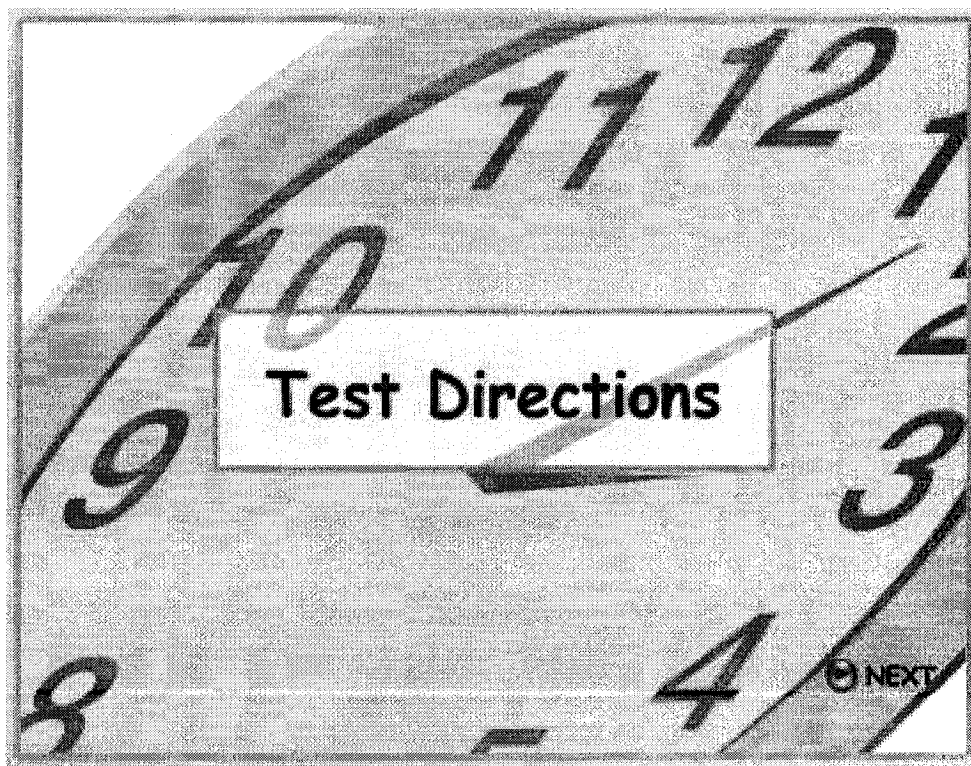
17. When you rate speech samples in this study, how many times did you listen to them, on average?

- ☐ Once
☐ Twice
☐ Three times
☐ More than three times

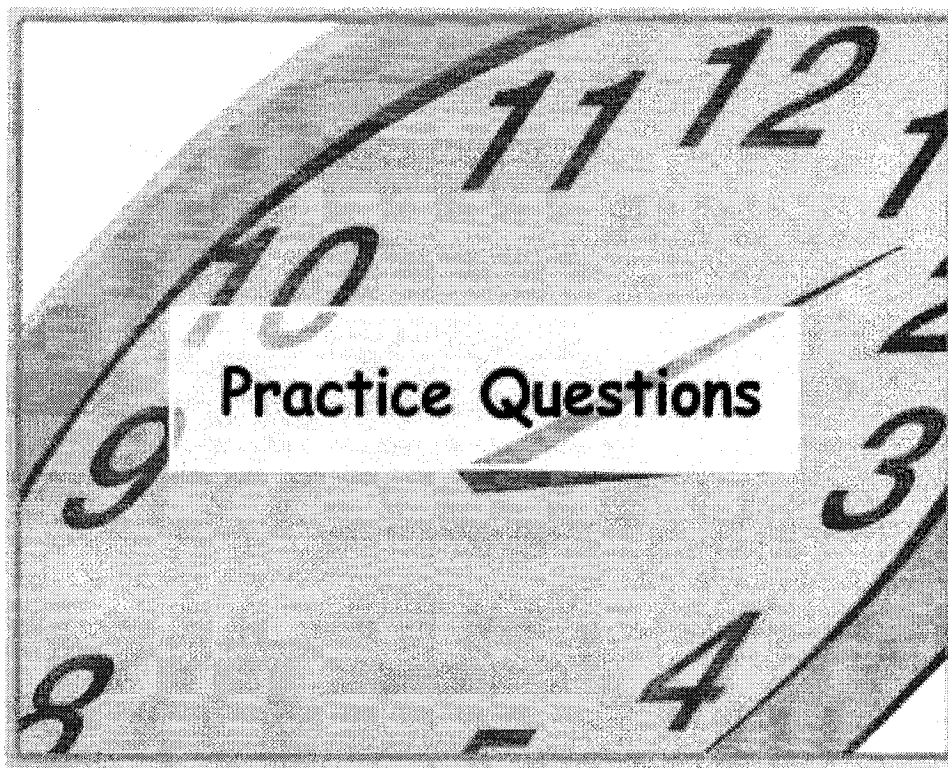
APPENDIX D: CATOE

(COMPUTER-ASSISTED TEST OF ORAL ENGLISH)

(Audio Prompt) Test directions. This test is designed to test your general spoken English proficiency. You will be asked questions in an interview format and your answers will be recorded. The test consists of eight questions and lasts approximately 20 minutes. It is recommended that you answer each question as completely as possible in the time allowed. The questions, and the time that you have to answer each one, will be shown on the computer screen. Your scores will be awarded based on your communicative ability in English. Be sure to speak loudly and clearly enough for the raters to hear you.

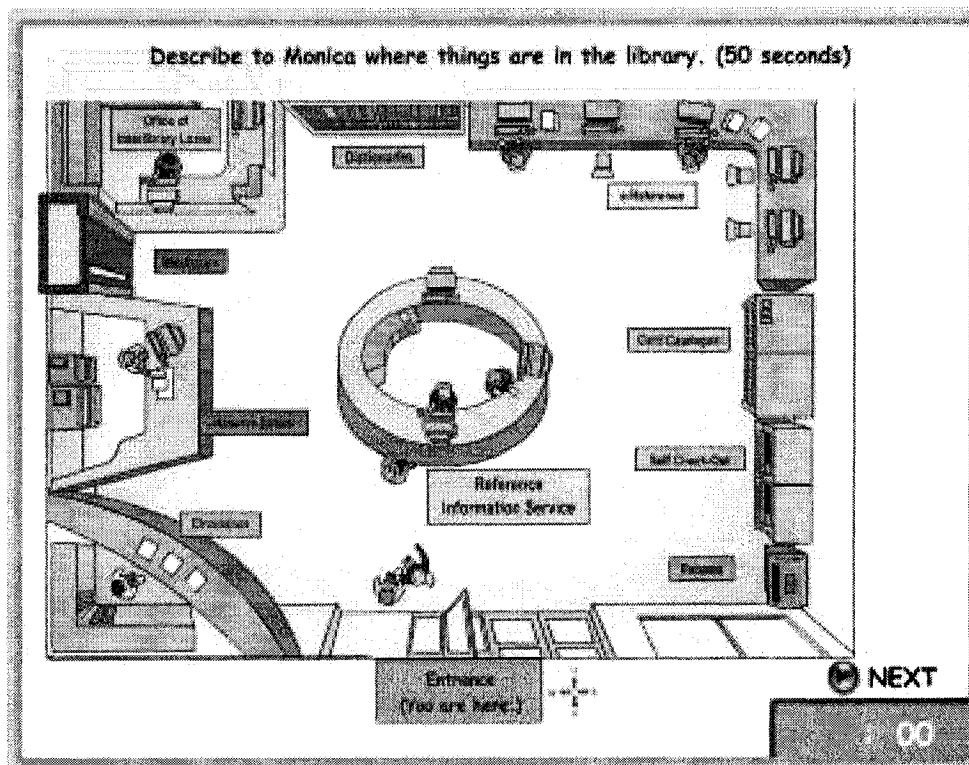


(Audio Prompt) Practice questions. You will be asked two practice questions. These questions are for practice only and will not affect your score, but you are encouraged to answer them.



(Audio Prompt) Practice questions.
How do you feel today? (10 seconds)
What are you studying? (10 seconds)
The test will now begin. Make sure to speak as clearly and completely as you can as you answer each question.

Task 1: (Audio Prompt) Suppose that you and Monica are friends, and Monica is going to visit your school library to borrow a book not available at her own school library. You would like to describe to Monica where things are in the library. You will have 20 seconds to look over the library map. Then, you will be asked to speak for 50 seconds.



Task 2: (Audio Prompt) After you describe the library, Monica says she wants to visit the library as often as possible. Based on the following information, explain to Monica the library services. You will have 1 minute to read the information and prepare your answer. Then, you will be asked to speak for 1 minute and 30 seconds.

5

Explain to Monica the library services. (1minute, 30seconds)

Library Service Information

Hours : 8:30 a.m. - 6:00 p.m.


Loan Periods : 2 weeks (undergraduates/external borrowers)
3 weeks (graduates/faculty)

Renewals* : 2 times (graduate/faculty/external borrowers)
3 times (undergraduates)

Returns : Circulation Desk (When library open)
Outside Returns Slot (When library closed)

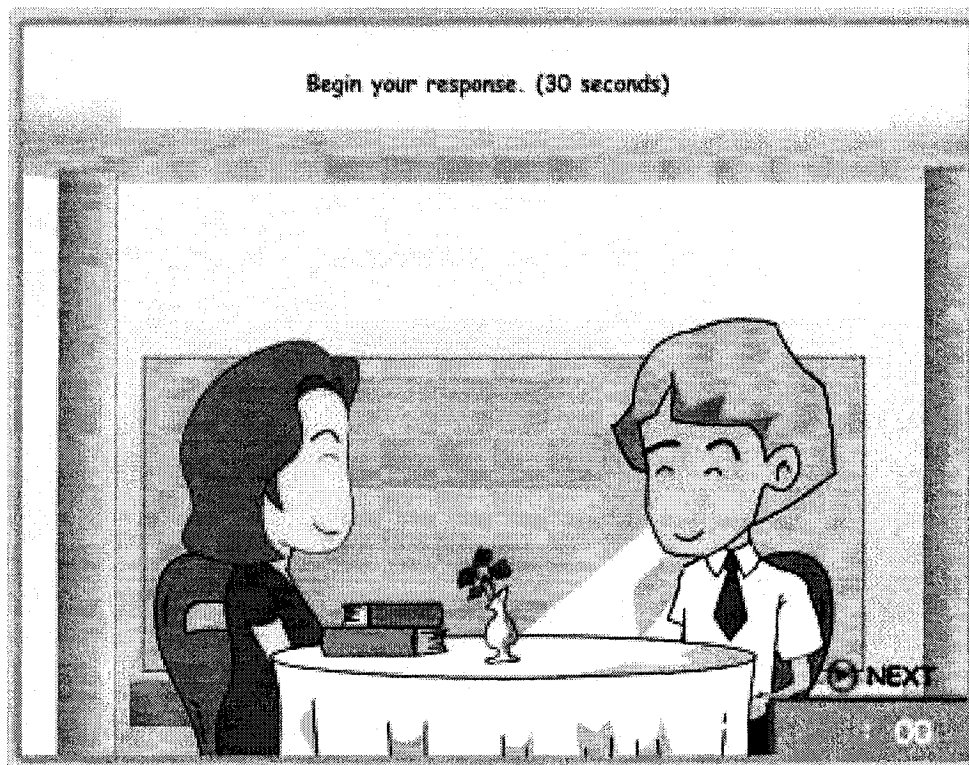
Fines : \$0.50/day (Books)
\$2.00/day (Periodicals)
\$2.50/day (Recalls)

* - If books are not required

 NEXT

00 : 00

Task 3: (Audio Prompt) After Monica has borrowed a book from the library, you go to the cafeteria together for lunch. Monica tells you that she has been accepted at a graduate school program that she really wants to attend. Knowing that Monica has worked very hard to be accepted into the program, you wish to congratulate her on her accomplishment. You will have 20 seconds to prepare what you will say to her. Then, you will be asked to speak for 30 seconds.



Task 4: (Audio Prompt) Look at the following six pictures. These pictures show what happened to John yesterday, beginning with picture one and continuing through picture six. I would like you to tell me the story shown in the six pictures. You will have 1 minute to study the pictures and prepare your answer. Then, you will be asked to speak for 1 minute and 30 seconds.

What happened to John yesterday? (1 minute, 30 seconds)

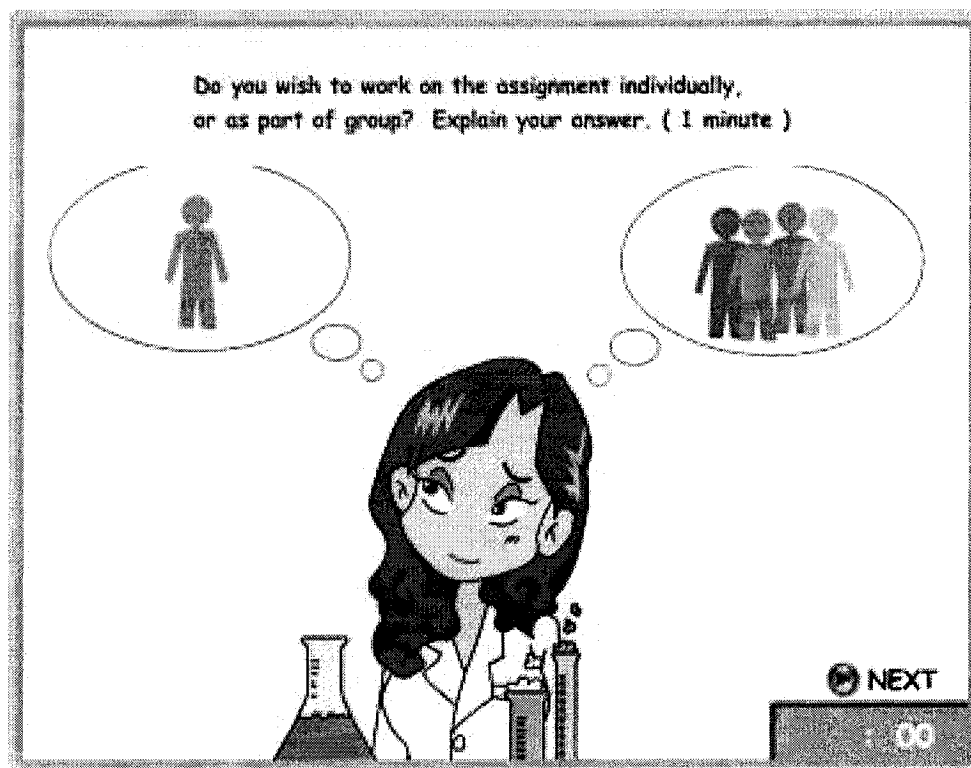
1 2 3

4 5 6

NEXT

00 : 00

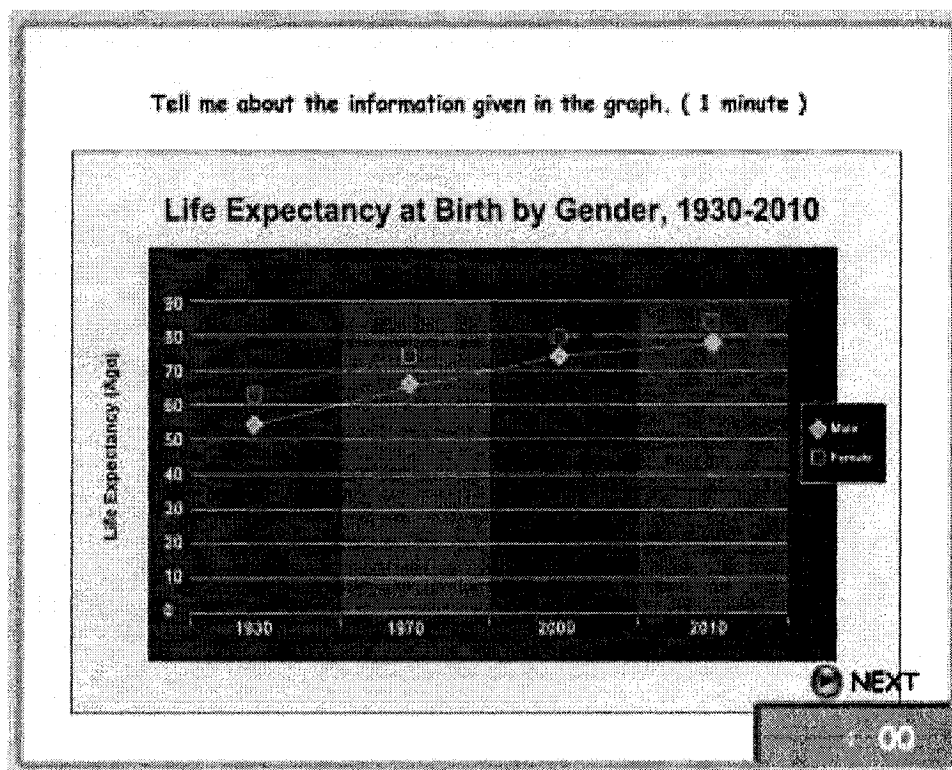
Task 5: (Audio Prompt) Imagine that you are taking a chemistry course. The instructor tells the class that students may complete next week's final laboratory assignment individually, or as part of a group. You must decide if you prefer to work individually or in a group. Your classmate, Monica, would like to know what you prefer and why. You will have 30 seconds to think about your answer. Then, you will be asked to speak for 1 minute.



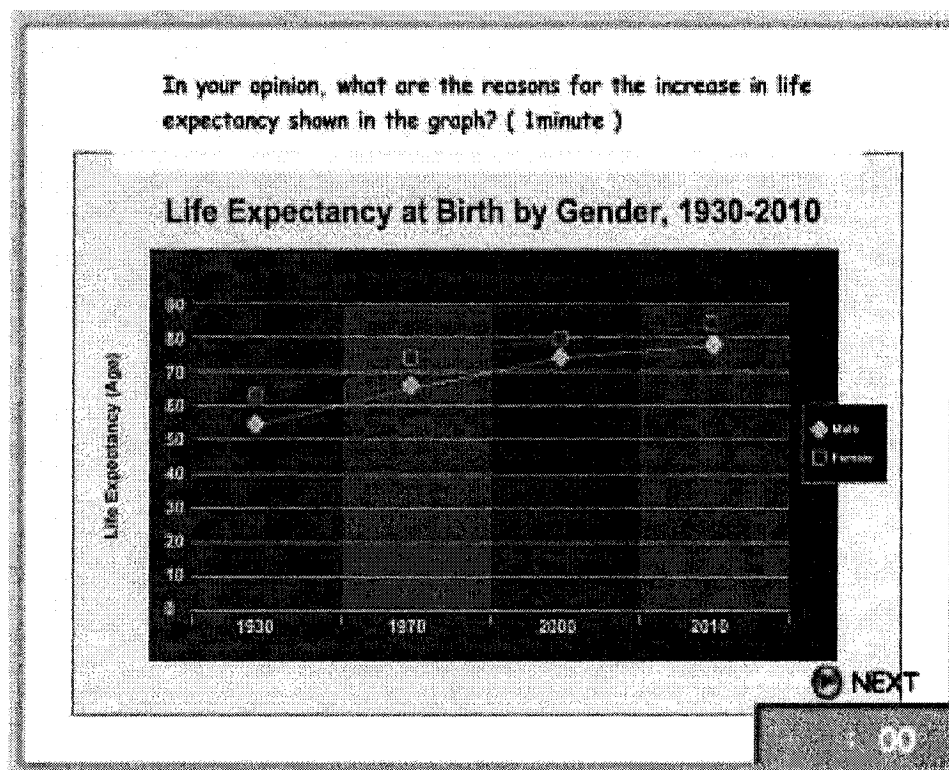
Task 6: (Audio Prompt) Imagine that you are taking a sociology course. In today's class, you are supposed to discuss the current dramatic increase in Internet use. The professor says that while the Internet is certainly a useful tool for accessing information, there are concerns that its use may have some harmful effects. He asks students to suggest what these harmful effects might be, and it is your turn to answer. You will have 30 seconds to think about your answer. Then, you will be asked to speak for 1 minute.



Task 7: (Audio Prompt) Imagine that you are attending a seminar about world population. The instructor shows a graph about life expectancy at birth by gender from 1930 to 2010. She would like you to describe the information given in the graph. You will have 30 seconds to look at the graph. Then, you will be asked to speak for 1 minute.



Task 8: (Audio Prompt) As the graph shows, life expectancy has increased over time. The instructor would like you to explain the reasons for this increase in life expectancy. You will have 30 seconds to think about your answer. Then, you will be asked to speak for 1 minute.



APPENDIX E: CATOE RATING SCALE*
(COMPUTER-ASSISTED TEST OF ORAL ENGLISH)

4

Overall communication is almost always successful; little or no listener effort is required.

3

Overall communication is generally successful; some listener effort is required.

2

Overall communication is less successful; more listener effort is required.

1

Overall communication is generally unsuccessful; a great deal of listener effort is required.

*

1. "Communication" is defined as both addressing a given task and getting a message across.
2. A score of 4 does not necessarily mean speech is comparable to that of native English speakers.
3. No response, or a response of "I don't know" is automatically rated *NR (Not Ratable)*.

APPENDIX F: CERTIFICATE OF ETHICAL ACCEPTABILITY

MCGILL UNIVERSITY
FACULTY OF EDUCATION

Received

CERTIFICATE OF ETHICAL ACCEPTABILITY FOR
FUNDED AND NON FUNDED RESEARCH INVOLVING HUMANS

NOV 11 2004

The Faculty of Education Ethics Review Board consists of 6 faculty members appointed by the Faculty of Education, an appointed member from the community, and the Chair of the Ethics Review Board.

The undersigned considered the application for certification of the ethical acceptability of the project entitled: An investigation into native and non-native teachers' judgments of oral English performance: Are they really different? as proposed by:

Applicant's Name Youn-Hee Kim

Supervisor's Name Carolyn E. Turner

Applicant's Signature/Date [Signature] Nov. 11, 2004

Supervisor's Signature [Signature] Nov 1, 2004

Degree / Program / Course M. A. in Second Language Education Granting Agency _____

Grant Title (s) _____

The application is considered to be:

A Full Review _____

An Expedited Review x

A Renewal for an Approved Project _____

A Departmental Level Review _____

Signature of Chair / Designate

The review committee considers the research procedures and practices as explained by the applicant in this application, to be acceptable on ethical grounds.

1. Prof. René Turcotte
Department of Kinesiology and Physical Education

4. Prof. Joan Russell
Department of Integrated Studies in Education

Signature / date

Signature / date

2. Prof. Ron Morris
Department of Integrated Studies in Education

5. Prof. Doreen Starke-Meyerring
Department of Integrated Studies in Education

Signature / date

Signature / date

3. Prof. Ron Stringer
Department of Educational and Counselling Psychology

6. Prof. Ada Sinacore
Department of Educational and Counselling Psychology

Signature / date

Signature / date

7. Member of the Community

Signature / date

Office of the Associate Dean (Research & Graduate Students)
Faculty of Education, Room 230
Tel: (514) 398-7039 Fax: (514) 398-1527

[Signature] Dec 1, 2004
Signature / date Chair of the Ethics Review Board

Office Use Only

REB #: 475-1104
(Updated September 2003)

APPROVAL PERIOD:

November 15 - December 1, 2004

APPENDIX G: FREQUENTLY ASKED QUESTIONS

1. Who are the ten students who took the CATOE?

- They are ten Korean students living in Montreal. Concerning their academic background or status, I cannot provide further information. This is because if you know their academic background, etc., your scoring may be influenced. If you could understand my efforts to minimize subject expectancy, it would be very much appreciated.

2. Do the four sample responses correspond to each level of the CATOE rating scale?

- No, as you may see in the handout that I passed out, these sample responses are only to familiarize teachers with the speech samples of the potential examinees. Depending on each teacher's personal judgment, the four sample responses may or may not correspond to each level of the CATOE rating scale.

3. Why is the CATOE rating scale so simple?

- The CATOE rating scale was developed to suit the unique purposes of my study. You may find that the rating scale plays a similar role to the Likert Scale, indicating a level of the general spoken English. The main reason that the scale is so simple is to derive the teachers' perceptions of the spoken English performance as much as possible as well as not to influence teachers' intact perceptions.

4. How can I score incomplete answers or irrelevant answers to the question?

- The rating scale does not address these cases. How to score these answers wholly depends on each teacher's personal decision.

5. How many comments should I make?

- You are encouraged to make comments as many as possible.

6. *When can I use NR (Not Ratable)?*

- Please ensure that NR is assigned to only the following two cases: no response, or a response of “I don’t know.”

7. *How many times can I listen to the speech samples?*

- There is no limitation when listening to the speech samples. You are allowed to listen as many times as you want.

APPENDIX H: CODING PROTOCOL OF COMMENTS

Major Categories & Definition	Sub-Categories	Examples of Comments
1. General Task Fulfillment: the degree to which the response fulfills the general demands of the task	Understanding the task	Didn't seem to understand the task. Didn't understand everything about the task.
	Overall task accomplishment	Generally accomplished the task. Task not really well accomplished. Successfully accomplished task.
2. Content Effectiveness: the degree to which the content of the response is of good quality and effectiveness in conveying an intended message	Strength/soundness of argument	Good range of points raised Good statement of main reason presented. Arguments quite strong
	Accuracy of transferred information	Some key information inaccurate Misinterpretation of information (e.g., graduate renewals for undergrads, \$50 a day for book overdue?) Incorrect information (e.g., "9pm" instead of "6pm")
	Topic relevance	Irrelevant content discussed. Not all points relevant Suddenly addressing irrelevant topic (i.e., focusing on physically harmful effects of laptops rather than on harmful effects of the internet)

3. Language Use: the degree to which language features of the response are of good quality and effectiveness in conveying an intended message	Overall language use	Generally good use of language Native-like language Very limited language
	Vocabulary	Limited vocabulary Good choice of vocabulary Some unusual vocabulary choices (e.g., he <i>crossed</i> a girl.)
	Pronunciation	Native-like pronunciation Pronunciation difficulty (e.g., l/r, d/t, vowels, i/e) Mispronunciation of some words (e.g., “circulation)
	Fluency	Choppy, halted Pausing, halting, stalling – periods of silence Smooth flow of speech
	Intelligibility	Hard to understand language (a great deal of listener work required) Almost always understandable language Almost impossible to understand any words
	Sentence structure	Cannot make complex sentences. Telegraphic speech Took risk with more complex sentence structure

3. Language Use (Continued)	General grammar use	Generally good grammar Some problems with grammar Few grammatical errors
	Specific grammar use	Omission of articles Incorrect or vague use of prepositions of place Good use of past progressive
4. Socio-contextual Appropriateness: the degree to which the response is appropriate and relevant to the intended communicative goals of a given situation	Socio-cultural appropriateness	Effectively communicate congratulations in a culturally appropriate manner. Cultural / pragmatic issue (a little formal to congratulate a friend) Little congratulations, more advice (culturally not appropriate)
	Contextual appropriateness	Appropriate language of a given situation Student response would have been appropriate if Monica had expressed worry about going to graduate school. Appropriate language for a decision-making situation
5. Organizational Development: the degree to which the response is developed and organized in a coherent and effective manner	Coherence	Good use of linking words Great time markers Organized answer
	Supplement of details	Provides enough details for effective explanation about the graph.

5. Organizational Development (Continued)	Supplement of details (Continued)	Student only made one general comment about the graph without referring specifics. Lacks enough information with logical explanation.
	Completeness of discourse	Incomplete speech No reference to conclusion End not finished.
	Elaboration of argument	Mentioned his arguments but did not explain it. Good elaboration of reasons Connect ideas smoothly by elaborating his arguments.

APPENDIX I
: TABLES OF CATOE SCALE CATEGORY STATISTICS AND
FIGURES OF CATOE SCALE STRUCTURES FOR TASKS 2 – 8

Table 1. CATOE Scale Category Statistics for Task 2

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-1.91	0.29
3	0.07	0.19
4	1.84	0.20



Figure 1. CATOE Scale Structure for Task 2

Table 2. CATOE Scale Category Statistics for Task 3

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-1.98	0.43
3	0.51	0.19
4	1.48	0.21



Figure 2. CATOE Scale Structure for Task 3

Table 3. CATOE Scale Category Statistics for Task 4

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-2.53	0.29
3	0.28	0.18
4	2.26	0.22

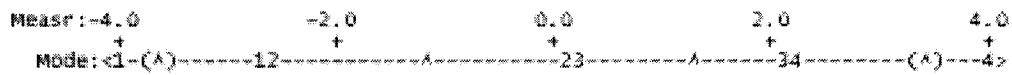


Figure 3. CATOE Scale Structure for Task 4

Table 4. CATOE Scale Category Statistics for Task 5

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-2.13	0.33
3	0.28	0.18
4	1.86	0.21

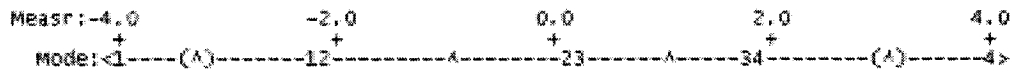


Figure 4. CATOE Scale Structure for Task 5

Table 5. CATOE Scale Category Statistics for Task 6

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-1.58	0.21
3	0.00	0.20
4	1.58	0.29



Figure 5. CATOE Scale Structure for Task 6

Table 6. CATOE Scale Category Statistics for Task 7

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-2.33	0.25
3	0.25	0.18
4	2.08	0.27

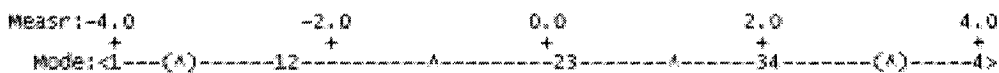


Figure 6. CATOE Scale Structure for Task 7

Table 7. CATOE Scale Category Statistics for Task 8

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-2.33	0.28
3	-0.34	0.17
4	2.67	0.28



Figure 7. CATOE Scale Structure for Task 8

APPENDIX J
: TABLES OF CATOE SCALE CATEGORY STATISTICS AND
FIGURES OF CATOE SCALE STRUCTURES
FOR SITUATION-BASED AND TOPIC-BASED TASKS

Table 1. CATOE Scale Category Statistics for Situation-Based Task

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-2.30	0.73
3	0.59	0.27
4	1.71	0.30

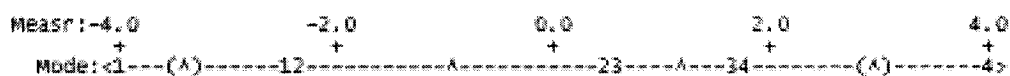


Figure 1. CATOE Scale Structure for Situation-Based Task

Table 2. CATOE Scale Category Statistics for Topic-Based Task

CATOE Scale Category	Step Calibrations	
	Measure (logits)	S.E.
1		
2	-1.94	0.20
3	0.02	0.15
4	1.92	0.20

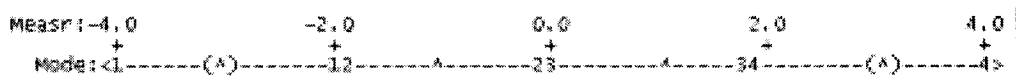


Figure 2. CATOE Scale Structure for Topic-Based Task

APPENDIX K
: TABLES OF CATOE SCALE CATEGORY STATISTICS AND
FIGURES OF CATOE SCALE STRUCTURES FOR TASKS 2 – 8
BY NS AND NNS GROUPS

Table 1. CATOE Scale Category Statistics for Task 2 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.23	0.41	-1.58	0.40
3	0.21	0.25	-0.10	0.27
4	2.02	0.30	1.69	0.27

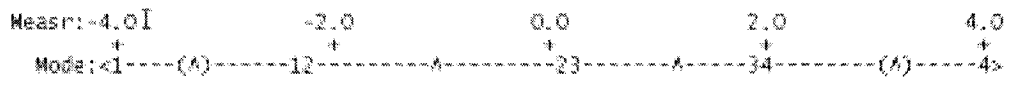


Figure 1. CATOE Scale Structure of NS Group for Task 2

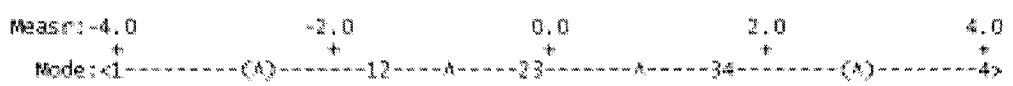


Figure 2. CATOE Scale Structure of NNS Group for Task 2

Table 2. CATOE Scale Category Statistics for Task 3 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.33	0.68	-1.72	0.57
3	0.59	0.26	0.46	0.28
4	1.75	0.31	1.26	0.28

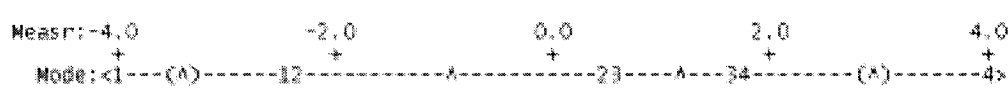


Figure 3. CATOE Scale Structure of NS Group for Task 3

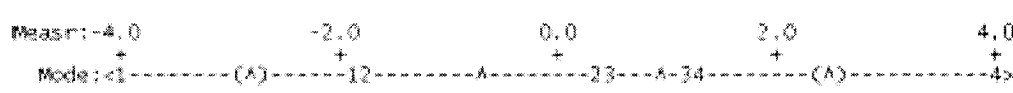


Figure 4. CATOE Scale Structure of NNS Group for Task 3

Table 3. CATOE Scale Category Statistics for Task 4 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.82	0.41	-2.29	0.40
3	0.20	0.25	0.35	0.25
4	2.63	0.34	1.94	0.30



Figure 5. CATOE Scale Structure of NS Group for Task 4



Figure 6. CATOE Scale Structure of NNS Group for Task 4

Table 4. CATOE Scale Category Statistics for Task 5 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-1.94	0.44	-2.35	0.50
3	0.30	0.26	0.26	0.25
4	1.64	0.29	2.09	0.30

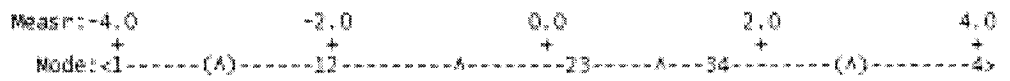


Figure 7. CATOE Scale Structure of NS Group for Task 5

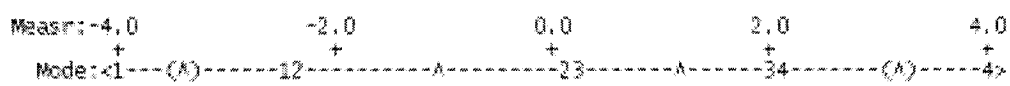


Figure 8. CATOE Scale Structure of NNS Group for Task 5

Table 5. CATOE Scale Category Statistics for Task 6 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-1.64	0.30	-1.55	0.30
3	0.03	0.29	-0.04	0.29
4	1.61	0.42	1.59	0.40

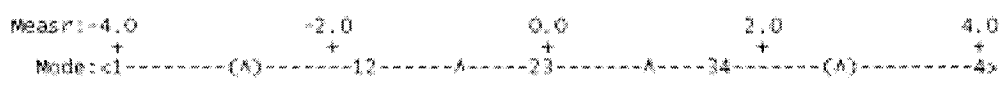


Figure 9. CATOE Scale Structure of NS Group for Task 6

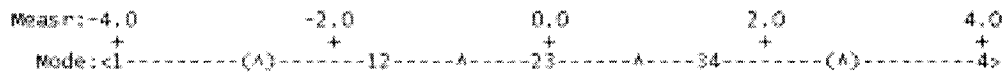


Figure 10. CATOE Scale Structure of NNS Group for Task 6

Table 6. CATOE Scale Category Statistics for Task 7 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.42	0.36	-2.27	0.34
3	0.06	0.25	0.45	0.25
4	2.36	0.39	1.81	0.37

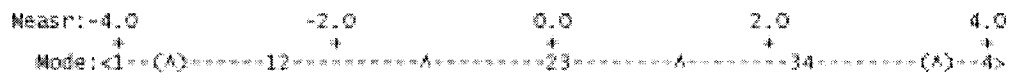


Figure 11. CATOE Scale Structure of NS Group for Task 7

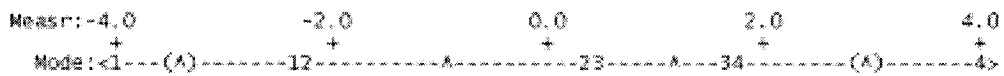


Figure 12. CATOE Scale Structure of NNS Group for Task 7

Table 7. CATOE Scale Category Statistics for Task 8 by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.38	0.39	-2.30	0.40
3	-0.17	0.25	-0.51	0.25
4	2.55	0.40	2.80	0.39

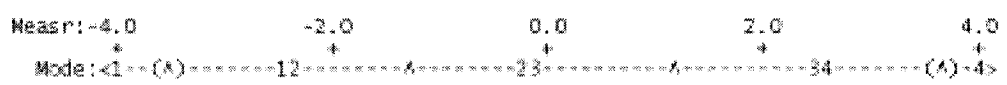


Figure 13. CATOE Scale Structure of NS Group for Task 8

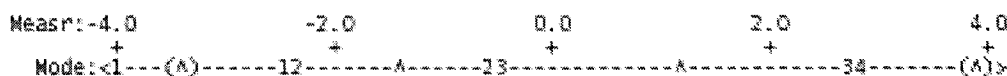


Figure 14. CATOE Scale Structure of NNS Group for Task 8

APPENDIX L
: TABLES OF CATOE SCALE CATEGORY STATISTICS AND
FIGURES OF CATOE SCALE STRUCTURES FOR SITUATION-
BASED AND TOPIC-BASED TASKS BY NS AND NNS GROUPS

Table 1. CATOE Scale Category Statistics for Situation-Based Task by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-2.31	0.73	-1.73	0.54
3	0.60	0.27	0.45	0.27
4	1.71	0.30	1.28	0.29



Figure 1. CATOE Scale Structure of NS Group for Situation-Based Task



Figure 2. CATOE Scale Structure of NNS Group for Situation-Based Task

Table 2. CATOE Scale Category Statistics for Topic-Based Task by NS and NNS Groups

Scale Category	Step Calibrations (NS)		Step Calibrations (NNS)	
	Measure (logits)	S.E.	Measure (logits)	S.E.
1				
2	-1.94	0.20	-1.88	0.21
3	0.04	0.15	-0.22	0.15
4	1.91	0.20	2.10	0.20

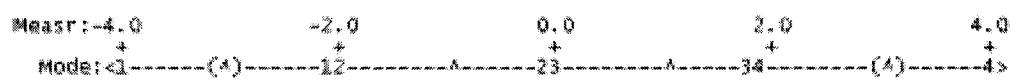


Figure 3. CATOE Scale Structure of NS Group for Topic-Based Task

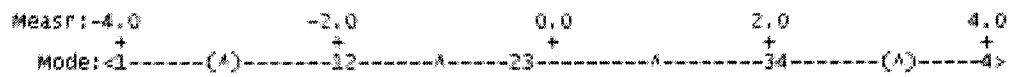


Figure 4. CATOE Scale Structure of NNS Group for Topic-Based Task

APPENDIX M
: TABLES OF NUMBER AND PERCENTAGE OF COMMENTS
FOR TASKS 2 – 8

Table 1. Number and Percentage of Comments for Task 2

	Number of Comments	Percentage of Comments
General Task Fulfillment	20	3.9%
Content Effectiveness	43	8.3%
Language Use	384	74.3%
Socio-contextual Appropriateness	0	0.0%
Organizational Development	70	13.5%
Total	517	100.0%

Table 2. Number and Percentage of Comments for Task 3

	Number of Comments	Percentage of Comments
General Task Fulfillment	11	4.0%
Content Effectiveness	0	0.0%
Language Use	149	54.8%
Socio-contextual Appropriateness	98	36.0%
Organizational Development	14	5.1%
Total	272	100.0%

Table 3. Number and Percentage of Comments for Task 4

	Number of Comments	Percentage of Comments
General Task Fulfillment	16	3.1%
Content Effectiveness	22	4.3%
Language Use	377	73.3%
Socio-contextual Appropriateness	0	0.0%
Organizational Development	99	19.3%
Total	514	100.0%

Table 4. Number and Percentage of Comments for Task 5

	Number of Comments	Percentage of Comments
General Task Fulfillment	13	3.2%
Content Effectiveness	68	16.7%
Language Use	254	62.3%
Socio-contextual Appropriateness	23	5.6%
Organizational Development	50	12.3%
Total	408	100.0%

Table 5. Number and Percentage of Comments for Task 6

	Number of Comments	Percentage of Comments
General Task Fulfillment	25	6.3%
Content Effectiveness	96	24.2%
Language Use	216	54.4%
Socio-contextual Appropriateness	0	0.0%
Organizational Development	60	15.1%
Total	397	100.0%

Table 6. Number and Percentage of Comments for Task 7

	Number of Comments	Percentage of Comments
General Task Fulfillment	9	2.5%
Content Effectiveness	48	13.2%
Language Use	225	61.6%
Socio-contextual Appropriateness	0	0.0%
Organizational Development	83	22.7%
Total	365	100.0%

Table 7. Number and Percentage of Comments for Task 8

	Number of Comments	Percentage of Comments
General Task Fulfillment	25	6.2%
Content Effectiveness	71	17.6%
Language Use	238	58.9%
Socio-contextual Appropriateness	0	0.0%
Organizational Development	70	17.3%
Total	404	100.0%

APPENDIX N
: TABLES OF NUMBER AND PERCENTAGE OF COMMENTS
FOR SITUATION-BASED AND TOPIC-BASED TASKS

Table 1. Number and Percentage of Comments for Situation-Based Task

	Number of Comments	Percentage of Comments
General Task Fulfillment	11	4.0%
Content Effectiveness	0	0.0%
Language Use	149	54.8%
Socio-contextual Appropriateness	98	36.0%
Organizational Development	14	5.1%
Total	272	100.0%

Table 2. Number and Percentage of Comments for Topic-Based Task

	Number of Comments	Percentage of Comments
General Task Fulfillment	63	5.2%
Content Effectiveness	235	19.4%
Language Use	708	58.6%
Socio-contextual Appropriateness	23	1.9%
Organizational Development	180	14.9%
Total	1209	100.0%

APPENDIX O
: TABLES OF NUMBER AND PERCENTAGE OF COMMENTS
FOR TASKS 2 – 8 BY NS AND NNS GROUPS

Table 1. Number and Percentage of Comments for Task 2 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	7	2.1%	13	7.1%
Content Effectiveness	21	6.3%	22	12.1%
Language Use	271	80.9%	113	62.1%
Socio-contextual Appropriateness	0	0.0%	0	0.0%
Organizational Development	36	10.7%	34	18.7%
Total	335	100.0%	182	100.0%
$\chi^2 (3, N = 517) = 23.69, p = 0.000 < 0.001$				

Table 2. Number and Percentage of Comments for Task 3 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	5	2.8%	6	6.5%
Content Effectiveness	0	0.0%	0	0.0%
Language Use	105	58.3%	44	47.8%
Socio-contextual Appropriateness	57	31.7%	41	44.6%
Organizational Development	13	7.2%	1	1.1%
Total	180	100.0%	92	100.0%
$\chi^2 (3, N = 272) = 10.60, p = 0.014 < 0.05$				

Table 3. Number and Percentage of Comments for Task 4 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS) (%)
General Task Fulfillment	12	3.6%	4	2.2%
Content Effectiveness	6	1.8%	16	8.7%
Language Use	248	75.2%	129	70.1%
Socio-contextual Appropriateness	0	0.0%	0	0.0%
Organizational Development	64	19.4%	35	19.0%
Total	330	100.0%	184	100.0%
$\chi^2(4, N = 514) = 14.28, p = 0.003 < 0.005$				

Table 4. Number and Percentage of Comments for Task 5 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	8	3.1%	5	3.4%
Content Effectiveness	55	21.1%	13	8.8%
Language Use	164	62.8%	90	61.2%
Socio-contextual Appropriateness	14	5.4%	9	6.1%
Organizational Development	20	7.7%	30	20.4%
Total	261	100.0%	147	100.0%
$\chi^2(4, N = 408) = 21.07, p = 0.000 < 0.001$				

Table 5. Number and Percentage of Comments for Task 6 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	20	8.5%	5	3.1%
Content Effectiveness	64	27.1%	32	19.9%
Language Use	126	53.4%	90	55.9%
Socio-contextual Appropriateness	0	0.0%	0	0.0%
Organizational Development	26	11.0%	34	21.1%
Total	236	100.0%	161	100.0%
$\chi^2 (3, N = 397) = 13.03, p = 0.005 < 0.01$				

Table 6. Number and Percentage of Comments for Task 7 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	5	2.1%	4	3.2%
Content Effectiveness	33	13.7%	15	12.1%
Language Use	156	64.7%	69	55.6%
Socio-contextual Appropriateness	0	0.0%	0	0.0%
Organizational Development	47	19.5%	36	29.0%
Total	241	100.0%	124	100.0%
$\chi^2 (3, N = 365) = 4.97, p = 0.174 > 0.05$				

Table 7. Number and Percentage of Comments for Task 8 by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	22	8.0%	3	2.3%
Content Effectiveness	49	17.8%	22	17.1%
Language Use	156	56.7%	82	63.6%
Socio-contextual Appropriateness	0	0.0%	0	0.0%
Organizational Development	48	17.5%	22	17.1%
Total	275	100.0%	129	100.0%
$\chi^2 (3, N = 404) = 5.30, p = 0.151 > 0.05$				

APPENDIX P
: TABLES OF NUMBER AND PERCENTAGE OF COMMENTS
FOR SITUATION-BASED AND TOPIC-BASED TASKS
BY NS AND NNS GROUPS

Table 1. Number and Percentage of Comments for Situation-Based Task by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	5	2.8%	6	6.5%
Content Effectiveness	0	0.0%	0	0.0%
Language Use	105	58.3%	44	47.8%
Socio-contextual Appropriateness	57	31.7%	41	44.6%
Organizational Development	13	7.2%	1	1.1%
Total	180	100.0%	92	100.0%
$\chi^2 (3, N = 272) = 10.60, p = 0.014 < 0.05$				

Table 2. Number and Percentage of Comments for Topic-Based Task by NS and NNS Groups

	Number of Comments (NS)	Percentage of Comments (NS)	Number of Comments (NNS)	Percentage of Comments (NNS)
General Task Fulfillment	50	6.5%	13	3.0%
Content Effectiveness	168	21.8%	67	15.3%
Language Use	446	57.8%	262	60.0%
Socio-contextual Appropriateness	14	1.8%	9	2.1%
Organizational Development	94	12.2%	86	19.7%
Total	772	100.0%	437	100.0%
$\chi^2 (4, N = 1209) = 23.37, p = 0.000 < 0.001$				