Title
Statistical Profiling of Academic Oral English Proficiency based on an ITA Screening Test

Author
Choi, Ick Kyu

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Statistical Profiling of Academic Oral English Proficiency based on an ITA Screening Test

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Applied Linguistics

by

Ick Kyu Choi

2013

ABSTRACT OF THE DISSERTATION

Statistical Profiling of Academic Oral English Proficiency
based on an ITA Screening Test

by

Ick Kyu Choi
Doctor of Philosophy in Applied Linguistics
University of California, Los Angeles, 2013
Professor Lyle F. Bachman, Chair

At the University of California, Los Angeles, the Test of Oral Proficiency (TOP), an internally developed oral proficiency test, is administered to international teaching assistant (ITA) candidates to ensure an appropriate level of academic oral English proficiency. Test taker performances are rated live by two raters according to four subscales. While the subscale scores have potential as valuable feedback to major stakeholders, only a weighted average of the four subscale scores are currently reported and used. This study presents a way of extracting valuable information from the TOP subscale scores. In particular, it investigates an approach to obtaining oral English proficiency profiles based on the subscale score patterns of 960 TOP test takers.

This study utilized item response theory and finite mixture modeling to investigate profiles of academic oral English proficiency in terms of the TOP subscales. A higher-order generalization of the graded response model was formulated to estimate subscale scores that accounted for

structural dependencies and rater effects in the observed TOP scores. The estimated scores were clustered using a multivariate normal mixture model to yield subscale score profiles. The mixture model suggested seven profile groups and classified the TOP test takers into the seven groups. The profile groups were then interpreted and labeled based on characteristic score patterns and linguistic background shared by group members.

To achieve a thorough understanding of the resulting profiles, discourse features of test taker performances sampled from different profile groups were closely examined. A small corpus was constructed based on the sampled test taker performances and compared to a reference corpus to explore the overall pattern of TOP test takers' language use. The comparison showed that the TOP test takers tended to use relatively fewer function words than speakers in the reference corpus. Characteristic features of each profile group's discourse were investigated through an identification and examination of discourse organizing lexical bundles. The results suggested that the use of metadiscourse and textual reference bundles with an explicit past reference point might be related to test takers' academic oral English proficiency.

The dissertation of Ick Kyu Choi is approved.

Li Cai

Susan J. Plann

Hongyin Tao

Lyle F. Bachman, Committee Chair

University of California, Los Angeles

2013

DEDICATION

To my parents, who have given me everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and committee chair, Dr. Lyle Bachman. Lyle has an uncanny ability to convert rough ideas into legitimate research questions through his insights, patience, and constructive criticism. He has been a constant source of inspiration and encouragement during my study at UCLA. I will always be grateful for his guidance and support.

I am extremely thankful to the members of my committee, Drs. Li Cai, Susan Plann, and Hongyin Tao. I am deeply indebted to Li for his untiring efforts to find ways to make this dissertation better. I thank Susan for her unwavering support and thought-provoking questions. I am grateful to Professor Tao for his detailed and practical suggestions.

I have been extremely fortunate to have wonderful teachers throughout my study. I would like to express my appreciation to Professor Sung-Ock Sohn, for her constant support and trust. I would also like to thank my professors at Seoul National University, Drs. Hynkee Ahn, Jin-Wan Kim, Byungmin Lee, Sun-Young Oh, Moonsu Shin, and Hyun-Kwon Yang. My special thanks go to Professor Oryang Kwon, who introduced me to the exciting field of language testing and encouraged me to become a researcher.

This dissertation would not have been possible without the support of the UCLA Office of Instructional Development. I would like to thank Drs. Kumiko Haas and Joanne Valli-Meredith, for being the best boss one can possibly hope. I have learned a lot and had enormous pleasure working for them.

I owe a debt of thanks to the KT&G Scholarship program for their generous financial support during my entire study at UCLA.

My four years at UCLA would have not been complete without my friends in the language assessment group, Hongwen Cai, Yujie Jia, Hsin-Min Liu, Jonathan Schmidgall, and Huan Wang. It was a great pleasure working with them. I am especially grateful to Dr. Youngsoon So who has been unbelievably generous and supportive even before I came here.

Lastly, I would like to thank my family. My parents have given me everything, and I am forever grateful for their love and support. My utmost appreciation goes to my dear loving wife Innhwa Park, who has always been there for me. I simply cannot thank you enough.

VITA

Degrees

2007 B.A. in English Education, Seoul National University

2009 M.A. in English Education, Seoul National University

Employment

2010-2011 Programmer/Analyst, Center for Public Health & Disasters, UCLA

2010-2013 Graduate Student Researcher, LAUSD Dual Language Test, UCLA

2011 Research Assistant, CTB/McGraw-Hill

2011-2013 Test Coordinator, Test of Oral Proficiency, UCLA

2011-2013 Graduate Student Researcher, Center for Educational Assessment, UCLA

Publications

Ockey, G., & Choi, I. (To Appear). Conducting structural equation model analyses and reporting results: Best practice guidelines for language assessment researchers. Language Testing.

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. Language Testing, 29(3), 603-619.

Baek, S., Park, H., Yoon, Y., Mo, K., Chung, S., Yoo, J., Kwon, O., Ham, E., Yoo, Y., Kil, H., Kim, S., Park, S., Ha, E., Lee, B., Kim, S., Kim, H., Choi, I., & Chang, G. (2008). Research on 2008 National Middle-school Subjects Learning Diagnostic Test. Seoul National University.

Presentations

Choi, I. (2013, July). Finite mixture models for extracting learner proficiency profiles based on analytically scored performance assessments. Paper to be presented at the 36th Language Testing Research Colloquium, Seoul, Korea.

Sohn, S., & Choi, I. (2013, July). Facilitating communication to stakeholders using statistical graphics. Paper to be presented at the 36th Language Testing Research Colloquium, Seoul, Korea.

Choi, I. (2013, May). Extending testlet models for language performance assessments. Paper presented at the 16th Southern California Association for Language Assessment Research, Fullerton, CA.

Choi, I. (2012, October). A multidimensional hierarchical rater model for language performance assessments. Paper presented at the 14th Midwest Association of Language Testers, Urbana–Champaign, IL.

Choi, I. (2012, May). Uses and misuses of specification search in applied linguistics studies. Paper presented at the 15th Southern California Association for Language Assessment Research, Los Angeles, CA.

Choi, I. (2012, April). Modeling the structure of passage-based tests: An application of a two-tier full information item factor analysis. Paper presented at the 35th Language Testing Research Colloquium, Princeton, NJ.

Choi, I.,& Schmidgall, J. (2011, June) A survey of methodological approaches employed to validate language assessments: 1999 – 2009. Paper presented at the 34th Language Testing Research Colloquium, Ann Arbor, MI.

Choi, I.,& Schmidgall, J. (2011, May). Frameworks for validity: A comparison of traditional and argument-based approaches for reviewing research. Paper presented at the 14th Southern California Association for Language Assessment Research, Los Angeles, CA.

Choi, I. (2011, March). Philosophical underpinnings and practical usefulness of dynamic and psychometric approaches to classroom assessment. Paper presented at the Annual Conference of American Association for Applied Linguistics, Chicago, IL.

Choi, I. (2010, October). Effects of item content characteristics on item difficulty of multiple choice test items in an EFL listening assessment. Paper presented at the East Coast Organization of Language Testers, Georgetown University, Washington DC.

Choi, I. (2008, June). The effects of the interlocutor on high proficiency level learners in foreign language oral proficiency tests. Paper presented at the Conference on the English Education of Korean Primary and Middle School Students in China, Yianji, China.

Chapter 1: The Problem

1.1 Background of the problem

As the proportion of international graduate students has grown in many U.S. universities, the number of international teaching assistants (ITAs) has also increased and become quite large (Constantinides, 1987; Rounds, 1987). This phenomenon has brought about predominantly negative reactions from major stakeholders, including American-born undergraduates and their parents, faculty members, and even ITAs themselves (e.g., Fox & Geneva, 1994; Mestenhauser, 1981; Nelson, 1991; Rubin & Smith, 1990).

Central to this negative reception has been the concern about the adequacy of ITAs' oral English proficiency (Bresnahan & Kim, 1993; Hendel et al., 1993; Hinofotis & Bailey, 1981). Communicating with undergraduate students is crucial to most, if not all, teaching assistant (TA) duties. Inadequate oral English proficiency could seriously undermine the quality of guidance and help undergraduate students may receive from their TAs. While it has been acknowledged that English oral proficiency should not be singled out as the sole concern for all ITAs (e.g., Bauer, 1991; Nelson, 1990; Thomas & Monoson, 1991), it still lies at the heart of the issues surrounding ITA (e.g., Gallego, 1990; Hendel et al., 1993; Hinofotis & Bailey, 1981). Thus, English oral proficiency poses a crucial barrier for ITAs trying to serve in their full capacity and for undergraduates seeking to receive the education that they deserve.

Therefore, universities with a large number of ITAs have institutionalized procedures to assure that ITAs possess a minimum level English oral proficiency. Two most commonly adopted procedures include training and screening. English as a Second Language (ESL) classes have assumed the training role in many institutions (Bauer & Tanner, 1994), while mandatory

oral proficiency tests have been put into practice to screen out international graduate students who do not meet the required English proficiency standards (Briggs, 1994; Johncock, 1991).

1.2 Statement of problem situation

1.2.1 Lack of high-quality feedback from ITA testing

In general, ESL classes last for a short period and the number of students per class is larger than ideal (e.g., Kaplan, 1989; Locastro, 2001; Warriner, 2007). Such limitations make it difficult for instructors to provide learners with individual feedback that would help them identify areas of weakness. It has been acknowledged that while detailed linguistic feedback is desirable for effective second/foreign language training (e.g., Davis & Tyler, 1994; Shohamy, 1992), it is difficult for language learners to receive such feedback. In fact, the lack of high quality linguistic feedback has been one of the biggest challenges in ESL teaching (e.g., Vardi, 2009). Furthermore, the lack of such feedback is not limited to ESL classrooms. Researchers have noticed that native speaker peers tend not to provide linguistic feedback to international graduate students even in a context in which one evaluates others' oral presentations (Jacoby & McNamara, 1996).

At the same time, information gathered from ITA screening tests is often not utilized as effectively as it could be. In particular, this information has rarely been employed as feedback. A number of universities have adopted an English oral proficiency test as a screening device. A prospective ITA is asked to produce a sample of spoken English which is then rated in several different aspects, such as grammatical accuracy and pronunciation, by multiple trained raters (Briggs, 1994). However, it is typical that only the final score, usually in the form of a weighted composite of different subscale scores, is considered in the pass-fail decision and reported to the

test takers. Other information including the subscale scores is disclosed to test takers who make an additional request for all their scores or is used only for internal review purposes.

1.2.2 Test of Oral Proficiency: Setting of the test program under study

At the University of California, Los Angeles (UCLA), an internally developed oral proficiency test, called the Test of Oral Proficiency (TOP), has been administered for the purpose of securing minimum oral English proficiency for ITAs since 2004. The TOP consists of two scored tasks: a syllabus presentation task and a simulated lecture task. Two trained undergraduate students are present as questioners in test rooms to ask questions of and to interact with test takers. Two trained raters are also present in the test rooms, but do not interact with test takers; they focus on rating and documenting evidence that could justify their ratings. Both tasks are rated live by the two raters according to four subscales: Pronunciation, Lexical Grammar, Rhetorical Organization, and Question Handling. A pass-fail decision is based on a weighted average of the four subscale scores, namely the total score. There are three decision categories: fail, provisional pass, and pass. In order for an international graduate student to start working as a TA without any restriction, he or she needs to pass the test. Test takers whose scores correspond to the provisional pass decision are allowed to work as a TA under the condition that they take a remedial ESL class of their choosing. When a test taker fails the test, he or she cannot start working as a TA until the TOP is re-taken, and the test taker (provisionally) passes the test.

While the test yields scores for all four subscales, only the total score and the corresponding decision category are reported to test takers. Although it is possible for a test taker to receive the subscale score information by arranging a test review session with a TOP coordinator, the proportion of the test takers who opt to arrange a review session is less than 20 percent of all test

takers, indicating that most test takers are not benefiting from more information that is available to them. At the same time, practical concerns prohibit the coordinators from providing the review session to every individual test taker. For instance, while there are approximately 100 test takers per quarter, a typical review session takes from 30 minutes to one hour. Considering other duties of the coordinators, such as training raters and questioners, monitoring raters' and questioners' performances, reporting test results to test takers and their departments, and examining and monitoring psychometric qualities of the test, they would not have enough time to provide the review session to every test taker, should all test takers request the service.

1.3 Purpose of the study

It is unfortunate that the information resulting from the TOP has not been fully utilized to provide feedback to test takers on their academic oral proficiency. The TOP subscale scores capture different aspects of test takers' academic oral proficiency, and are provided by trained raters with documented justification. These scores thus have considerable potential to provide valuable feedback to both TOP test takers and ESL instructors. However, most test takers are not currently benefiting from this additional information, and thus, it is desirable to have a means to provide the additional information to the TOP test takers.

This study investigates a way of extracting information from the TOP subscale scores. In particular, it attempts to produce oral English proficiency profiles of TOP test takers based on their subscale score patterns. The score patterns will be clustered and interpreted to yield meaningful profiles of academic oral English proficiency. To achieve a thorough understanding of the resulting profiles, discourse features of test taker performances sampled from different profile groups will be examined and compared.

1.4 Research questions

An investigation of TOP subscale score profiles can yield a broad picture of test takers' performance in terms of the four subscales. The total score currently reported to the test takers does not provide any information for each subscale. The weighted-average method of computing the total score implies a compensatory scoring scheme; the same total score can be obtained from a number of different subscale score combinations. However, subscale score patterns might indicate that some test takers would benefit from more specific information to guide their learning. For example, when a test taker receives high scores for Pronunciation and Lexical Grammar, but low scores for Question Handling, it might be more beneficial for the test taker to focus on comprehension and interactive aspects of oral proficiency, rather than individual phonemes or grammar patterns. Therefore, the subscale score profile can provide useful information that is not available in the current score report.

Another utilization of the subscale score profile involves an examination of discourse features in test taker performances sampled from different subscale score profiles groups. Characteristics of ITA language use have been studied based on a relatively small sample; comparisons between ITAs with different levels of oral proficiency in terms of their characteristic language use features have often relied upon holistic judgment of the participants' proficiency or standardized test scores (e.g., Robinson, 1993; Tyler, 1992; Williams, 1992). The TOP subscale score profile will provide a more grounded sampling scheme for describing and comparing characteristic features of language use by ITAs from different levels of English oral proficiency. In addition, if distinct characteristics of language use features are observed across

different profile groups, those characteristics can also be provided as linguistic feedback to test takers.

With a focus on the potential benefits of the TOP subscale score profile, this study attempts to obtain the profiles of the TOP test takers based on their subscale scores and to examine features of language use by the members of different profile groups. In particular, the following research questions will be investigated:

1. What different subscale score profiles of academic English oral proficiency can be meaningfully produced by the TOP?
2. What features of oral discourse characterize test takers who have different subscale score profiles?

1.5 Assumptions

This study assumes that the TOP subscale scores are valid indicators of academic oral English proficiency. That is, this study pursues the goal of creating subscale score profiles under the assumption that the four TOP subscales validly measure prospective ITAs' English academic oral proficiency. It is beyond the scope of this study to investigate the quality of the four subscales as indicators of academic oral English proficiency of prospective ITAs. Furthermore, it will not be claimed that the four subscales form an exhaustive set of indicators of English academic oral proficiency.

Another assumption this study makes is that the TOP test taker population consists of heterogeneous subpopulations that cannot be observed directly. While there are several background variables that divide the total test taker population into different groups, including

gender, first language (L1), and academic major, these background variables do not provide direct information about potentially different subscale score profiles of TOP test takers. This study assumes that a more direct grouping of TOP test takers in terms of the subscale score patterns cannot be made using the existing observed grouping variables, but can be estimated based upon the data.

1.6 Delimitations of the study

In essence, this study is an exploratory attempt to extract information from the available sources of the TOP data and use the extracted information to provide test takers with detailed feedback on their oral English proficiency. It is not claimed that the results of this study will generalize beyond the local setting of the TOP. In fact, any attempts to generalize to other oral tests or other populations of test takers based on this study, should such attempts be made, must be preceded by a rigorous cross-validation and thorough consideration of the context of both this study and the target domain of the generalization.

This study does not involve an empirical investigation of the effectiveness of the resulting profiles as feedback. While it will be argued that the additional information included in the resulting profiles has a great potential to make effective feedback, whether it would function effectively as feedback for ITAs and their potential ESL instructors is beyond the scope of this study. The effectiveness of feedback depends on the interaction among a number of external variables, including learner attitudes and motivation (Rea-Dickins, 2007). Whether the profile information resulting from this study could be put into use or how effective this would be in practice are both valid and interesting research questions, but will not be pursued in this study.

Lastly, this study, in its use of statistical models, embraces the view that statistical models are never true (Box, 1979). Consequently, this study will make no attempt to confirm or prove a certain proposition based on the results. This study aims to provide useful and efficient descriptions and explanations of the available data with statistical models.

1.7 Importance of the study

1.7.1 Theoretical importance

This study attempts to produce academic English oral proficiency profiles in terms of the TOP subscale scores. While the generalizability of the resulting profiles to other contexts will not be investigated, the profiles based on the local setting will contribute to an understanding of the multifaceted construct of academic oral proficiency. Through the exploration of different subscale score profiles, this study will present a structure of English academic oral proficiency emerged from a local test setting. Furthermore, the results of this study could provide insights into the relatively less studied discipline of academic spoken language (Bowles, 2006; Hyland, 2002) through a close investigation of test taker performances on the TOP.

This study examines the possibility of using finite mixture (FM) models for analyzing performance assessment data to provide detailed feedback to test takers, as an alternative to cognitive diagnostic assessment (CDA) models. While CDA models are widely used to yield detailed skill profiles of test takers, their rather strict identification conditions render them difficult to adopt in most performance assessment context. If FM models, which are the main statistical methodology of this study, turn out to be capable of producing meaningful profiles, such models will provide applied linguists and language testing specialists who hope to provide

detailed feedback based on language test data from a performance assessment with another useful methodological tool.

1.7.2 Practical importance

The resulting subscale score profiles can provide additional information beyond what is currently reported to test takers. The additional information can be readily put to use to facilitate teaching and learning of academic English and ITA training. The ability of the TOP to provide detailed feedback will bring about positive impact, and therefore add additional evidence for the use of the test.

In addition, this study explores a means of extracting additional information based on the readily available results from the TOP administrations using statistical clustering techniques and discourse analytic methods. In other words, it does not necessitate additional data collection which often requires time and other resources. If successful, therefore, the methods employed in this study can be easily adopted to other contexts to yield similar profiles of test takers.

Finally, this study will provide empirical evidence of how the Rhetorical Organization subscale has been interpreted by raters. Little empirical evidence has been available about the characteristic features of Rhetorical Organization. Therefore, raters have relied upon intuitions and expectations rather than empirical description of test taker practices in assigning Rhetorical Organization scores. A concrete description of linguistic characteristics of test taker performances, which will be produced as the result of this study, could help improve the quality of rubric description, rater training, and rating itself (Biber, Conrad, Reppen, Byrd, & Helt, 2002).

Chapter 2: Review of Literature

2.1 The ITA problem

In the 1980's, the increasing numbers of ITAs stimulated discussions about roles of, and expectations and misconceptions towards ITAs. Bailey (1982) recognized this complex issue and labeled it the "ITA problem". Based on a comprehensive description of the difficulties ITAs faced and the discords among major stakeholders, she claimed that serious studies of the ITA problem should be conducted. Bailey's suggestion resonated among researchers in the field of higher education and applied linguistics. A number of studies and presentations which focused on the ITA problem followed, resulting in a series of conferences and books (Bailey, Pialorsi, & Zukowski/Faust, 1984; Madden & Myers, 1994; Nyquist, Abbott, Wulff, & Sprague, 1991).

Researchers have conceptualized oral English language proficiency required in the U.S. higher education context as a multi-componential construct (e.g., Bauer, 1991; Hoekje & Williams, 1992; Nelson, 1990; Robertson, 1983; Seo, 1989). Over the years, a list of components pointed out by researchers has grown; this includes American culture in general (Nelson, 1990; Thomas & Monoson, 1991), expected classroom behavior (Bauer, 1991; Thomas & Monoson, 1993; Young, 1989), and teaching skills (Thomas & Monoson, 1993; Young, 1989). However, researchers in general have been in agreement that inadequate English oral proficiency lies at the heart of the ITA problem. English oral proficiency has been the primary concern in the ITA problem of researchers and practitioners alike (e.g., Ard, 1987; Bauer, 1991, 1992; Smith, Byrd, Nelson, Barret, & Constantinides, 1992), and the inadequacy of English oral proficiency has been found as one of the most crucial barriers for ITAs (Jacobs & Friedman, 1988; Williams, Barnes, Finger, & Ruffin, 1987).

While the specific roles of ITA vary across institutions and academic fields, early surveys of ITA roles have identified the core tasks ITAs are expected to perform: providing undergraduate students with teaching, guidance, and help (Fox & Geneva, 1994; Williams et al., 1987). Considering the central role of communication in such core tasks, concerns about ITAs' inadequate oral English proficiency have arisen (e.g., Bailey, 1982, 1984; Feetham, 1988; Thomas & Monoson, 1993). If ITAs are incapable of providing undergraduate students with help and guidance due to inadequate oral English proficiency, ITAs should be trained to overcome the language hurdle (Smith et al., 1992; Thomas & Monoson, 1993). It has been argued that ITAs should be able to meet a minimum standard of English proficiency required to perform TA duties (Dick & Robinson, 1994; Jacobs & Friedman, 1988; Thomas & Monoson, 1991, 1993). Higher education institutions responded by establishing procedures that could address ITAs' communication problem, the most popular of which include ITA training programs and ITA screening tests (e.g., Bauer & Tanner, 1994; Kaplan, 1989; Smith et al., 1992).

2.2 ITA training and screening

A large number of ITA training programs, often consisting of ESL classes for (prospective) ITAs, began with an emphasis on pronunciation, especially on accent reduction at the individual phoneme level (Bauer & Tanner, 1994; Kaplan, 1989; Smith et al., 1992). As researchers have realized that accent reduction at the segmental level alone could not guarantee satisfying results (e.g., Derwing, 2008), ITA training programs have also diversified to include other components of oral English proficiency. For example, Anderson-Hsieh (1990) has argued that explicit instructions about suprasegmental features should be included in ITA training programs. Other components of oral proficiency beyond pronunciation have also garnered interest, including

vocabulary use (e.g., Rubin, 1993; Taner, Selfe, & Wiegand, 1993; vom Saal, Miles, & McGraw, 1988) and pragmatic knowledge (e.g., Davies, Tyler, & Koran, 2002; Jenkins, 1997), leading to a variety of aspects of oral English proficiency that have been added to the curriculum. Furthermore, ITA training programs have expanded to include non-linguistic components, such as American cultural knowledge (e.g., Althen, 1991) and familiarity with the daily lives of typical undergraduate students (e.g., Jia & Bergerson, 2008).

However, several issues have been raised with regard to the effectiveness of ITA training programs. A critical issue has been that trainees are not given enough attention or practice time due to large class sizes (e.g., Green, Christopher, & Lam, 1997; Warriner, 2007). The duration of ITA training programs has also been pointed out as a potential problem. Researchers have emphasized that second language learning and development is a time-consuming process (e.g., Ellis, 1994; Gass & Selinker, 2001), while ESL programs often do not last longer than a semester or a quarter (Bauer & Tanner, 1994; Kaplan, 1989). This has yielded a side effect that has been noticed by several researchers. Specifically, some ITA training programs appear to have a tendency to teach ITAs what can be easily taught, rather than what should be taught (Halleck & Moder, 1995).

In addition to the ITA training component, a number of universities have administered a screening test for prospective ITAs (Bauer & Tanner, 1994; Kaplan, 1989; Smith et al., 1992). Most early ITA screening tests took the form of a standardized English proficiency test, which sometimes also functioned as a gatekeeper for all incoming international students (e.g., Dunn & Constantinides, 1991; Johncock, 1991). However, researchers have argued that these standardized test kits might not suit the ITA screening purpose properly. It has been pointed out

that the standardized tests were not flexible enough to be optimized for each institution's needs (e.g., Plakans & Abraham, 1990) and did not engage academic oral English proficiency of (prospective) ITAs to a desired degree (e.g., Bier & Friedman, 1982). As these drawbacks have been recognized, institutions have begun to develop their own ITA screening test (see, e.g., Briggs, 1994). These internally developed ITA tests often take the form of oral English proficiency test and involve a simulation of a lecture or classroom discussion (Bailey, 1985; Briggs, 1994; Plakans & Abraham, 1990). Surveys of the internally developed ITA tests have also shown that they have typically adopted scoring procedures that utilize multiple raters and analytic scoring methods (Bauer & Tanner, 1994; Johncock, 1991). Researchers who have surveyed ITA screening tests are generally in agreement that the tests have positively contributed to achieving minimum oral English proficiency of ITAs (Briggs, 1994; Johncock, 1991).

However, several concerns have been raised about ITA screening tests including construct under-representation (e.g., Fox & Geneva, 1994; Thomas & Monoson, 1993; Plough, Briggs, & Bonn, 2010). In addition, results from ITA screening tests including analytic scores have rarely been put into use for the ITA training purpose. This is an unfortunate practice considering the high potential of the resulting information from ITA screening tests; the test results are typically obtained from trained raters under a highly focused situation (see, e.g., Kaplan, 1989; Smith et al., 1992). ITA screening tests provide an advantageous context in terms of generating rich information. This clear potential of ITA testing results has recently begun to attract the interests of researchers. For example, Schmidgall (2012a) described test takers' experience with receiving feedback and their attitude towards receiving detailed feedback based on an ITA assessment, which turned out to be largely positive.

2.3 Profiles of ITAs' English oral proficiency

The idea of utilizing test results for improving learning is certainly not new. The formative assessment movement has arisen focusing on the very same idea. The main tenet of formative assessment centers on the idea that test results could, and perhaps should, function as feedback to improve learning (see, e.g., Black & Wiliam, 1998 for a comprehensive review). The increasing interest in diagnostic assessment (see, e.g., Bejar, 1984) is also closely related to the idea of utilizing test results to enhance learning processes. If the main objective of an assessment is to provide diagnostic information about test takers' current mastery of skills, abilities or knowledge components, the assessment can be viewed as a diagnostic assessment (Bachman & Palmer, 2010). The focus of diagnostic assessment has been identifying different profiles of skill mastery observed across different learners (Bejar, 1984). With the skill mastery information, the users of diagnostic assessment aim to provide more contextualized and beneficial feedback and input to learners.

The important first step of extracting profile information from test results is to decide on the components of the profile. In the context of ITA screening tests, the decision could be made based upon theoretical and empirical models of second language oral proficiency in an academic context. Different models of second language oral proficiency exist (e.g., Bachman and Palmer, 2010; Foster, Tonkyn, & Wigglesworth, 2000; Iwashita, Brown, McNamara, & O'Hagan, 2008; Poehner, 2008), based on different foci or perspectives, or sometimes reflecting different philosophical backgrounds. However, they all appear to agree on one issue: second language oral proficiency consists of multiple components. While not all of the suggested models explicitly discuss the relationships among multiple components of language proficiency, it is reasonable to

conjecture that different learners possess different profiles in terms of different language proficiency components. Anecdotal evidence of different profiles of language learners abounds; famous examples include Henry Kissinger and Joseph Conrad, who, despite their accented pronunciation, showed a mastery of the oral and written grammar of English, respectively. Indeed, different profiles of language learners have attracted the interests of a number of researchers (e.g., Gibbons, 1984; Kagan, 2005; Weissberg, 2000)

Pronunciation has been one of the most popular research areas when it comes to different language learner profiles. Over the years, a number of studies have reported cases of language learners who were deemed proficient in many other aspects of target language including mastery of oral and written grammar, but had distinct accents and therefore suffered from compromised intelligibility (e.g., Han & Odlin, 2006; Oyama, 1976; Schwartz, 1997). Those pronunciation effects have typically been studied under the framework of fossilization or stabilization, as well as under the critical period hypothesis (see Han, 2004; Han & Odlin, 2006, for a review). Grammar has been another area which language learners have found difficult to master even after achieving proficiency in other aspects of the target language (e.g., Patkowsky, 1980). In particular, there has been intensive research about which areas of grammar are difficult to acquire for learners with a specific L1. Some researchers have argued that, based on the results from so-called "morpheme order" studies (e.g., Dulay & Burt, 1973, 1974, 1975), there exists a specific set of developmental stages, often defined in terms of the acquisition of specific morphological/syntactical elements, applicable to all learners of the same target language (Pienemann, 1998; Zobl & Liceras, 1994). However, this claim is not without criticism (see Doughty & Long, 2003; Ellis, 1994; Hudson, 1993, for a review).

While pronunciation and grammar have been the most productive research areas that focus on different components of second language oral proficiency, there are other areas that have attracted researchers' attention. Highly relevant to the ITA context includes what Bachman and Palmer (2010) call textual knowledge, which has to do with "producing or comprehending the sequence of units of information in text" (p. 45). Researchers have found a strong positive relationship between the comprehensibility of academic lectures and the use of organization markers, lexical ties, and cohesiveness (e.g., Douglas & Myers, 1989; Flowerdrew & Tauroza, 1995; Rounds, 1987; Tyler, 1992). The lack of proper organization in ITA speech could impede undergraduate students' understanding of substantive materials (Duerksen, 1994; Tyler, 1992; Williams, 1994). Furthermore, the distinction between comprehension and production is also relevant to the ITA context. For example, one's use of communication strategies could compensate for his lack of comprehension (Douglas & Myers, 1989). This discrepancy could be addressed by real-time communications such as the question and answer structure of classrooms. Indeed, studies have shown that ITAs could lack proper understanding of the questions, and that this could lead to misunderstanding undergraduates' linguistic and non-linguistic signals, which would ultimately result in frustration for undergraduates (e.g., Chiang, 2009).

Taken together, research investigating several different components of language proficiency strongly suggests that ITAs could possess different profiles of oral English proficiency in terms of pronunciation, grammar, organization, and question handling. Some ITAs could be regarded as proficient in grammar and organization but lacking intelligibility. Others might have proper organization in their speech but suffer from comprehension problems that could impede communication. These different profiles of ITAs, then, could play an important role in deciding how to improve their oral English proficiency. For example, if one is good in pronunciation but

not in organization, taking an ESL class focusing on phoneme distinctions in the target language would not be effective; on the other hand, if one's heavy accent is the primary concern, learning about the target culture would not be the best use of his or her time. When it comes to ITAs, the efficiency of instruction and training is an important concern given the already heavy burden a graduate student carries (Hoekje & Williams, 1992; Sequeira & Costantino, 1989). Therefore, specifying different profiles of language proficiency would improve the effectiveness and efficiency of language learning and teaching.

2.4 Statistical approaches to obtain profile information

2.4.1 Cognitive diagnostic assessment models

Acknowledging the potential of language proficiency profile information for enhancing language learning, numerous researchers have made attempts to extract skill profiles from test results (e.g., Jang, 2009; Kim, 2011; Lesaux & Kieffer, 2010). In this regard, the formative and diagnostic assessment movements have provided a substantive framework to utilize test results for learning purposes (e.g., Black, Harrison, Lee, Marshall, & Wiliam, 2003; Sadler, 1989). On the other hand, psychometric models have provided methodological tools to extract language skill profiles based on information obtained from test administrations. A popular psychometric model for this approach is a class of restricted latent class models, often called cognitive diagnostic assessment (CDA) models (Leighton & Griel, 2007). Combining a content analysis of test items and an investigation of test takers' response patterns, CDA models yield a set of skill profiles that provides probabilistic judgments regarding the mastery of each skill measured by the test.

The goal of the content analysis stage of CDA is to yield a zero-one matrix called a Q matrix (Tatstuoka, 1983), which specifies the relationships between test items and skills. The item-skill mapping in the Q matrix functions as the model specification which informs the subsequent decomposition of test takers' response patterns to estimate item parameters and yield each test taker's skill mastery information. A variety of CDA models have been proposed depending on specific model parameterizations and suggested response processes (DiBello, Roussos, & Stout, 2007). Detailed discussions of individual CDA models are beyond the scope of this study. Rupp and Templin (2008) give a comprehensive description and review of popular CDA models. Comprehensive coverage of CDA models can also be found in DiBello et al. (2007) and Leighton and Griel (2007). In the language testing context, a special issue of Language Assessment Quarterly (Lee & Sawaki, 2009) covered CDA extensively, focusing on the theory and use of CDA models in the second language assessment context. Overall, CDA models have provided useful methodological tools for language testing researchers and practitioners who aim to yield skill profile information based on test results.

However, CDA models are not without limitations. The most relevant limitation in the context of this study involves the models' identification condition. In particular, CDA models require that each skill has at least one item that measures only that specific skill in order to empirically distinguish different skills (Rupp & Templin, 2008). This identification condition would not pose much difficulty when a test consists of a number of items but is designed to engage a relatively small number of skills. For example, if a test contains 50 items and involves five skills, the above identification condition could easily be met. However, the identification condition is not normally satisfied in most performance assessment situations. In general, language performance assessments consist of a relatively small number of tasks (or items) that

are supposed to engage a number of different skills (Luoma, 2004; Wiegle, 2002). In addition, skills required for each item are often difficult to specify, for there could exist different combinations of different skills that enable one to perform well on an item. Under these circumstances, it becomes extremely difficult to satisfy the identification condition of CDA models. The paucity of applications of CDA models in the performance assessment context can be accounted for, at least partly, by this identification condition.

Another limitation involves CDA models' heavy reliance upon content analyses. CDA models yield model-based estimates and skill profiles. However, if the model is not properly specified, the validity of the resulting estimates and profiles would be compromised. When a test is originally designed for diagnostic purposes, that is, when a test is developed under the CDA framework, the skills involved in the test can be identified in a straightforward manner. However, not all applications of CDA have been based upon tests designed to yield diagnostic information; CDA models have been adopted to investigate existing tests, which were often not developed for diagnostic purposes (e.g., Jang, 2005; Kim, 2011). As a result, researchers have relied upon reverse-engineered content analyses to create the Q matrix. Under those circumstances, at least partial misspecifications of the model are suspected, and results should be interpreted with caution. In general, researchers advise against the use of CDA models for making high-stakes diagnostic decisions based on a test not developed for diagnostic purposes. (Rupp & Templin, 2008).

2.4.2 Finite mixture models

CDA models belong to a restrictive class of latent class models in the sense that the model is specified based on the Q matrix (Rupp & Templin, 2008). A more general family of models

which subsumes latent class models are finite mixture (FM) models. FM models assume that the data under study consist of heterogeneous groups from different population distributions (McLachlan & Peel, 2000). The main goal of FM models is to estimate the proportion of each group (called the mixing proportion) in the data as well as parameters of each group's population distribution (called the component distribution). For example, if one suspects that the data consist of a mixture of two normal distributions with different means and variances, FM models can be applied to estimate the mean and variance for each component distribution and the mixing proportion of the two distributions in the data. Therefore, free parameters of general FM models include the parameters of each component distribution and the mixing proportion of each component distribution, with an obvious restriction that the mixing proportions should add up to unity.

FM models share similar objectives with traditional cluster analysis techniques when the main goal is to decompose the data into clusters of homogeneous subgroups. In this light, FM models can be viewed as a type of cluster analysis (Everitt, Landau, Leese, & Stahl, 2011). A crucial feature of FM models that distinguishes them from traditional cluster analysis models is their parametric nature; FM models assume that each component distribution follows a known mathematical distribution, which puts them in the category of parametric models. On the other hand, traditional cluster analysis techniques often do not impose a parametric form for each cluster and rely upon distance functions to identify clusters (For a detailed account of clustering techniques, see Everitt et al., 2011).

The idea of finding heterogeneous distributions in data has been around for a long time. A well-known early application of a FM model goes back more than a century to Pearson (1894).

Using the method of moments, Pearson had to perform daunting computations to obtain his solution. For many years, this computational burden had prevented researchers from employing FM models (McLachlan & Peel, 2000). The introduction of computers and competent algorithms such as expectation-maximization (EM) (Dempster, Laird, & Rubin, 1977) and Markov Chain Monte Carlo (MCMC), largely freed researchers from the heavy computational burden related to FM models (McLachlan & Peel, 2000). Currently, FM models are widely used in a variety of disciplines including medicine (e.g., Schlattmann, 2009), marketing (e.g., Wedel & Desarbo, 2002), and political science (e.g., Hill & Kiresi, 2001).

It is possible to expand FM models by imposing further structures to component distributions. A mixture of generalized linear models has been popular in social science applications (McLachlan & Peel, 2000). Psychometricians have actively implemented factor analysis models into mixture component distributions. Yung (1997) has suggested a mixture of confirmatory factor analysis models, and Muthén (2001) has argued that FM modeling is a potent methodology to model atypical distributions and to deal with cases in which one has substantive reasons to believe that the data consist of heterogeneous groups. Muthén and Asparouhov (2006) have successfully applied a mixture of factor analysis models to address addictive behaviors. Another productive area of FM models has been a mixture of IRT models. Rost (1997) has shown that the mixture of logistic regressions can be easily estimated, and Cho, Cohen and Kim (2013) present how the model parameters could be estimated using a full Bayesian approach via MCMC. However, to date, few applications of FM models in applied linguistics or language testing research exist, which is unfortunate considering that profiles of language learners have been the interest of a number of applied linguistics and language testing researchers.

2.5 Corpus linguistics and analysis of ITA oral proficiency

2.5.1 Real language use and corpus linguistics

Corpus linguistics is based on the notion that intuitions about language use, which has been the main instrument of traditional theoretical linguistics, is not free from fallacy (Biber, Conrad, & Reppen, 1998). Everyone has a unique style of language use to a certain degree (Halliday, 1975) and these idiosyncrasies could affect how one views and regards language use. Consequently, studies of language use based solely on intuition might lead to inaccurate interpretations and conclusions. This limitation can be mitigated by investigating a large collection of instances of language use (i.e., corpus). For many years, technical problems and practicality issues were the main obstacles for building corpora that could be used for research purposes. However, powerful computers and cheap storage have diminished the hurdles these obstacles presented. With the advent of large corpora, corpus linguistics has been widely utilized by applied linguists who want to construct or test hypotheses involving real language use (Biber et al., 1998). Corpus linguistics has provided empirical evidence for claims about language use and has become a very productive tool.

An important advantage of corpus linguistics involves the capability of handling real language use data at various levels (McEnery, Xiao, & Tono, 2006; Stubbs, 1996). Corpus linguistics provides a systematic way of analyzing language use data in a quantitative way. Qualitative linguistic methodologies such as discourse analysis and conversation analysis have been invaluable tools for many linguists who need to investigate rich data obtained from a relatively small sample of participants. However, these qualitative methods are of limited use in aggregating individual evidence to capture important patterns underlying the phenomena under

study at a large scale. Corpus linguistics provides researchers with a way of aggregating qualitative descriptions to allow more general claims (Biber et al., 1998; McEnery et al., 2006). However, corpus linguistics is not limited to a large scale analysis; small scale corpora designed to address specific research objectives have been developed and provided promising results (e.g., Aguilar, 2004; Farr, 2003; Rounds, 1987). While the use of a small corpus inevitably constrains generalizability, such a corpus can still be highly useful in finding patterns that are not clear at the individual sentence or text level.

2.5.2 Spoken and learner corpora

Most corpus studies over the last several decades have focused on written texts (Biber, Conrad, Reppen, Byrd, & Helt, 2002), which, at least partly, can be explained by the limited availability of oral language corpora. Building an oral language corpus is clearly more difficult than building a corpus consisting of texts. Recently, however, several corpora consisting of oral language data have been compiled (e.g., Michigan Corpus of Academic Spoken English; Corpus of Contemporary American English). Those oral language corpora have already begun to attract the interest of applied linguists who have long desired to analyze spoken language use (e.g., Csomay, 2007; Fortanet, 2004). Analyses of oral language corpora have provided important observations about spoken language use (e.g., Leech, 2002) as well as about differences between spoken and written language use (e.g., Biber, 1986; Leech, Rayson, & Wilson, 2001).

Another area of research to which corpus linguistics has recently made an influential contribution is the target language use of nonnative speakers (NNS). In the early days of corpus linguistics, the majority of available corpora consisted of language use data obtained from native speakers (NS) (Granger, 2003a). NS corpora could be helpful for language learning purposes in

an indirect fashion; for example, they could help develop teaching materials, provide grounded descriptions of language use, and validate grammatical rules. However, researchers have pointed out that direct investigations of learner language use could benefit from corpora consisting of learner language data (Gabrielatos, 2005; Granger, 2003b; Nesselhauf, 2004). Several corpora containing learners' target language use, sometimes exclusively, have recently been made available (see Pravec, 2002, for a survey of learner corpora). These learner corpora have been effective in understanding how learners use the target language and whether and how their language use is different from that of NSs (e.g., Gabrielatos, 2005; Granger, 2003a, 2003b).

2.5.3 Analysis of ITA language use based on corpus linguistics

Corpus linguistics has been an invaluable tool in studying language use in the academic context. A number of researchers have investigated characteristics of academic language use based on academic corpora (see, e.g., Biber et al., 2002; Csomay, 2007; Hyland & Tse, 2004; Reppen, 2004). ITA language use is no exception; analyses based on ITA spoken corpora have been utilized to examine differences between language use of ITAs and native speaker TAs (Reinhardt, 2010) and to describe discourse features of ITA language use (Liao, 2009), to name a few. However, considering that ITA corpora consist of spoken language obtained from learners of English, both of which are relatively new to corpus studies, there still remain areas of research that could benefit from more such studies. The above corpus-based studies of ITA language use have shown that corpora could be very useful in studying ITA language use. Nevertheless, they have been relatively narrow in terms of their research focus in that they have investigated a specific, sometimes isolated feature of ITA language use. At the other extreme, there have been very broad and general approaches such as Biber's (1998) multidimensional analysis. Using a set

of dimensions identified from an exploratory factor analysis based on a large corpus, Biber and his colleagues analyzed other corpora to investigate empirical differences across different discourse genres (e.g., Biber et al., 2002). While this type of approach could provide helpful insights at a very general level, its unit of analysis might not be appropriate when there are potentially heterogeneous subgroups within the same discourse genre, which is the case of this study.

In order to properly address the language use of prospective ITAs consisting of heterogeneous subgroups, it is desirable to base the investigation upon a representative set of characteristic language use in each subgroup. Furthermore, using an extensive set of discourse features at once could lead to an overly complex picture of ITA language use, leading to a model that is not generalizable. There have been few studies that investigate the language use in the ITA testing context. A systematic description of ITA test taker performances in terms of characteristic discourse features could lead to a better understanding of ITA language use.

2.5.4 Lexical bundles and their discourse organizing functions

Prefabricated multiword expressions, often called lexical bundles, have attracted interests of applied linguists as an essential feature of language use (e.g., DeCarrico & Nattinger, 1988; Nattinger & DeCarrico, 1992). The recent advent of large corpora has led to empirical findings that attest to the frequent occurrences of such lexical bundles in both oral and written language (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999). Lexical bundles have been studied under many different names, including lexical phrases (e.g., Nattinger & DeCarriro, 1992) and formulaic language (e.g., Ellis, Simpson-Vlach, & Maynard, 2008), each of which is associated with a slightly different definition. However, despite such differences, they all share a common

core, which can be defined as "sequences of word forms that commonly go together in natural discourse" (Biber et al., 1999, p. 990).

Even before large corpora became widely available to enhance the identification of lexical bundles in various real language use settings, researchers acknowledged their important role in organizing academic discourse, especially in academic contexts. Chaudron and Richards (1986) investigated the impact of macro-organizers, which are essentially lexical bundles that signal important marks of given discourse, on the comprehensibility of a lecture. The availability of a large quantity of lexical bundles identified in academic corpora has led to a systematic synthesis of pragmatic functions served by lexical bundles. Biber, Conrad, and Cortes (2004) present a functional taxonomy of lexical bundles in a university setting and classify lexical bundles into three categories: stance expressions, referential expressions, and discourse organizers. Their classification results show that discourse organizers are most frequently observed in lectures. Nesi and Basturkmen (2009) use Biber et al.'s (2004) taxonomy to investigate how the frequent use of discourse organizing lexical bundles in university lectures contributes to overall cohesion. Nesi and Basturkmen suggest that discourse organizing lexical bundles in academic lectures serve as a facilitating mechanism that reduces the information processing loads of students.

Simpson-Vlach and Ellis (2010) extend Biber et al.'s (2004) taxonomy based on the findings from a large academic corpus. In Simpson-Vlach and Ellis' categorization, the function of discourse organizers can be further divided into four subcategories: metadiscourse and textual reference, topic introduction and focus, topic elaboration, and discourse markers. Out of the four subgroups, they find lexical bundles that function as metadiscourse and textual references (MTR) are genre-specific, in that frequently occurring MTR bundles in lectures do not overlap with

frequently occurring MTR bundles in textbooks. Considering this result, MTR bundles in lectures can be regarded as having a unique position among lexical bundles in that they constitute a characteristic feature of academic oral language use as a part of discourse organizers while retaining formal distinctiveness from their counterparts that normally occur in written texts. This unique position of MTR bundles in academic lectures makes them a natural contender for an investigation of the relationship between discourse organization and academic oral English proficiency of NNS lecturers.

Chapter 3: Methodology

This chapter discusses analytic procedures employed to address the research questions. It begins with an overview of the analysis plan, which describes the combination of different analytic approaches used for each research question. It then describes the dataset that was analyzed, including the participants-test takers and raters-who generated the data. Finally, the analytic procedures that were used for the first and the second research questions are described in detail.

3.1 Overview

To answer the research questions raised in the first chapter, this study utilized item response theory (IRT)-based score estimation, statistical clustering techniques, and discourse/corpus analytic methods. Major sources of data included the TOP subscale scores and videotaped performance of a subset of the test takers who were selected as representative samples of each subscale score profile group. The sampling scheme and the detailed description of the data for each stage are presented in the following sections.

This study consisted of two data analysis stages, each of which corresponds to one of the two research questions. In the first stage, the TOP subscale scores of the test takers were estimated using an IRT model. This estimation process accounted for structural dependency and rater effects contained within the observed rater scores. The resulting estimates will be referred to as the model-based subscale scores in the remainder of this study. Finite mixture (FM) models were fitted to the model-based subscale scores to explore groups of TOP test takers who exhibited homogeneous subscale score profiles. Based on the FM modeling result, each test taker was

assigned a profile group membership. The resulting profile groups were interpreted and labeled based on the shared score patterns of the group members.

The second stage focused on features of language use that characterize the test performances of the TOP test takers who belong to different score profile groups. Videotaped performances of a sample of test takers from each of the profile groups identified in the FM modeling were transcribed and analyzed to examine characteristic language use features. The transcripts constituted a small oral corpus of the test taker performances. The resulting corpus was then compared to a reference corpus to examine salient features in the overall language use of the TOP test takers. In addition, the use of discourse organizing lexical bundles in the corpus was investigated to explore relationships between discourse organizers and subscale score profiles.

3.2 Stage 1: TOP subscale score profiles

3.2.1 Objectives

The goal of the subscale score profile analysis was threefold. First and foremost, it aimed to find distinct groups of TOP test takers who share similar patterns of TOP subscale scores. Based on the resulting groups, it attempted to develop a suitable interpretation and label for each group and to assign group membership probabilities to the test takers. The resulting groups were also to be used as grouping variables for the second analysis stage.

3.2.2 Data

The data analyzed in this stage included the observed TOP subscale scores and the model-based subscale scores. The latter was estimated using an IRT model that can account for structural dependencies and bias arising from the TOP scoring procedure involving multiple

tasks and raters. The IRT model utilized in this stage will be described in detail in the next section. All subscale scores share the same range, from one to four. Since there are two raters who give a score for each of the four subscales for each of the two tasks, each test taker receives a total of 16 (4 subscales × 2 tasks × 2 raters) scores. However, the 16 scores do not provide independent information; a test taker receives four scores (2 tasks × 2 raters) per each subscale. In this study, the four scores within a subscale were weighted and combined to produce the model-based subscale scores. Therefore, each test taker was given a model-based score for each subscale, resulting in four model-based subscale scores per test taker. The model-based subscale scores were used as inputs for the FM analysis.

From the introduction of the TOP in 2004 to the end of academic year 2011-2012, a total of 2,450 tests were administered. Out of these, 361 were repeaters who took the test more than once because they did not "pass" in the first test. Of these 361 repeaters, only the scores from the first attempt were used for the analysis. While this reduced the total sample size, including multiple data from the repeaters would have created additional dependencies among data points. Furthermore, since all other test takers' scores were obtained from their first attempt, it was reasonable to retain the results from only the first attempt of the repeaters. An additional filtering process was used to obtain stable estimates of the model-based subscale scores.

There are 56 raters who have rated at least once since the introduction of the TOP. Based on their productivity (i.e., the number of test takers he or she rated), 18 out of the total 56 raters were selected. The remaining 38 raters were excluded due to the small number of test takers they have rated. In addition, different pairings of the remaining 18 productive raters were considered. Only the test takers who were assigned to rater pairs who rated at least five test takers were

included for the analysis. This process resulted in a total of 960 test takers who were analyzed in this stage.

3.2.2.1 Test taker language background

TOP test takers are asked to provide several pieces of background information at registration, including their first language (L1). In the dataset used for this study, there were a total of 53 different L1s. These L1s were grouped into nine different language groups based on linguistic commonalities among them. Table 3-1 provides the numbers and proportions of TOP test takers in each group, and the individual languages included in each group.

Table 3-1

Proportions of Major Language Groups in the TOP Dataset

Language Group	Percent Prop. (Raw Count)	Included Languages
Arabic (ARA)	5. 73 (55)	Arabic, Farsi, Persian
Chinese (CHI)	40.10 (385)	Cantonese, Mandarin Chinese
East Asian (EAL)	16.98 (163)	Japanese, Korean
English (ENG)	2.92 (28)	English
Germanic (GER)	2.71 (26)	Danish, Dutch, Finnish, German, Swedish
Indian (IND)	8.44 (81)	Bengali, Gujarati, Hindi, Kannada, Marathi, Oriya, Tamil
Romance (ROM)	11.04 (106)	French, Italian, Portuguese, Spanish, Romanian
Southeast Asian (SEA)	2.50 (24)	Filipino, Malay, Thai, Vietnamese
Slavic (SLA)	4.06 (39)	Armenian, Bulgarian, Czech, Macedonian, Polish, Russian, Serbo-Croatian, Slovenian, Ukrainian

Table 3-1 shows that the proportions of the language groups differed widely, with Mandarin Chinese being the L1 of the largest number of test takers, accounting for approximately 40 percent (372 out of 960). Other L1s with large populations in the TOP dataset included Spanish, Korean, Portuguese, and languages spoken in India. Thirty-nine out of the 53 L1s belonged to one of the nine groups. Speakers of languages belonging to the nine major language groups

comprised the vast majority (907 out of 960) of the entire sample. The remaining 14 languages had only few speakers each in the TOP dataset and these languages did not share enough commonalities with the languages included in the nine groups. The Chinese languages, including Cantonese and Mandarin Chinese, had the largest number of speakers, accounting for approximately 40 percent of the entire sample. East Asian languages, namely Japanese and Korean, comprised the second largest group, followed by Romance languages. There were a small number of English speakers. This was due to the definition of an international student at UCLA, which includes students from other English speaking countries such as Australia, Canada, and United Kingdom. The university policy mandates that every international student must pass the TOP before he or she starts working as a TA, regardless of L1. In addition, test takers who listed English as their second language were also counted as English speakers.

3.2.2.2 Test taker decision categories

As described in Chapter 1, TOP test takers are assigned one of three decision categories, namely Pass, Provisional Pass, and Fail. The TOP decision categories of the 960 test takers are given in Table 3-2.

Table 3-2

Proportions of TOP Decision Categories in the TOP Dataset

	Pass	Provisional Pass	Fail
Proportion (%)	63.64	20.94	15.42

Table 3-2 shows that approximately 65 percent of the entire sample passed the test, while approximately 15 percent of all test takers failed the test. The remaining test takers belonged to the Provisional Pass category, which mandates them to take a remedial ESL course.

3.2.2.3 Model-based subscale scores as the data for the mixture analysis

The unit of analysis for the FM model was the model-based subscale scores. The model-based subscale scores were entries in a data matrix, **X**, which has test takers as rows and the subscales as columns and was the input to the FM modeling. Therefore, an element x_{ih} of **X** represented *i*th test taker's model-based score on *h*th subscale, where $i = 1,..., 960$ and $h = 1,..., 4$. The structure of **X** is given in Table 3-2.

Table 3-3

The Structure of the Data Matrix

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Test Taker #1	x_{11}	x_{12}	x_{13}	x_{14}
Test Taker #2	x_{21}	x_{22}	x_{23}	x_{24}
…	…	…	…	…
Test Taker #960	$x_{960,1}$	$x_{960,2}$	$x_{960,3}$	$x_{960,4}$

3.2.3 Motivation for an IRT model to obtain the model-based scores

The TOP employs a fully crossed analytic scoring scheme with two tasks and two raters. A scoring procedure involving multiple raters and tasks is common in language proficiency performance assessments (Luoma, 2004; Weigle, 2002), but poses a challenge to psychometric modeling because of its complex data structure. In particular, residual dependencies are expected among scores produced by the same rater and among scores based on the same task. It is well known that bias in point estimates and their standard errors are introduced when underlying dependencies are ignored in model-based scoring and inferences (e.g., Bock, Brennan, & Muraki, 2002).

Another issue relevant to the TOP scoring procedure involves differences between raters. As noted before, each test taker is rated by two randomly assigned raters. This leads one to suspect a possible rater effect due to potential differences in raters' severity as well as their discriminating competency. For example, if some raters tend to be harsher than other raters in the TOP rater pool, it is reasonable to expect that, on average, test takers rated by those harsh raters would receive lower scores than they would have received had they been rated by other raters.

Given the incomplete block design of the TOP rater assignment, it was not possible to directly compare raters' severity and discrimination. In particular, each test taker in the dataset was rated by two raters, and scores from 16 out of 18 raters were missing. This makes the proportion of missing data in the TOP dataset almost as high as 90 percent (16 out of 18). Had all 18 rater scores been available, it would have been straightforward to evaluate raters' severity and discrimination through a direct comparison. An indirect and crude approximation to the direct comparison under the lack of that complete data would be to investigate univariate frequencies of scores given by each rater. Figure 3-1 provides the univariate frequencies of Pronunciation scores from all 18 raters in this study. Scores from the two tasks were summed for a clear presentation. Since each task is rated on a four-point scale staring from one, the summed scores were distributed on a six point scale from two to eight. Figure 3-1 shows differences among the raters' score frequencies. Similar differences were observed in the other three TOP subscale scores.

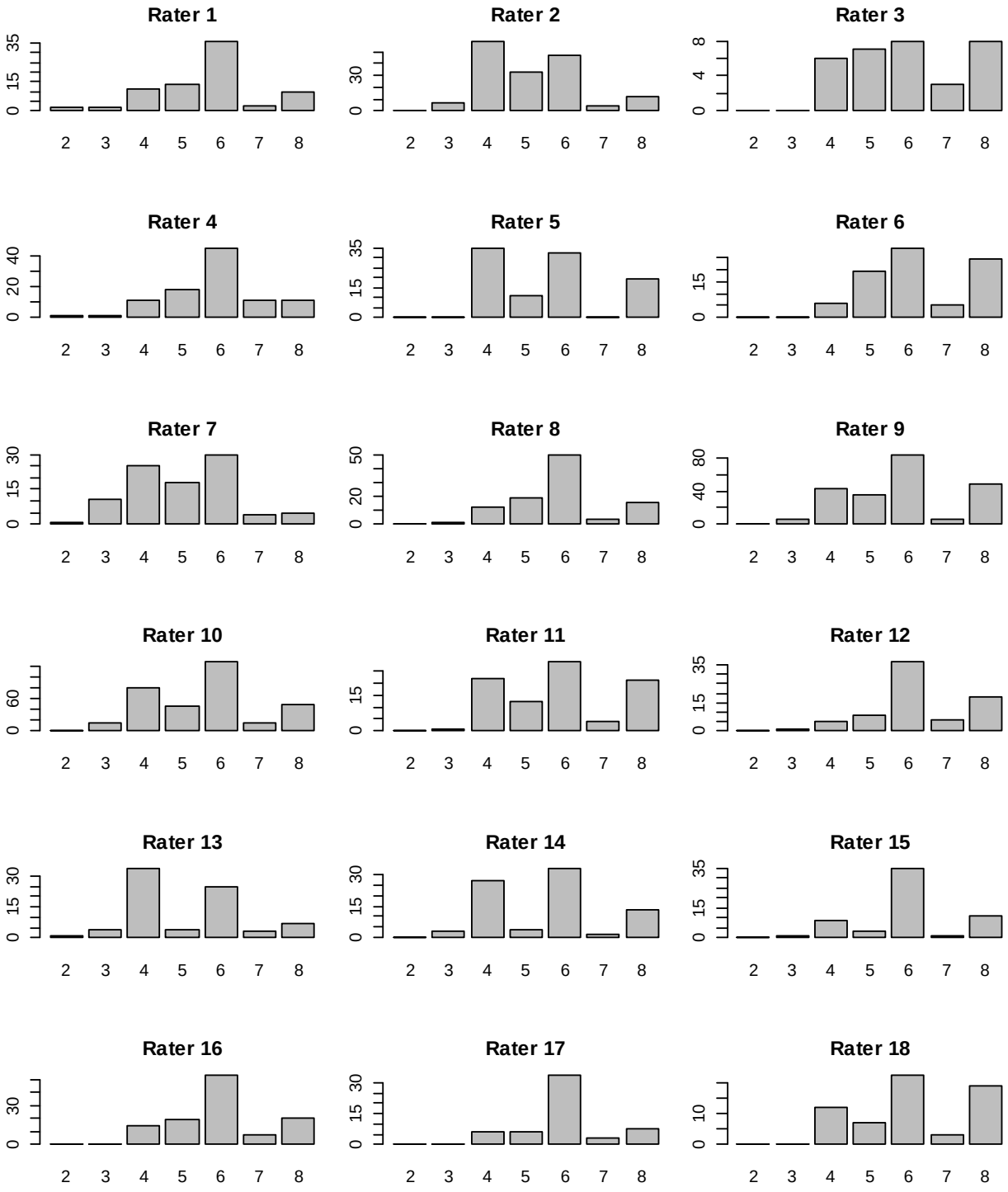


Figure 3-1. Univariate Pronunciation score frequencies from all 18 raters. Scores from two tasks were summed for a clearer presentation.

Another indirect way of comparing rater severity would be to examine the score frequencies of test takers who were rated by the same rater pairs. One rater pair, Rater 2 and Rater 10, marked a total of 74 test takers together, and their score distributions from the 74 test takers are given in Figure 3-2. Figure 3-2 suggests that the two raters yielded different score distributions despite the same performances they rated. Other pairs also presented similar discrepancies.

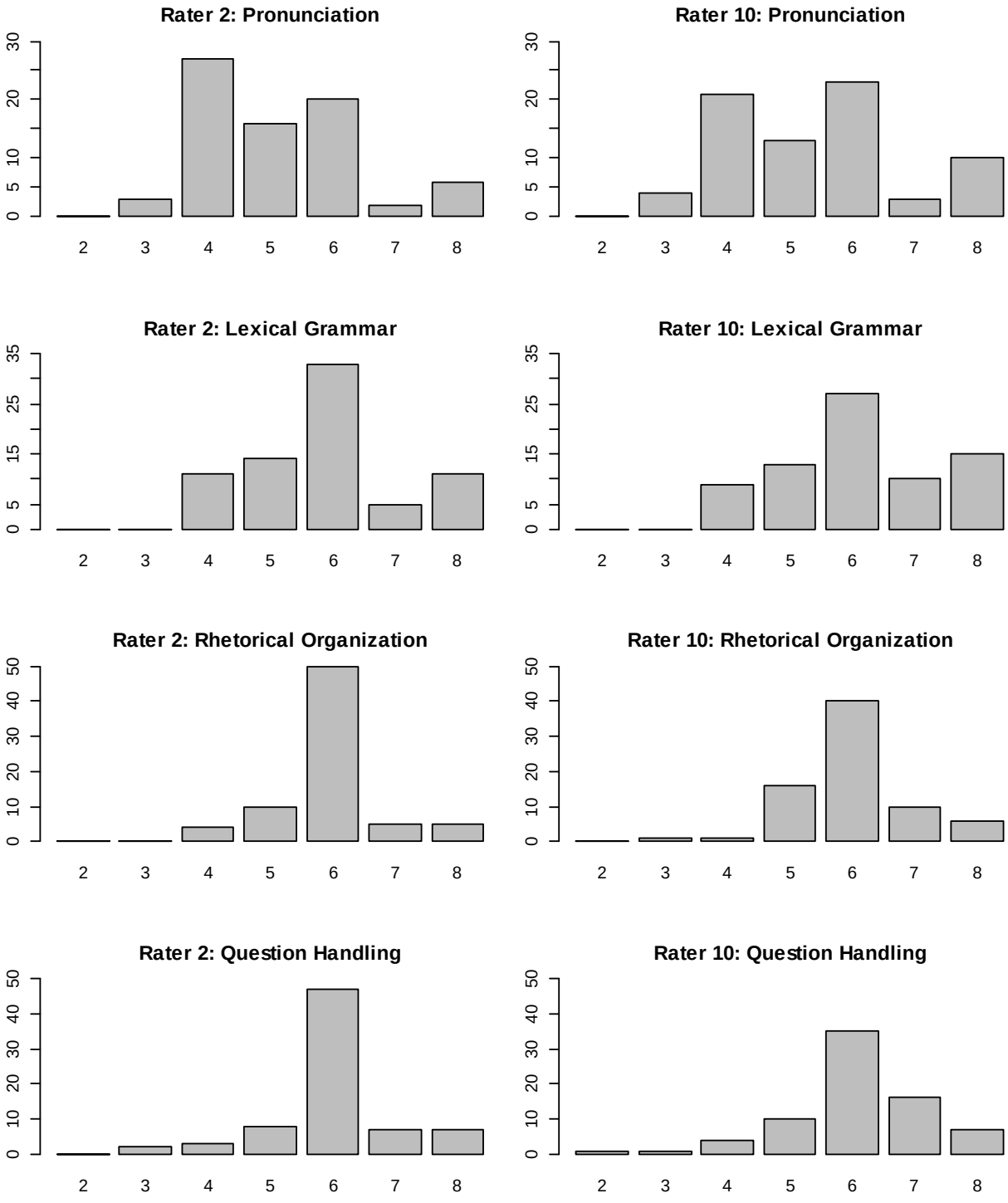


Figure 3-2. All subscale score frequencies from 74 students rated by Rater 2 and Rater 10.

Rater effects are of substantial importance, considering that the ultimate goal of this study was the estimation of subscale scores. The model-based subscale scores would be used as data

which, in turn, would be clustered to yield subscale score profile groups. Since the TOP uses a four-point scale for all subscales, there is only a finite number of possible score patterns. In addition, from Figures 3-1 and 3-2, it is evident that the lowest score was seldom assigned, which resulted in an even smaller number of effective score patterns. Even a one-point classification error from a rater could bring about substantial shifts in classification results. Therefore, rater effects should be properly accounted for in order to obtain accurate profile groups of the TOP subscale scores.

This study formulated an IRT model to address the structural dependency and the rater effects in the observed TOP scores. In particular, this study extended the multi-rater ordinal model of Johnson (1996) to a second-order graded response model (GRM) (Samejima, 1996) for analyzing the TOP data structure. In the following sections the formulation of a second-order graded response model will be presented in detail. A variant of the multi-rater ordinal model will be considered first as the essential component and will be followed by the rationale for having a higher-order structure to account for the dependencies and rater effects in the TOP scores. The mathematical formulation of the second-order GRM will be provided next. Then, the second-order GRM will be reparameterized into a restricted testlet model (Wainer, Bradlow, & Wang, 2007) as well as a restricted bifactor model (Gibbons & Hedecker, 1992), following Rijmen (2010). The relationship between the second-order GRM and the testlet model provides an insight into the role of raters and tasks in multi-rater, multi-task contexts such as the TOP. Furthermore, the relationship between the second-order GRM and the bifactor model allows the use of an efficient maximum likelihood estimation method, namely the bifactor dimension reduction, in estimating the second-order GRM.

3.2.4 The second-order GRM

The multi-rater ordinal model (Johnson, 1996) was developed to effectively combine multiple rater scores on a single essay task to estimate test takers' writing proficiency. In addition, Baldwin, Bernstein, and Wainer (2009) utilized a Bayesian GRM in a similar "multi-rater, single-task" context for a medical application. These two models form the basis of the second-order GRM proposed in this study.[1] Since the TOP involved multiple tasks rated by multiple raters, the two models were expanded to combine information from multiple tasks. The extension utilized a second-order factor structure to combine multiple task scores, hence its name, the second-order GRM. While the extension of the base models to the second-order GRM is straightforward, real-world applications of the extension have yet to be reported in the literature to the knowledge of the author.

3.2.4.1 Model notations and assumptions

The second-order GRM was formulated considering testing situations in which there exist multiple tasks or items, each of which is marked by multiple raters to yield ordinal scores. Therefore, it is necessary to establish a general notation system that takes account of test takers, tasks, raters, and scores with finite ordinal categories. Suppose there are n test takers who perform on J tasks, which are rated by all or some of R raters. The number of raters for each test taker × task combination can vary, while in the TOP scoring scheme it was fixed at two. Test takers will be denoted by subscript i, tasks by subscript j, and raters by subscript r. That is, $i = 1, \ldots, n$; $j = 1, \ldots, J$; and $r = 1, \ldots, R$. Furthermore, suppose raters classify test taker

[1] It should be noted that the two base models are essentially equivalent in that one can be obtained from the other with an appropriate reparameterization. However, they assumed different priors for some parameters.

performances to K ordered categories. When the number of scoring categories differs across tasks, the task subscript should be added. That is, K_j should be used instead of K. Let y_{ijr} denote the observed score of test taker i performing on task j marked by rater r. It follows that, assuming the same number of scoring categories for all tasks, $y_{ijr} \in \{1, \ldots, K\}$. The ultimate interest in this setup is to estimate each test taker's proficiency, based on observed scores assigned by raters (i.e., y_{ijr}).

A fundamental model assumption, which is shared by all IRT models, is that test taker proficiency is a latent trait that governs test taker performances on test tasks (or test items, using a more standard IRT terminology). A standard IRT model makes an additional assumption that task scores are observed with certainty. This additional assumption is reasonable when a test taker response can be objectively scored. A dichotomously scored multiple choice question is a quintessential example of an objectively scored task. However, when human raters are involved in assigning task scores, scoring errors cannot be ignored. In other words, task scores cannot be taken as observables for tasks involving human raters. An additional assumption is needed to remedy this situation. In particular, it is assumed that there exists an unobserved quantity that represents the task score free from all rater errors. It is further assumed that this "rater-free" task score is continuous on an arbitrary scale.[2] For simplicity, this rater-free task score will be referred to as the true task score.

With the true task score in mind, observed scores can be conceptualized as its error-prone ordinal indicators. That is, a score assigned by a human rater can be understood as the result of

[2] Not all rater models make this continuity assumption. For example, the hierarchical rater model (Patz, Junker, Johnson, & Mariano, 2002) and its variants (Mariano & Junker, 2007; DeCarlo, Kim, & Johnson, 2011) assume discrete "true-task" scores on the same scale as the observed scores.

error-prone discretization of the true task score on an ordinal scale based on rater-specific cutoff thresholds on the true task score scale. When there are multiple tasks, test taker proficiency can be estimated from true task scores, which, in turn, can be estimated from observed rater scores. Following the standard notation in the IRT literature, let θ_i denote test taker i's proficiency. Furthermore, let ξ_{ij} denote test taker i's true task score for task j. This two-step procedure implies a second-order structure illustrated in Figure 3-3.

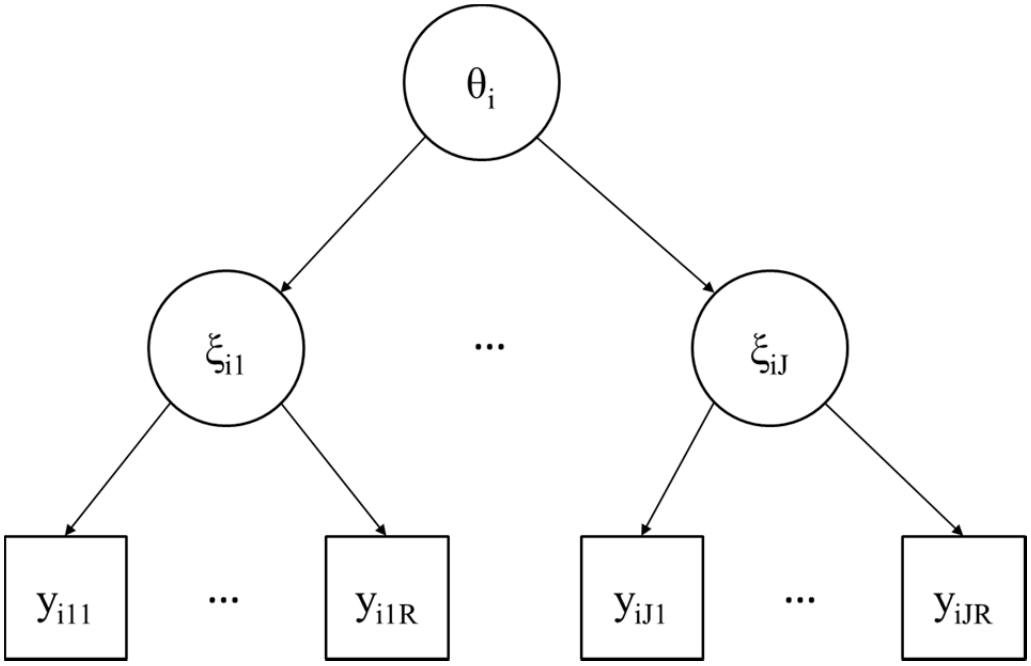


Figure 3-3: A path diagram representing the two-step procedure to estimate test taker proficiency from observed rater scores.

3.2.4.2 Modeling the rating procedure

A rater's scoring procedure is assumed to follow a two-stage process. The first stage involves a rater's perception of the true task score of test taker i on task j, namely ξ_{ij}. Raters are known to differ in terms of their harshness or leniency (e.g., McNamara, 1996), and such differences can

be modeled at this level. In particular, let τ_r denote rater *r*'s overall harshness. In addition, due to the inherent subjectivity and inconsistency involved in human rating, it is reasonable to assume a random error term attached to the rater perception of true task scores. This error term captures the random perturbation of rater *r*'s perception across different ξ_{ij}, and will be denoted with ε_{ijr}. The aforementioned formulation decomposes rater *r*'s perception of ξ_{ij} into three different parts: the true task score (ξ_{ij}), the fixed overall harshness effect specific to rater *r* (τ_r), and the random error term (ε_{ijr}). Formally put,

$$v_{ijr} = \xi_{ij} - \tau_r + \varepsilon_{ijr}, \quad (1)$$

where v_{ijr} denotes rater *r*'s perception of ξ_{ij}.

In most practical cases, rater scores are an ordinal variable with more than or equal to two categories. Therefore, it follows that observed rater scores are an ordinal variable with *K* categories, where $K \geq 2$. The corresponding model assumption is that raters classify their perception of a test taker's performance on a task into *K* ordered categories, based on $K - 1$ cutoff points. It is further assumed that rater *r* classifies test taker *i*'s performance on *j* as *k*, where $k = 1, 2, \ldots, K$, if and only if v_{ijr} lies between the $k\text{-}1$th largest and the *k*th largest category cutoffs. Let γ_{rk} denote the *k*th largest cutoff point of rater *r*. The discretization process can then be represented as

$$y_{ijr} = k \Leftrightarrow \gamma_{r,k-1} < \xi_{ij} - \tau_r + \varepsilon_{ijr} \leq \gamma_{rk}, \quad (2)$$

subject to a weak order constraint $\gamma_{r1} \leq \ldots \leq \gamma_{rK-1}$. It is standard and reasonable to set $\gamma_{r0} = -\infty$ and $\gamma_{rK} = \infty$. When it is reasonable to assume that raters use different category cutoffs across *J* tasks, triple subscripts should be used for γ, such as γ_{jrk}.

The above rating procedure can be made into a stochastic model when a distributional assumption is made on the random error term. Following the standard choice in the IRT literature, let the random error term follow a normal distribution. In particular, it is assumed that $\varepsilon_{ijr} \sim N(0, \sigma_r^2)$, where σ_r^2 denotes rater *r*'s overall lack of discrimination. That is, ε_{ijr} is assumed to be normally distributed around zero, with the only source of variation being rater *r*'s lack of discrimination. σ_r^2 has a clear substantive interpretation; when the random error of a rater fluctuates widely across different ξ_{ij}, the rater's perception is less likely to be accurate, hence low discrimination. Given the distributional assumption on ε_{ijr} , its distribution function is

$$\Phi\left(\frac{\varepsilon_{ijr}}{\sigma_r}\right), \tag{3}$$

where $\Phi(.)$ denotes the standard normal distribution function. It follows from (2) and (3) that

$$\Pr\left(y_{ijr} = k\right) = \Phi\left(\frac{\gamma_{rk} - \left(\xi_{ij} - \tau_r\right)}{\sigma_r}\right) - \Phi\left(\frac{\gamma_{r,k-1} - \left(\xi_{ij} - \tau_r\right)}{\sigma_r}\right), \tag{4}$$

which is essentially identical to Johnson's (1996) multi-rater ordinal model except for the overall rater bias term τ_r and the additional subscript for tasks (i.e., *j*). It is also very similar to a Bayesian graded response model of Baldwin, Bernstein, and Wainer (2009), and their measure of within- and between-rater consistency can be directly adopted here.

The model in (4) is not identified. For example, let $\gamma_{rk}^{=} = \gamma_{rk} + z$, and $\tau_r^{=} = \tau_r + z$, for all *k* = *1*, … , *K*−*1*, *r* = *1*, . . . , *R*, and some constant *z*, the exact same form as (4) is obtained. Therefore, for identification, let $\tau_1 = 0$, such that all other τ_r represent an overall harshness of rater *r* relative to rater *1*. The choice of rater *1* as the reference is completely arbitrary. Also for identification, a sum-to-zero constraint is made on γ. That is, $\gamma_{r,K-1} = -\sum_{k=1}^{K-2} \gamma_{rk}$ for all *j* and *r*.

3.2.4.3 Combining information from multiple tasks

Figure 3-3 shows that test taker *i*'s proficiency, θ_i, can be obtained via a confirmatory factor analysis model based on true tasks scores, ξ_{ij}. Let $\boldsymbol{\xi}_i$ denote the vector of true task scores for test taker *i*. That is, $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{iJ})^{\mathrm{T}}$. A factor analytic representation of the relationship between $\boldsymbol{\xi}_i$ and θ_i is given by

$$\boldsymbol{\xi}_i = \boldsymbol{\Lambda}\theta_i + \boldsymbol{u}_i, \tag{5}$$

where $\boldsymbol{\Lambda}$ is a factor loading matrix and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iJ})^{\mathrm{T}}$ is a $J \times 1$ error vector. Following the factor analysis tradition, it is assumed that $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a $J \times J$ diagonal matrix. It is also assumed that θ_i follows a normal distribution. In particular, $\theta_i \sim N(0, \sigma^2_\theta)$. Depending on a specific modeling context, elements of $\boldsymbol{\Psi}$ or σ_θ should be fixed for identification. Assuming independence between $\boldsymbol{u}_i$ and θ_i, it follows that the variance-covariance matrix of $\boldsymbol{\xi}_i$, Σ_ξ, has a structure given by

$$\boldsymbol{\Sigma}_\xi = \sigma^2_\theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\boldsymbol{T}} + \boldsymbol{\Psi}. \tag{6}$$

From (5) and (6), it is clear that the second-order GRM utilizes a unidimensional confirmatory factor analysis model to combine information from multiple tasks, and the resulting factor scores are the estimates of test taker proficiency.

From (4) and (5), it follows that the second-order GRM has the following model likelihood. Assuming independence, the first-order model likelihood is given by

$$\mathrm{L}(\{\boldsymbol{\xi}_i\}, \{\gamma_{rk}\}, \{\tau_r\}, \{\sigma_r\}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \prod_{j=1}^{J} \prod_{r \in R_{ij}} \left[\Phi\left(\frac{\gamma_{r,y_{ijr}} - \xi_{ij} - \tau_r}{\sigma_r} \right) - \Phi\left(\frac{\gamma_{r,y_{ijr}-1} - \xi_{ij} - \tau_r}{\sigma_r} \right) \right], \tag{7}$$

where R_{ij} denotes the set of raters who evaluated test taker i's performance on task j. The first-order model has essentially the same likelihood as the multi-rater ordinal model of Johnson (1996), but with two deviations: the addition of the overall rater bias term τ_r and the product operator over J tasks. Furthermore, it follows from (5) that

$$\boldsymbol{\xi}_i|\theta_i \sim \mathrm{N}(\boldsymbol{\Lambda}\theta_i, \boldsymbol{\Psi});$$

$$\theta_i \sim \mathrm{N}\left(0, \sigma_\theta^2\right).$$

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$. It follows that the second-order model likelihood is given by

$$\mathrm{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \prod_{i=1}^{n}[|2\pi\boldsymbol{\Psi}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\boldsymbol{\xi}_i - \boldsymbol{\Lambda}\theta_i)^T\boldsymbol{\Psi}^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\Lambda}\theta_i)\} \times \frac{1}{\sqrt{2\pi}\sigma_\theta}\exp\{\frac{-\theta_i^2}{2\sigma_\theta^2}\}]. \quad (8)$$

3.2.4.4 Reparameterization of the second-order GRM into a bifactor model

The second-order GRM in the previous sections can be estimated in multiple ways. A full Bayesian approach can be employed when proper prior distributions for all parameters can be specified. Maximum likelihood is another option for the model estimation. However, the likelihood for the second-order GRM involves $J + 1$ dimensions, and its maximization can be computationally challenging even with a relatively small J. Therefore, it is desirable to have an alternative representation of the second-order GRM that could be more efficiently estimated. To this end, this section discusses a reparameterization of the second-order GRM into a restricted bifactor model. The reparameterization relies on Rijmen's (2010) proof that shows the relationships between a bifactor model, an extended testlet model (Wainer, Bradlow, & Wang, 2007), and a second-order IRT model. Through the reparameterization, the model parameters can be efficiently estimated using the bifactor dimension reduction (Gibbons & Hedecker, 1992).

The reparameterization begins with transformations of the existing parameters. First, define $a_r = 1/\sigma_r$, $a_r v_{ijr} = v^*_{ijr}$, and $a_r \varepsilon_{ijr} = \varepsilon^*_{ijr}$, and multiply both sides of (1) by a_r to obtain

$$v^*_{ijr} = a_r(\xi_{ij} - \tau_r) + \varepsilon^*_{ijr}. \tag{9}$$

Recall the previous definition of σ^2_r as the lack of discrimination in rater r. The above transformation gives a_r a more straightforward interpretation as rater r's discrimination. It follows from (5) and (9) that

$$\begin{aligned} v^*_{ijr} &= a_r(\lambda_j\theta_i + u_{ij} - \tau_r) + \varepsilon^*_{ijr} \\ &= a_r\lambda_j\theta_i + a_r u_{ij} - a_r\tau_r + \varepsilon^*_{ijr} \qquad (10) \\ &= a^{\#}_r(\theta_i + C^{\#}_j u_{ij}) - \tau^*_r + \varepsilon^*_{ijr}, \qquad (11) \end{aligned}$$

where $a^{\#}_r = a_r\lambda_j$, $C^{\#}_j = 1/\lambda_j$, and $\tau^*_r = a_r\tau_r$. It is noteworthy that (11) is formally identical to the second-order IRT model of Rijmen (2010). Rijmen shows that a second-order IRT model such as the one in (11) is identical to the extended testlet response model (Wainer et al., 2007), which, in turn, is a bifactor model with proportionality constraints. To make the connection clear, a general bifactor model can be represented as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a_{jg}\theta_g + a_{js}\theta_s - b_j, \tag{12}$$

where p_{ij} is the probability of test taker i gets item j correct, and a_{jg} and a_{js} are slope parameters for general (θ_g) and specific (θ_s) dimensions, respectively. The proportionality constraints become clear when (10) and (12) are compared. Assuming the standard logistic distribution

rather than the standard normal distribution for ε^*_{ijr}, the probability of test taker *i* receiving score *k* from rater *r* on task *j* is

$$\Pr(y_{ijr} = k) = \Psi(\gamma^*_{rk} - \{a^{\#}_r(\theta_i + C^{\#}_j u_{ij}) - \tau^*_r\}) - \Psi(\gamma^*_{r,k-1} - \{a^{\#}_r(\theta_i + C^{\#}_j u_{ij}) - \tau^*_r\}), \quad (13)$$

where $\gamma^*_{rk} = a_r \gamma_{rk}$ and Ψ(.) denotes the standard logistic distribution function. From the relationship between (10) and (12), the reparameterized second-order GRM in (13) belongs to the family of generalized bifactor models (Cai, Yang, & Hansen, 2011) with proportionality constraints on the slope parameter for u_{ij}.

3.2.4.5 The conceptual analogy between the second-order GRM and the extended testlet model

The close algebraic relationship between the second-order GRM and the extended testlet model provides an insight into the interesting similarity between the role of raters in a multi-rater, multi-task context and that of items in a testlet-based test. Raters in the former provide scores that function as indicators of true task scores, which are, in turn, indicators of test taker proficiency. Rater scores are based on the test taker performance on a task, and therefore, rater scores within the same tasks are expected to be more highly correlated with one another than with rater scores across different tasks. This relationship between raters and tasks in the second-order GRM is analogous to the relationship between items and testlets in the testlet model. In the latter, items are indicators of test taker proficiency, and within-testlet items are structurally dependent upon one another by the virtue of belonging to the same testlet. The analogy also holds for the parameters of the two models. Each rater has its own severity and discrimination parameters in the second-order GRM. Similarly, each item has its own difficulty and discrimination parameters in the extended testlet model. In this light, it is reasonable to consider

the second-order GRM as a testlet model that treats raters as items and tasks as testlets. An important difference between the two models lies in the feasibility of additional constraints on rater/item parameters. In a typical testlet-based test, each testlet consists of different items; it is difficult to imagine two different testlets sharing an item. However, it is common to have the same set of raters evaluate multiple tasks of the same test takers. Therefore, the second-order GRM has a potential to become more parsimonious. In particular, it can become a testlet model with additional equality constraints on rater severity and discrimination of the same raters across different tasks.

Treating raters as if they were items is not entirely new. Both the multi-rater ordinal model (Johnson, 1996) and the Bayesian GRM (Baldwin et al., 2009) adopt this analogy. Johnson (1996) does not make explicit the relationship between raters and items. On the other hand, Baldwin et al. (2009) justify their approach of treating raters as items in that rater judgment (in their case, physicians) about the same stimuli or performances (in their case, radiographs) are not supposed to differ widely (p. 2282). Their logic is pertinent to a typical multi-rater, multi-task situation in which one does not expect large fluctuations among rater scores on the same performance. Raters in most operational testing contexts are trained and normed in that, despite frequent and expected disagreements, their scores do not, in general, deviate too much from one another.

In addition to the computational advantage in model estimation, the relationship between the second-order GRM and the extended testlet model facilitates the conceptualization of complicated real-world rater assignment scenarios. For practical reasons, operational tests seldom adopt a complete-block design in which every rater scores every test taker on every task. Instead, an incomplete-block design, in which a subset of raters marks only a subset of test takers

performing on a subset of tasks, has been much more commonly employed. Such an incomplete design poses a challenge to many existing models, but can be easily conceptualized and estimated with the second-order GRM. Keeping the raters-as-items analogy, the entire rater pool can be conceptualized as a test form. For example, a rater pool consisting of *R* raters can be regarded as a test form with *R* items. Then, an incomplete-block design reduces to a missing data problem, in which test takers respond to only a subset of items (i.e., raters). If test takers were rated by two raters in the example, there will be *R-2* missing responses per test taker. Raters are often randomly assigned to test takers or to test taker and item combinations, and therefore missing completely at random (MCAR) (Rubin, 1976) is a reasonable assumption. While missing data still present a challenge for estimation, this conceptualization provides a straightforward way to model complicated rater assignment schemes based on an incomplete-block design.

3.2.4.6 Estimation

Given the likelihoods in (7) and (8), multiple estimation methods can be utilized to estimate the second-order GRM. This study employed two such methods, namely maximum likelihood and Markov chain Monte Carlo (MCMC), to fit the model to the TOP data. Parameter recovery capability of the second-order GRM was evaluated based on a simulation study, which will be presented in detail in Chapter 4. The maximum likelihood approach was used to obtain sound initial values for the MCMC approach, which was the main estimation method in this study. Identification conditions and equality constraints specific to the TOP data will be described in detail in Chapter 4. Prior distributions for the MCMC estimation will also be discussed in Chapter 4. The maximum likelihood estimation was conducted using flexMIRT version 1.86

(Cai, 2012). The reparameterization in the previous section, combined with flexMIRT's built-in dimension reduction functionality, facilitated the estimation process. The MCMC estimates were obtained using WinBUGS 1.4.3 (Lunn, Thomas, Best, & Spiegelhalter, 2000).

3.2.5 FM model analysis

The model-based subscale scores estimated using the MCMC approach comprised the data matrix **X**. Based on that data matrix, a mixture of four dimensional multivariate normal distributions were employed to identify distinct subscale score profiles. Given the lack of previous knowledge about the number of distinctive profiles, models consisting of different numbers of component distributions were fitted to the estimated TOP proficiency. The considered models were estimated using an EM algorithm with non-parametric hierarchical clustering results as initial values. The final model was selected based on statistical goodness of fit and substantive interpretability. The selected model provided estimates of the proportion of each profile group in the 960 TOP test takers, as well as the probability of a test taker belonging to each profile group. The composition of each profile group was examined in terms of members' linguistic and academic background, and their TOP decision categories. The profile groups were then interpreted and labeled based on the score patterns and background shared by members. All FM model analyses were conducted using the mclust package (Fraley, Raftery, Murphy, & Scrucca, 2012) in R (R Core Team, 2012).

3.3 Stage 2: Features of oral discourse across subscale score profiles

3.3.1 Objectives

In this stage, videotaped performances of representative test takers from each subscale score profile group were transcribed and analyzed to examine characteristic features of language use in each profile group. This analysis also serves as an indirect validation of the subscale score profiling, in that different score patterns, especially on the Rhetorical Organization and Question Handling subscales, are expected to be related to differences in salient discourse features.

3.3.2 Data

All TOP test administrations have been videotaped and stored in a digital medium. The duration of the test varies from one test taker to another and typically ranges from ten to twenty five minutes. The first several minutes of the test contain a warm-up task which is not scored, and thus were excluded from all analyses in this study. The syllabus presentation task was also excluded from the subsequent analysis because of its limited variability across different test takers in terms of discourse structures. In particular, performances on the syllabus presentation task are expected to be structured based on the syllabus given to test takers. The remaining performances of test takers (i.e., the lecture task performances) were transcribed and the transcripts constituted the main data for the analysis.

The large amount of time required for the transcription and building a corpus prohibited a large sample size. Consequently, it was decided to select a subset of the TOP test takers and focus on the language use of the selected subset. The representativeness of the sample was crucial in this stage, and therefore, efforts were made to achieve representativeness of the

selected sample. The sampling scheme for obtaining a representative subset of each subscale score profile group was based on the FM modeling results in the previous stage, and therefore, will be reported in detail in Chapter 5. A total of 82 TOP test takers' performances were selected for transcription. Profile group membership and background variables of the 82 test takers will be presented in Chapter 5.

The sampled subset of the videotaped performances was transcribed following the notations and conventions used in the conversation analysis (CA) literature (Sacks, Schegloff, & Jefferson, 1974). Utterances from test takers and questions from the TOP questioners, as well as overlaps between them, were separately transcribed and marked. The resulting transcriptions were used to construct a small corpus of the TOP test takers' performances following the procedures described in Biber et al. (1998) and McEnery et al. (2006). Finally, a group-specific sub-corpus was constructed for each subscale score profile group. All analyses in this stage were conducted based on the resulting corpus and group-specific sub corpora.

3.3.3 Analysis

The second research question, "What features of oral discourse characterize test takers who have different subscale score profiles?", was addressed by the two following corpus-based analyses. First, the lecture task performances of the 82 selected test takers were compared to a reference corpus consisting of academic lectures. The specific corpus used for this comparison was the Michigan Corpus of Academic Spoken English (MICASE) (Simpson, Briggs, Ovens, & Swales, 2002). This corpus is described in detail in Chapter 5. Second, lexical bundles that serve a discourse organizing function were identified and analyzed to explore their patterns of distribution across the subscale score profile groups. The objective of the first analysis was to

investigate the overall language use of the TOP test takers by comparing their performances to a reference collection of academic lectures given by native speakers of English. The second analysis, on the other hand, was designed to provide an insight to the discourse structures of the TOP test takers.

The comparison between the TOP performances and the reference corpus was made by examining a set of words that were used significantly more or less frequently by the selected TOP test takers than by speakers in the reference corpus. It was expected that the differences in frequency of word use between the two corpora would help explain the characteristics and peculiarities of the TOP test taker performances. In addition, each group-specific sub-corpus was compared to the reference corpus in the same manner. This group level analysis was designed to find potential differences in overall language use of test takers who had different subscale score profiles. The keyword component of AntConc 3.2.4 (Anthony, 2011) was used in comparing the corpora.

The analysis of lexical bundles in the TOP test taker performances was motivated by recent findings in the corpus linguistics literature. This research suggests that multiword expressions are an integral discourse organizing device that contributes to the overall coherence of a given text or speech (Biber et al., 1999; Halliday & Hasan, 1976; Nattinger & DeCarrico, 1992). Biber et al. (2004) provide a functional taxonomy of lexical bundles including discourse organization. Biber (2006) later suggested discourse organization bundles (e.g., *going to talk about*, *let's have a look at*, and *to go ahead*) as a characteristic feature of academic lectures. Nesi and Basturkmen (2009) use Biber et al.'s taxonomy to classify lexical bundles occurring in a corpus consisting of academic lectures and investigate the discourse signaling role of such bundles. Based on an

investigation of another large academic corpus, Simpson-Vlach and Ellis (2010) expanded Biber et al.'s taxonomy and present a comprehensive list of lexical bundles that serve different functions in the expanded taxonomy. Furthermore, Simpson-Vlach and Ellis suggest four subcategories of the discourse organizing function, namely metadiscourse and textual reference, topic introduction and focus, topic elaboration, and discourse markers.

The lexical bundle analysis in this study focused on the first subcategory of Simpson-Vlach and Ellis' (2010) taxonomy, namely metadiscourse and textual reference (MTR), which they define as discourse organizing bundles that explicitly refer "to prior or upcoming discourse" (p. 507). The choice to focus on the MTR category was dictated by both substantive and practical considerations. Simpson-Vlach and Ellis suggest that the MTR bundles are genre-specific in that MTR bundles that are commonly used in spoken academic discourse did not overlap with popular MTR bundles in written discourse, and vice versa. Given the focus on academic oral English proficiency in this study, investigating features specific to academic oral discourse was believed to be a logical choice. In addition, other functions were mainly handled by a function word rather than by lexical bundles. For example, the vast majority of topic elaboration were handled by *so* and *because* in the TOP test taker performances. From a practical point of view, lexical bundles that belonged to the other subcategories were either very difficult to identify and classify or oftentimes not relevant. Topic introduction and elaboration were not always exclusive, and therefore, distinguishing the two functions was not straightforward. Discourse markers in Simpson-Vlach and Ellis' (2010) taxonomy include interactional bundles such as *oh my god* and *thank you very much*, which were not pertinent to the testing context.

MTR bundles in the TOP test taker performances were identified using a combination of search queries and manual reading of the CA transcripts. The identification procedure will be reported in detail in Chapter 5. The identified MTR bundles were then classified in terms of their timeframe. As Simpson-Vlach and Ellis (2010) noted, the MTR bundles refer to a previous or upcoming point in discourse. The timeframe coding used this dichotomous distinction between past and upcoming events in discourse, and categorized all MTR bundles into either the past- or future-reference subgroup. Furthermore, whether a given MTR bundle was grammatically well-formed or not was recorded to examine the formal appropriateness of MTR bundle use in the TOP test taker performances. The distribution of different MTR bundles across the subscale score profile groups was then investigated to explore potential relationships between MTR bundle use and subscale score profiles.

Chapter 4: TOP Subscale Score Profile Groups

This chapter reports the findings of analyses addressing the first research question, "What different subscale score profiles of academic English oral proficiency can be meaningfully produced by the TOP?" In particular, it presents the procedures and the results of model-based subscale score estimation and clustering. This chapter is organized according to the order of analysis. First, the results of a simulation study evaluating the parameter recovery of the second-order GRM are reported. The estimation of model-based subscale scores is presented next, followed by the exploration of subscale profile groups based on the model-based scores. The resulting profile groups are further examined in terms of the TOP decision categories and the linguistic background of group members.

4.1 Parameter recovery of the second-order GRM

Parameter recovery, which refers to a model's capability of accurately estimating true parameter values, is essential for any statistical model. The use of a model cannot be justified when its ability to recover parameters is unknown or doubtful. While the second-order GRM introduced in Chapter 3 is not entirely new in that it is a straightforward extension of existing IRT models, the lack of its prior use in the literature means that it has not been properly evaluated in terms of parameter recovery. Therefore, a small-scale simulation study was conducted to investigate the parameter recovery of the second-order GRM. This section describes the procedures and the results of the simulation study.

4.1.1 Simulation study setup

The parameter recovery of the second-order GRM was evaluated in three different settings. The settings were designed to match the characteristics of typical language performance assessment contexts, with a specific focus on the TOP dataset. All three settings had 960 test takers and four scoring categories per task, but they differed in terms of the number of tasks and raters. The first setting (hereafter Setting 1) was designed to be the most favorable situation. In particular, Setting 1 specified a complete block design with five tasks and five raters fully crossed, so that all five tasks were rated by each and every rater. The second setting (hereafter Setting 2) also specified a complete block design with fully crossed tasks and raters, but with fewer tasks than Setting 1. In particular, only two tasks were specified in Setting 2. Due to the smaller number of tasks, Setting 2 was a less favorable condition for the model than Setting 1. The third setting (hereafter Setting 3) shared the number of tasks and raters with Setting 2, but presented a more challenging scenario in that it adopted an incomplete block design with missing data. In particular, it was specified in Setting 3 that only two raters (out of five) were randomly assigned to each test taker. Operationally, the data for Setting 3 were obtained by randomly deleting three out of five rater scores per each test taker from the Setting 2 data. Setting 3 was designed to be highly comparable to the TOP dataset, which had the same number of test takers (i.e., 960), tasks (i.e., two) and raters per test taker (i.e., two), as well as a large amount of missing data due to the incomplete block design. In both Setting 3 and the TOP dataset, the raters were missing completely at random (MCAR) (Rubin, 1976) in that raters were assigned to test takers in a completely random fashion. The differences between the three settings are summarized in Table 4-1.

Table 4-1

Settings for the Parameter Recovery Simulation Study

	Setting 1	Setting 2	Setting 3
Test takers (N)	960	960	960
Scoring categories (K)	4	4	4
Tasks (J)	5	2	2
Raters per test taker (R)	5	5	2
Missing data	No	No	Yes
Rater assignment	Complete block	Complete block	Incomplete block

True parameter values for data generation were set in the following way. Both the test taker proficiency (i.e., θ_i) and the task-specific error (i.e., u_{ij}) were randomly drawn from a normal distribution centered at zero. The ratio of the task-specific error variance (i.e., diagonal elements of $\boldsymbol{\Psi}$) to the proficiency variance (i.e., σ_θ^2) was roughly based on a previous Generalizability theory analysis of another TOP dataset (Schmidgall, 2012). In particular, all task-specific variances were set to one, and the proficiency variance was set to eight. To retain the variance ratio between the task-specific variances and the proficiency variance, all loadings from the proficiency factor to the task factors (i.e., λ_j) were set to one. The rater discrimination parameters (i.e., a_r) were generated from the standard normal distribution truncated below zero. The overall rater severity parameters (i.e., τ_r) were drawn from the standard normal distribution, except for τ_1 which was set to one for identification. Lastly, the rater cutoff parameters (i.e., γ_{rk}) were drawn from the standard normal distribution with the weak order constraints and the sum-to-zero constraints described in Chapter 3. Based on the true parameter values, data for all three settings were generated using the second-order GRM as the data generation mechanism in R (R Core Team, 2012). The R syntax for the data generation is provided in Appendix A.

This simulation study estimated the model parameters using the same procedures as were intended to be used in the actual TOP model-based subscale score estimation. First, maximum likelihood (ML) estimates were obtained without separating the overall rater severity and the rater cutoff parameters. The resulting ML point estimates and standard errors were then used to obtain starting values for subsequent MCMC estimation of test taker proficiency. Prior distributions for the MCMC estimation were set to be equal to the generating distributions of the true parameter values, except for the proficiency variance (i.e., σ_θ^2), whose prior distribution was set to follow an inverse-Gamma distribution. Three parallel MCMC chains were constructed. The first chain contained the ML point estimates as its starting values. The second and third chains began with the ML point estimates plus and minus, respectively, two point half times the corresponding standard errors as starting values. Since the ML estimation did not distinguish the rater severity and the rater cutoff parameters, starting values for these parameters were randomly drawn from the standard normal distribution. Each chain was run for 20,000 iterations. The flexMIRT (Cai, 2012) and WinBUGS (Lunn et al., 2000) syntax for the ML and MCMC estimation, respectively, are provided in Appendix B.

4.1.2 Simulation results

Before presenting the parameter recovery results from the three settings, a note on model fit evaluation for the second-order GRM is in order. Both the simulation study in this section and the main study described in the later sections adopted a full Bayesian framework and utilized MCMC as the main estimation method. Model fit in Bayesian modeling is often assessed based on relative fit indices such as the Akaike Information Criterion (AIC; Akaike, 1974), the Bayesian Information Criterion (BIC; Schwartz, 1978), and the Deviance Information Criterion

(DIC; Spiegelhalter, Best, Carlin, and van der Linde, 2002) or through comparisons between the data and simulated predictions by the model (Gelman, Carlin, Stern, & Rubin, 2003). However, absolute indices that can evaluate the degree of model fit against a known distribution can be very convenient. Therefore, whether absolute model fit indices obtained in the ML estimation stage could be used to evaluate the goodness of fit of the second-order GRM was also investigated in the simulation study.

While full-information fit statistics such as the Pearson's chi-square statistic and the likelihood-ratio statistic are available, their performance under sparse underlying contingency tables is known to be inferior to limited-information fit statistics such as M_2 (Maydeu-Olivares & Joe, 2005). The original M_2 statistic was developed for dichotomous response variables, and Cai and Hansen (2012) suggest an extension of M_2 that is more suitable when response variables are polytomous. Consequently, the Cai-Hansen M_2 statistic and the corresponding *p*-value obtained in the ML estimation stage were examined to evaluate the model fit in the three simulation settings. In addition, the corresponding Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993) was also monitored.

Since the fitted model was identical to the data generating model in this simulation study, the fit indices should indicate that the model fit the data very well in all three settings. That is, one would expect to observe non-significant *p*-values and near zero RMSEA values across all settings. Table 4-2 provides the Cai-Hansen M_2 statistics, the *p*-values, and RMSEA in Settings 1, 2, and 3.

Table 4-2

Cai-Hansen M_2 Statistics and RMSEA Values for All Three Settings

Setting	Cai-Hansen M_2	M_2 D.F.	M_2 *p*-value	RMSEA
Setting 1	2582.82	2754	1.00	.00
Setting 2	404.05	414	.63	.00
Setting 3	32140.16	414	.00	.21

As can be seen in Table 4-2, the model fit results from Settings 1 and 2 were in line with the expectation. In particular, the Cai-Hansen M_2 statistics and the corresponding RMSEA values suggested almost perfect fit in both settings. However, Setting 3 presented a deviation from the expectation in that both indices suggested poor fit. The only difference between Setting 2 and Setting 3 was the missing data introduced in the latter. Therefore, it is reasonable to infer that the large amount of missing data could have caused problems for the Cai-Hansen M_2 statistic and the corresponding RMSEA[3]. The performance of the Cai-Hansen M_2 and RMSEA in Setting 3 suggested that these indices could be misleading for the actual TOP model-based subscale score estimation, which involved an even larger amount of missing data than Setting 3. Consequently, the goodness of fit of the second-order GRM to the actual TOP dataset was evaluated using a relative fit index.

All three settings involved a total of five raters. Consequently, there were five discrimination parameters that were estimated. Table 4-3 provides the deviation of the estimated values from the true generating values. The corresponding 95 percent Bayesian credible intervals are also given in the deviation form.

[3] It should be noted that this was a small-scale simulation study involving only one dataset per each setting. The performance of the Cai-Hansen M_2 and the corresponding RMSEA under a large amount of missing data deserves a much more rigorous investigation based on extensive replications.

Table 4-3

Deviations of the Estimated Discrimination Parameters from the True Values

Parameter	Setting 1 Deviation	Setting 2 Deviation	Setting 3 Deviation
a_1	-.003 (-.022, .016)	-.018 (-.046, .012)	.019 (-.032, .075)
a_2	.006 (-.084, .099)	-.047 (-.186, .102)	.189 (-.152, .610)
a_3	.011 (-.073, .100)	-.033 (-.177, .120)	.108 (-.214, .479)
a_4	-.034 (-.146, .086)	-.121 (-.293, .065)	.067 (-.353, .600)
a_5	-.008 (-.136, .129)	-.029 (-.233, .200)	.090 (-.372, .689)

Note. The 95 percent Bayesian credible intervals are provided in the parentheses.

As can be seen in Table 4-3, the estimates from the second-order GRM were highly comparable to the true values, indicated by the small magnitude of the deviations. The discrimination parameters were successfully recovered across all three settings in that the 95 percent Bayesian credible intervals contained the corresponding true values in every case. However, there was a clear hierarchy of model performance among the three settings. Setting 1 yielded the most accurate and precise parameter estimates in that its point estimates were closest to the true values and the corresponding intervals were narrowest. Furthermore, the comparison between Setting 2 and Setting 3 showed that the missing data in the latter setting resulted in lower precision indicated by wider credible intervals.

Since there were four scoring categories in each task, all settings involved fifteen rater cutoff parameters (three cutoff parameters for each of the five raters). However, only ten of them were freely estimated due to the sum-to-zero constraint on the third cutoff of each rater. Table 4-4 gives the deviations of the freely estimated cutoff parameters from the generating values. The corresponding 95 percent Bayesian credible intervals are also given in the deviation form.

Table 4-4

Deviations of the Estimated Cutoff Parameters from the True Values

Parameter	Setting 1 Deviation	Setting 2 Deviation	Setting 3 Deviation
γ_{11}	-.003 (-.053, .047)	.035 (-.039, .106)	.059 (-.061, .172)
γ_{12}	-.005 (-.033, .024)	.012 (-.041, .067)	.005 (-.077, .086)
γ_{21}	-.002 (-.097, .093)	-.070 (-.199, .048)	-.235 (-.550, .021)
γ_{22}	-.016 (-.078, .045)	.011 (-.07, .091)	-.010 (-.154, .133)
γ_{31}	-.021 (-.103, .057)	-.047 (-.204, .094)	-.088 (-.398, .214)
γ_{32}	-.021 (-.074, .030)	-.044 (-.135, .045)	-.029 (-.179, .120)
γ_{41}	-.033 (-.126, .056)	.093 (-.030, .203)	.074 (-.169, .280)
γ_{42}	.027 (-.023, .078)	.021 (-.055, .095)	.073 (-.047, .186)
γ_{51}	-.007 (-.092, .076)	-.184 (-.447, .049)	-.101 (-.653, .346)
γ_{52}	-.005 (-.049, .040)	.104 (-.012, .237)	.040 (-.175, .300)

Note. The 95 percent Bayesian credible intervals are provided in the parentheses.

Table 4-4 shows that Setting 1 yielded point estimates with the smallest absolute deviations from the true values, as well as the narrowest credible intervals. While the point estimates from Settings 2 and 3 were similar in terms of the absolute deviation from the true values, Setting 2 resulted in more precise estimation indicated by narrower credible intervals. Overall, however, the estimated cutoff parameters were very close to the true values. Furthermore, the 95 percent credible intervals contained the true value in every case across all settings.

The severity of the first rater, τ_1, was fixed at one to identify the model. The severity parameters of the remaining four raters were freely estimated. The deviations of the resulting point estimates from the true values, as well as the corresponding 95 percent Bayesian credible intervals in the deviation form, are given in Table 4-5.

Table 4-5

Deviations of the Estimated Severity Parameters from the True Values

Parameter	Setting 1 Deviation	Setting 2 Deviation	Setting 3 Deviation
τ_2	-.020 (-.214, .175)	-.106 (-.380, .155)	.004 (-.395, .334)
τ_3	-.028 (-.238, .168)	.048 (-.165, .262)	.142 (-.126, .404)
τ_4	-.001 (-.210, .195)	-.109 (-.409, .173)	.282 (-.146, .637)
τ_5	-.039 (-.231, .138)	-.070 (-.281, .142)	-.081 (-.340, .177)

Note. The 95 percent Bayesian credible intervals are provided in the parentheses.

In general, the point estimates were very close to the true values. In addition, the credible intervals successfully contained the true values every time across all three settings. The most favorable conditions in Setting 1 resulted in the smallest absolute deviations of the point estimates and the narrowest intervals.

The person proficiency variance was freely estimated while the variances of task-specific factors were all fixed at one. Since WinBUGS parameterizes the reciprocal of variance for normal distributions, the reciprocal of the person proficiency variance was set up as a free parameter. Table 4-6 provides the deviations of the resulting estimates from the true generating values from the three settings. The corresponding 95 percent Bayesian credible intervals are also given in the deviation form.

Table 4-6

Deviations of the Estimated Person Proficiency Variance Parameters from the True Values

Parameter	Setting 1 Deviation	Setting 2 Deviation	Setting 3 Deviation
$1/\sigma_\theta^2$	.000 (-.017, .019)	-.010 (-.034, .018)	.007 (-.037, .059)

Note. The 95 percent Bayesian credible intervals are provided in the parentheses.

Table 4-6 shows that the second-order GRM successfully recovered the reciprocal of the proficiency variance parameter in that the true value was contained in the credible intervals in every setting. The point estimates were highly comparable to the true values. Setting 1 yielded the most precise estimate indicated by the narrowest interval.

The aforementioned parameter recovery results showed a consistent pattern. The second-order GRM was capable of recovering every parameter across all three settings. The number of tasks as well as the presence of missing data affected the precision of recovery, which is in line with the original expectation. In particular, the 95 percent Bayesian credible intervals were the narrowest in Setting 1, which had the largest number of tasks (i.e., five) without missing data. On the other hand, Setting 3, which had only two tasks as well as a large proportion of missing data, yielded the least precise estimates indicated by the widest credible intervals. Figure 4-1 presents an illustration of this pattern.

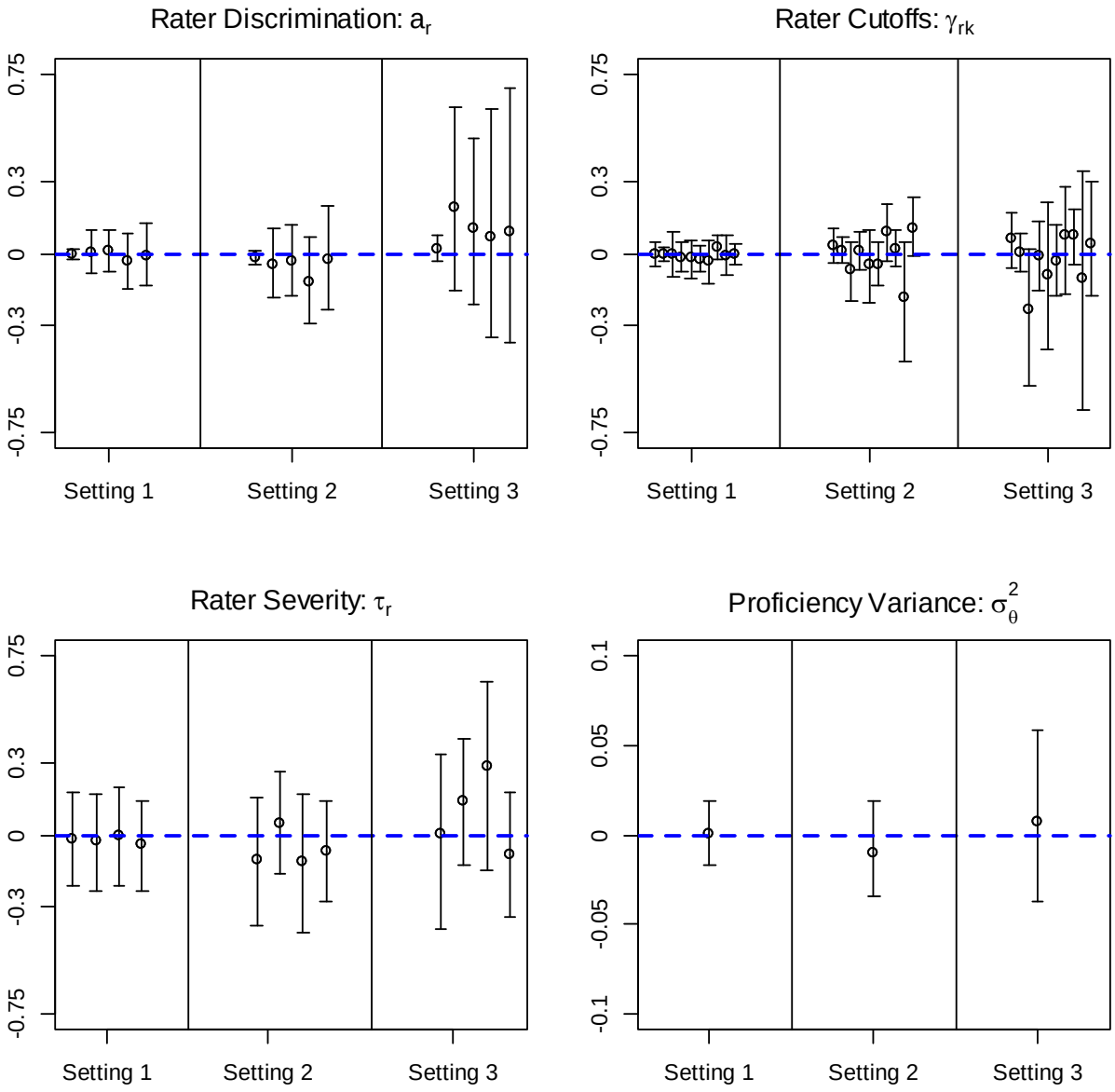


Figure 4-1. Deviations of the second-order GRM estimates from the true values. The circles represent the point estimates from the model, while the intervals stand for the 95 percent Bayesian credible intervals. The dotted lines represent the true values (i.e., zero deviation).

The recovery of test taker proficiency parameters (i.e., θ_i) was evaluated by examining the correlation between the true and estimated values. The correlations between the true values and the standardized sum scores, which were calculated by standardizing the sum of all observed rater scores, were also evaluated. The standardized sum scores are equivalent to the results of a

simple average scoring method. The comparison between the two sets of correlations was made to assess the added-value of the model in estimating the true proficiency. In particular, for the second-order GRM estimation to have added-value over a simple average scoring method, the correlation between the true values and second-order GRM estimates should be higher than the correlation between the true values and the standardized sum scores. The difference between the two sets of correlations can be interpreted as the added-value of using the second-order GRM for estimating the proficiency of each test taker, instead of a simple average scoring method. Table 4-7 gives the correlations of the second-order GRM estimates and the standardized sum scores with the true values across all three settings.

Table 4-7

Correlations with the True Values: the Second-order GRM Estimates and the Standardized Sum Scores

Settings	True vs. Model-based Estimates	True vs. Standardized Sum Scores
Setting 1	.97	.93
Setting 2	.92	.89
Setting 3	.86	.80

Setting 1 yielded estimates that were closest to the true test taker proficiency values. Both the reduction of the number of tasks and the introduction of the incomplete block design for rater assignment negatively affected the proficiency parameter recovery. Table 4-7 shows that the use of the second-order GRM led to better results than using a simple average scoring method. The model-based parameter estimates were more highly correlated with the true values than the standardized sum scores were across all three settings. Furthermore, the added value of using the second-order GRM became larger as the simulation setting became more realistic. In particular,

the largest difference between two sets of correlation coefficients was observed in Setting 3, which involved the smallest number of tasks and a large proportion of missing data.

Figure 4-2 presents bivariate scatter plots between the true proficiency values and the second-order GRM estimates in the three settings. The relationships between the true values and the standardized sum scores are also given in the same format in Figure 4-2.

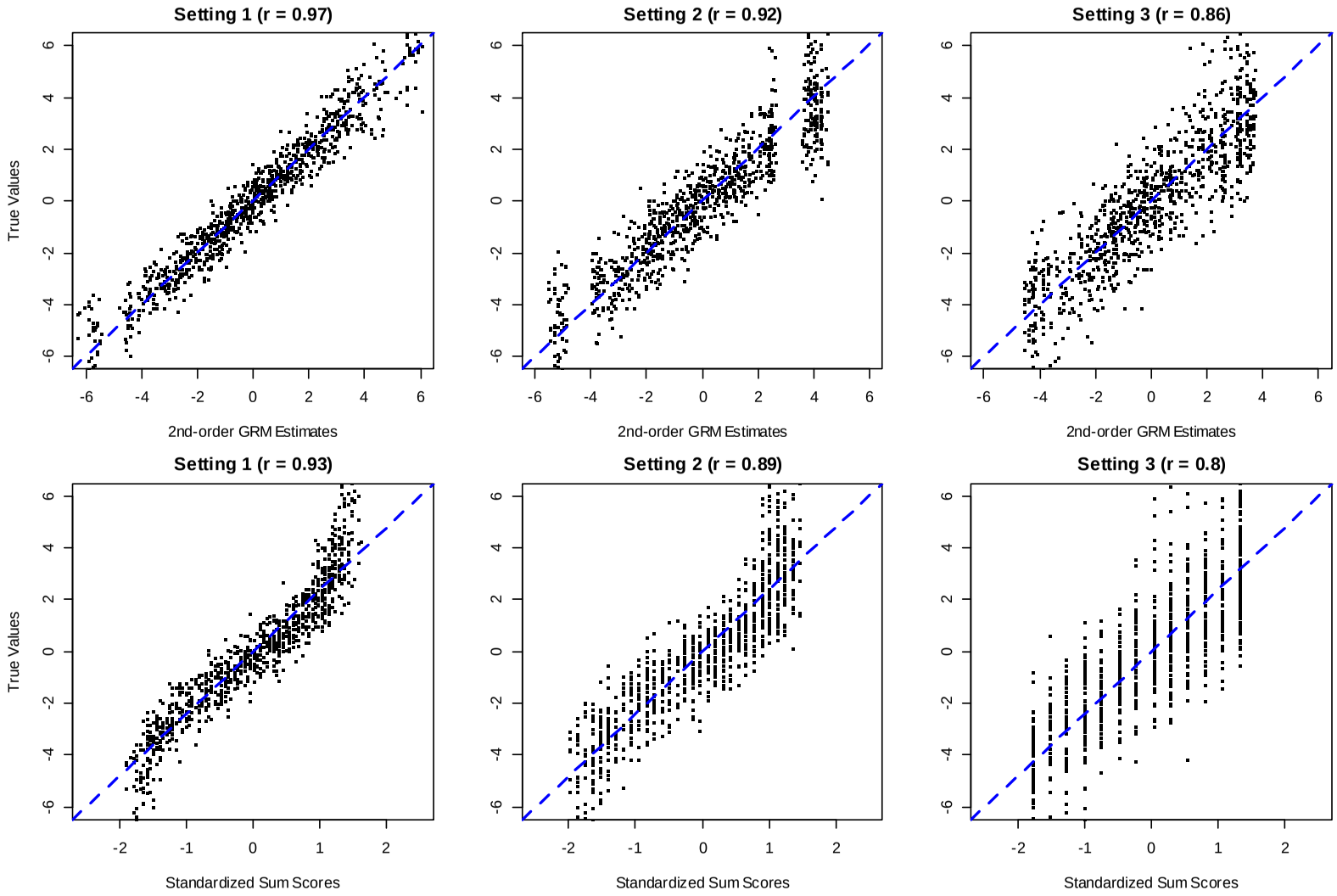


Figure 4-2. Bivariate relationships between the true proficiency values (on the y-axes) and the estimated values (on the x-axes). The first row presents the estimates from the second-order GRM on the x-axes, while the second row presents the standardized sum scores on the x-axes.

The scatter plots for the second-order GRM estimates illustrate that the model functioned well except for extreme proficiency values, which is expected from an IRT model. The scatter plots involving the standardized sum scores show severe nonlinearity at the extreme proficiency values across all three settings. Furthermore, the standardized sum scores were close to discrete variables in Setting 3. This tendency of generating discrete scores can be problematic when the goal of score estimation involves clustering of the estimated scores, which was the case in this study. The second-order GRM estimates did not suffer from this issue, making the model-based proficiency estimates more desirable for this study. In sum, the correlations in Table 4-7 and the scatter plots in Figure 4-2 illustrate the advantage of using the second-order GRM over a simple average scoring method in the context of this study.

It should be noted that the simulation study reported in this section was limited in that it only investigated a single case from each setting. More extensive simulation studies are needed to rigorously evaluate the performance of the second-order GRM under various settings. However, despite this limitation, the results were encouraging. The second-order GRM successfully recovered the true parameters under all three settings. The 95 percent Bayesian credible intervals estimated by the model always contained the true parameter values. This was the case even under the most challenging setting (i.e., Setting 3) in which there were only two tasks and a large proportion of missing data were present. Setting 3 was designed to be highly comparable to the TOP dataset, and the performance of the second-order GRM in Setting 3 provided a justification for its use with the TOP dataset.

4.1.3 Summary of the section

This section reported the results of a simulation study that evaluated the parameter recovery of the second-order GRM. Three simulation settings were specified to present the characteristics of typical language performance assessment contexts. The settings presented varying levels of challenge to the model, and the most challenging setting, namely Setting 3, was designed to be comparable to the TOP dataset in that they both shared the same number of tasks and observed rater scores, as well as the same missing data mechanism. Across all three settings, model parameters were estimated using MCMC as the main estimation method. The starting values of the MCMC estimation were obtained using ML estimates and standard errors. The simulation results showed that the second-order GRM successfully recovered all rater parameters and the test taker proficiency variances across all three settings. In addition, in each setting, the second-order GRM generated test taker proficiency estimates that were highly correlated with the true values. The added value of using the second-order GRM was evaluated by comparing the correlation between the true values and the model estimates to the correlation between the true values and the standardized sum scores. The second-order GRM estimates were more highly correlated with the true values than the standardized sum scores were across all three settings. Overall, the results of the simulation study showed that the second-order GRM was capable of recovering its parameters in practical settings, and that the use of the second-order GRM had a clear advantage over a simple average scoring method in estimating the model-based subscale scores.

4.2 Estimation of the model-based subscale scores

Given the positive results of the simulation study, the model-based subscale scores of the 960 TOP test takers were estimated using the second-order GRM. The model estimation involved two approaches: maximum likelihood (ML) with the reparameterization described in Chapter 3 and MCMC based on a full Bayesian framework. The latter was the main estimation approach for the model-based subscale scores. The ML approach was used to obtain sensible starting values for the MCMC estimation and to facilitate a better understanding of the target distributions of model parameters. The choice of adopting the MCMC approach was based on practical grounds. Due to the incomplete block design of the TOP rater assignment rule, it was necessary to estimate the rater parameters based on a relatively small sample. The lowest cutoff parameters were particularly problematic considering the rare use of the lowest score by the raters, as depicted in Figures 3-1 and 3-2 in Chapter 3. However, the distinction between the lowest and the second lowest scores was important. The lowest score indicates a severe communication failure, whereas the second lowest score implies difficulties in communication. Therefore, estimating the lowest cutoff parameters, as well as distinguishing the two lowest scores, was of substantive interest. A Bayesian framework can be used to regularize parameter estimation through the specification of prior distributions, and therefore, was preferred over the ML approach which can fail to estimate parameters under the lack of data.

This section reports the results of the model-based subscale score estimation. The organization of this section follows the order of analysis. First, the procedures of the ML approach are described. The main MCMC estimation procedures are presented next, followed by the report of the resulting estimates.

4.2.1 ML estimation procedures

The purpose of the ML estimation phase was twofold: to obtain sound starting values and to facilitate an understanding of the target distribution for the subsequent MCMC estimation. As with any iterative algorithm, MCMC benefits from good starting values (see, e.g., Spiegelhalter, Best, Gilks, & Inskip, 1996). In addition, it is desirable to achieve a solid understanding of the target distribution before conducting an MCMC analysis (Gelman et al., 2003). ML estimates are recommended as good candidates for sound starting values and as a means to obtain a concrete understanding of the target distribution (Gelman et al., 2003).

Some modeling details were specific to the ML estimation, and therefore, were not shared by the subsequent MCMC estimation. As previously mentioned, the lowest scoring category was seldom used, and not enough data were available to estimate the cutoff that divides the lowest and the second lowest scores. Consequently, the two lowest score categories were collapsed for the ML estimation, resulting in two, instead of three, cutoff parameters per rater. Collapsing the lowest two categories could introduce a downward bias in estimating the variance of model-based subscale scores, since the distinction between the collapsed categories was lost. In addition, the overall rater severity parameter, τ_r, for rater r, involved a reparameterization of the standard GRM, which is difficult to implement with an IRT software package. Therefore, the overall rater severity parameters were absorbed in the corresponding rater cutoff parameters. The resulting parameterization for the ML estimation presented a form highly comparable to a bifactor GRM (Cai, Yang, & Hansen, 2011; Gibbons et al., 2007).

ML estimation, in general, does not require prior distributions for model parameters. However, it is possible to specify prior distributions for all or a selected subset of parameters to

help stabilize estimation when parameters are estimated based on a relatively small sample. As noted in Chapter 3, when viewed from a second-order GRM perspective, the TOP dataset contained an overwhelming amount of missing data due to its incomplete block design. Consequently, rater parameters were estimated based on sparse data. Prior distributions for rater discrimination parameters were specified to ameliorate this issue. Since the discrimination parameter is seldom negative, a prior distribution that is frequently used for this in the IRT literature is a lognormal distribution (e.g., Patz & Junker, 1999). Following this convention, all rater discrimination parameters were given a lognormal prior with mean equal to zero and standard deviation equal to 0.25.

Additional constraints were imposed to identify the model. Since the TOP consists of two tasks, the model-based subscale score of test taker i (i.e., θ_i) is a second order factor based on two first order factors (i.e., ξ_{i1} and ξ_{i2}). Therefore, an equality constraint was needed on the loadings from the second-order factor to the first-order factors (i.e., $\lambda_1 = \lambda_2$). In addition, both loadings were set to be equal to one. When translated into the standard bifactor model terminology, these conditions implied that the slopes on the general factor and the specific factors were all set to be equal. This can be easily shown by inspecting the equation (10) in Chapter 3. In particular, when $\lambda_j = 1$, $a_r \lambda_j = a_r$. In addition, the variances of task-specific errors (i.e., diagonal elements of $\boldsymbol{\Psi}$) were fixed at one. The above conditions allowed the variance of model-based subscale scores (i.e., σ_θ^2) to be freely estimated and interpreted in a straightforward manner. In particular, the resulting model-based subscale score variance reflected its relative size compared to the task-specific error variances. The flexMIRT syntax implementing the aforementioned setup is given in Appendix C.

The second-order GRM introduced in this study is a univariate model in that it involves only one proficiency factor (in this study, the model-based subscale score). Since the TOP employs an analytic scoring scheme consisting of four subscales, four parallel analyses were conducted. That is, each scale was investigated using a univariate model. The aforementioned modeling details were universally applied to all four parallel analyses. It is theoretically possible to model the TOP analytic scoring scheme as a multivariate second-order GRM. The resulting multivariate model can be reparameterized into a two-tier model (Cai, 2010; Cai, Yang, & Hansen, 2011) with appropriate restrictions. Fitting this multivariate second-order GRM model to the TOP dataset would have been more desirable in theory than running four parallel univariate models.

The multivariate second-order GRM, however, would involve heavy computation even with the two-tier dimension reduction. For the TOP dataset, it involves as many as four general level factors. In addition, such a multivariate model cannot be built without a thorough understanding of rater behaviors under analytic scoring contexts. For example, little is known about the extent to which raters tend to remain equivalently harsh or lenient across different subscales. A similar point can be made for rater discrimination across different subscales. The lack of relevant substantive knowledge made it difficult to propose a complicated model such as the multivariate second-order GRM. Therefore, in this study, it was decided to run four parallel univariate analyses. Nevertheless, a multivariate version of rater models for analytic scoring schemes presents a promising line of research considering the popularity of analytic scoring schemes in the language testing field (Luoma, 2004; Weigle, 2002).

4.2.2 MCMC estimation of model-based subscale scores

This part of the study adopted a full Bayesian framework and utilized MCMC as a means to estimate the model-based subscale scores of the 960 TOP test takers. The starting values of Markov chains were obtained based on the ML estimates. The modeling details for the ML estimation largely remained unchanged except for prior distributions and identification conditions.

4.2.2.1 Prior distributions

The specification of prior distributions is an integral part of Bayesian modeling. Prior distributions reflect the nature and amount of prior knowledge available for model parameters; strong prior information is incorporated through the use of a strong prior distribution, while weak prior distributions are often employed to indicate the lack of available information (Gelman et al., 2003; Gill, 2007). In general, little was known about the parameters of the second-order GRM in the TOP setting. While the results of the ML estimation were available, it would have been premature to specify strong prior distributions based solely on these. Furthermore, information about rater parameters was particularly difficult to obtain a priori. For example, it was not clear how to gauge a rater's discrimination power before fitting an IRT model. Consequently, it was decided to specify relatively vague prior distributions for the second-order GRM parameters.

In the IRT literature, it is standard procedure to set conjugate priors when they are reasonable (see, e.g., Fox, J-P, 2010; Patz & Junker, 1999). This study followed this in adopting conjugate priors where feasible. In particular, prior distributions for rater cutoff (i.e., γ_{rk}), overall rater severity (i.e., τ_r), task-specific error (i.e., u_{ij}), and subscale score (i.e., θ_i) parameters were all specified as normal distributions. However, a normal distribution was not appropriate for the

rater discrimination parameter (i.e., a_r) because discrimination is seldom negative. Therefore, this study specified a normal distribution truncated below zero as the prior distribution of rater discrimination parameters.[4] Lastly, the variance of model-based subscale scores (i.e., σ_θ^2) was given an Inverse-Gamma prior distribution. To reflect the lack of prior knowledge for the parameters, the prior normal distributions were given large variances. The Inverse-Gamma prior distribution for the model-based score variance was also intentionally vague. The prior distributions of the model parameters are summarized in Table 4-8.

Table 4-8

Prior Distributions for the Second-order GRM Parameters

Parameter	Prior Distribution
Discrimination (a_r)	$N(0, 10)$, truncated below 0
Severity (τ_r)	$N(0, 10)$
Cutoff (γ_{rk})	$N(0, 4)$
Model-based score (θ_i)	$N(0, \sigma_\theta^2)$
Model-based score variance (σ_θ^2)	Inverse-Gamma(1,1)
Task-specific error (u_{ij})	$N(0, 1)$

It is noteworthy that the MCMC estimation utilized a different approach to model identification from that used for the ML estimation. The ML approach involved fixing the variances of task-specific error terms at one. The MCMC approach, on the other hand, achieved the same goal through the prior distribution of u_{ij}, as can be seen in Table 4-8. In particular, the prior distribution of u_{ij} had a fixed variance equal to one. The difference between the variance of task-specific error terms and the variance of model-based subscale scores is also noteworthy. While the former was fixed a priori, the latter was freely estimated based on the Inverse-Gamma

[4] Another popular choice for the discrimination parameter in the literature is a lognormal distribution (e.g., Patz & Junker, 1996). A later trial with a lognormal prior on the rater discrimination parameters suggested that the impact of a choice between the two prior distributions was trivial.

prior and the data, relative to the size of the former. This is analogous to their relationship in the ML estimation context.

4.2.2.2 Markov chain construction and run

The model likelihood specified in Chapter 3 and the prior distributions in Table 4-8 were used to construct Markov chains for the parameter estimation. The aforementioned modeling details were directly translated into WinBUGS 1.4.3 (Lunn et al., 2000). The resulting WinBUGS syntax used to construct Markov chains is provided in Appendix D.

Monitoring and evaluating the convergence of Markov chains is one of the most important and difficult parts of any MCMC application (Cowles & Carlin, 1996). Researchers have proposed several analytical approaches to convergence checking (e.g., Rosenthal, 1993, 1995a, 1995b), as well as convergence diagnostic measures (e.g., Gelman & Rubin, 1992; Geweke, 1992; Raftery & Lewis, 1992). While the analytical approaches are more desirable in theory, it is often difficult to adopt them in an applied setting. The convergence diagnostic measures are more straightforward to implement. When the evaluation of convergence relies on diagnostic measures, it is recommended to base the evaluation on multiple measures (Cowles & Carlin, 1996; Gelman & Shirley, 2011). In this study, the convergence of chains was monitored using the shrinkage factor (Gelman & Rubin, 1992). Other diagnostic measures suggested by Geweke (1992) and Raftery and Lewis (1992) were also examined.

The shrinkage factor (Gelman & Rubin, 1992) requires running multiple chains in parallel since it is calculated based on an ANOVA-like comparison of within- and between-chain variances. Running multiple chains has other benefits, including the mitigation of concerns about dependence between consecutive draws (i.e., autocorrelation). Large autocorrelation is

problematic since it reduces the number of effective sample draws from which inferences can be made. However, if draws from multiple chains are randomly mixed, the concern for high autocorrelation is ameliorated (Gelman, 1996; Gelman & Shirley, 2011). The number of parallel chains was determined by both theoretical and practical considerations. Theoretically, the more chains the better. However, more chains means more computing time, and it is often advised that three to five chains are sufficient for most practical applications (Gelman et al., 2003). Therefore, in this study, it was decided to run five parallel chains per subscale.

It is of substantial importance to arrange Markov chains such that they could travel through the entire joint posterior distribution of model parameters. Gelman and Rubin (1992) suggest a two-step process in which one first establishes an over-dispersed version of the target distribution, and then draws starting values from the over-dispersed distribution. Their approach, however, can be difficult to implement in practice, and researchers instead have relied upon different sets of starting values that are well separated from one another to ensure the appropriate coverage of the target distribution (Gelman & Shirley, 2011). This study used the point estimates and the standard errors from the ML estimation to obtain sets of starting values that were distant from each other. In particular, five sets of starting values, one for each of the five parallel chains, were obtained in the following manner. Chain 1 starting values were equal to the ML point estimates. Chain 2 and 3 starting values were equal to the ML point estimates plus and minus, respectively, 1.5 times the corresponding standard errors. Similarly, the ML point estimates plus and minus three times the corresponding standard errors comprised of Chain 4 and 5 starting values, respectively. There were two sets of parameters in the MCMC setup that did not have a direct counterpart in the ML estimation, namely rater cutoff and severity parameters. For those two parameters, randomly generated starting values were used. Lastly, the missing values in the

observed rater scores were given the rounded averages of available rater scores as their starting values.

Each chain was run for 15,000 iterations. The coda package (Plummer, Best, Cowles, & Vines, 2006) in R (R Core Team, 2012) was used to monitor the convergence of chains. All chains appeared to converge after approximately 5,000 draws. In particular, the Gelman-Rubin shrinkage factor for each and every parameter became smaller than the usual cutoff of 1.1 after the first 5,000 draws. Other diagnostic measures also agreed that the chains had reached convergence after 5,000 draws. Consequently, the first 5,000 draws were discarded as "burn-in" samples. This left 10,000 effective sample draws per chain. The five parallel chains were then combined, resulting in a total of 50,000 sample draws. All 50,000 draws were used to summarize the joint posterior distributions of model parameters.

4.2.2.3 MCMC estimation results

It is of little use to interpret estimates from a model that does not fit the data well. The model-based subscale scores were dependent upon the second-order GRM, and thus the model goodness of fit was important. Unfortunately, however, in Bayesian modeling a convenient absolute fit statistic such as the limited information M_2 statistic (Cai & Hansen, 2012; Maydeu-Olivares & Joe, 2005) is not available. Consequently, the fit of the second-order GRM was instead evaluated by comparing it to the fit of a similar model. In particular, a hierarchical rater model (HRM; Patz, Junker, Johnson, & Mariano, 2002) was selected for the comparison due to its similarity to the second-order GRM in terms of model parameters. The fit of the second-order GRM and the HRM to the TOP dataset were compared using a set of relative model fit indices.

The second-order GRM and the HRM are not nested models. That is, one cannot impose a set of constraints to obtain the second-order GRM from the HRM or vice versa. Therefore, the usual chi-square difference test was not appropriate in comparing the goodness of fit of the two models. The comparison of model goodness of fit was instead conducted based on the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwartz, 1978), both of which can be employed to compare non-nested models. A model with the smallest AIC or BIC is preferred. Both AIC and BIC are calculated based on the maximized log-likelihood given a model and the number of free parameters of the model. In particular, let L_{max} be the maximized log-likelihood, p be the number of free parameters, and n be the sample size. AIC and BIC are obtained as

$$AIC = -2\,L_{max} + 2p;$$

$$BIC = -2\,L_{max} + p\log n.$$

The above equations show that both AIC and BIC consist of two parts. The first term on the right hand side represents the model goodness of fit, while the second term penalizes model complexity. The maximized log-likelihood of the HRM was obtained by fitting the HRM to the TOP dataset. The HRM parameterization in this study was the same as described in Patz et al. (2002). Three parallel chains were constructed for each subscale HRM run. Each chain was run for 15,000 iterations, and the first 10,000 iterations were discarded as burn-in. Table 4-9 gives minus two log-likelihood, the number of free parameters, AIC, and BIC of the second-order GRM and the HRM for each subscale.

Table 4-9

Model Fit Comparison between the Second-order GRM and the HRM

Subscale	Model	# Parameters	$-2\,L_{max}$	AIC	BIC
PR	2nd-order GRM	72	3216.88	**3360.88**	3711.30
	HRM	43	3306.96	3392.96	**3602.24**
LG	2nd-order GRM	72	3509.27	**3653.27**	**4003.69**
	HRM	43	3887.66	3973.66	4182.94
RO	2nd-order GRM	72	3902.34	**4046.34**	**4396.76**
	HRM	43	4120.77	4206.77	4416.05
QH	2nd-order GRM	72	3536.06	**3680.06**	**4030.48**
	HRM	43	3862.88	3948.88	4158.16

Note. Boldfaced entries represent the preferred model.

Table 4-9 shows that the second-order GRM was the better fitting model across all subscales, as indicated by the smaller minus two times log-likelihood values. The second-order GRM, however, was the more complicated model in that it had more free parameters than the HRM. Both AIC and BIC contain a penalty term for model complexity measured by the number of free parameters. Since log*(960)* is larger than 2, BIC imposed a heavier penalty on the second-order GRM in the TOP dataset. Despite the penalty, however, the second-order GRM was the preferred model by both AIC and BIC for Lexical Grammar, Rhetorical Organization, and Question Handling. AIC and BIC yielded mixed results for Pronunciation. The second-order GRM was preferred by the former, whereas the HRM was preferred by the latter. However, it should be noted that the HRM was preferred by BIC for Pronunciation due to its simplicity, not because of better fit. In sum, the second-order GRM was superior to the HRM in terms of model fit across all subscales, albeit with added complexity.

The TOP dataset consisted of 960 test takers' TOP subscale scores given by 18 raters. Consequently, the second-order GRM yielded the estimates of 18 rater discrimination parameters,

17 rater severity parameters,[5] and 36 rater cutoff parameters[6] per subscale. All rater parameter estimates were within a reasonable range. Since the rater parameters were not of central interest in this study, the resulting estimates are not reported or interpreted here. Instead, the rater parameter estimates and the corresponding standard errors are provided in Appendix E.

The model also estimated the variance of model-based subscale scores for each subscale. Recall that every task-specific error variance, as well as every loading, was fixed at one for identification. The estimated subscale score variances can be interpreted in relation to the task-specific error variances, which were all equal to one. For example, if the resulting estimate of a subscale score variance is 20, the estimated model-based subscale score variance is twenty times larger than the task-specific error variance for that scale. Table 4-10 gives the resulting estimates of model-based score variances and the corresponding 95% Bayesian credible intervals. As previously mentioned, WinBUGS parameterizes precision (i.e., the reciprocal of variance) for normal distributions. Therefore, the reciprocal of the estimated variance of each subscale is given in the first row of Table 4-10 with the corresponding intervals. The second row of Table 4-10 presents the point estimates on the original variance scale.

Table 4-10

Model-based Subscale Score Variance Estimates

	PR	LG	RO	QH
$1/\sigma_\theta^2$	0.029 (0.017, 0.042)	0.048 (0.034, 0.068)	0.165 (0.110, 0.230)	0.079 (0.054, 0.114)
σ_θ^2	34.98	20.68	6.05	12.58

Note. The 95 percent Bayesian credible intervals are presented in the parentheses.

[5] The first rater's severity was fixed at zero for identification.
[6] Two cutoff parameters were estimated for each of 18 raters due to the sum-to-zero constraint on the third cutoff.

Table 4-10 shows that the magnitude of score variance relative to the task-specific error variance differed across the four subscales. Pronunciation showed the largest variance, implying the smallest impact due to task-specific fluctuations. The estimated variance for Rhetorical Organization was the smallest at 6.05, which suggested that, once rater effects were accounted for, the ratio of the model-based Rhetorical Organization score variance to the task-specific error variance was approximately 6:1. The difference among the variance estimates was in line with the test design. In particular, given the different organizational nature between the two TOP tasks (a syllabus presentation task and a short lecture task), it was expected that the most notable between-task difference would occur in Rhetorical Organization. On the other hand, it was reasonable to expect to observe little difference in terms of Pronunciation across the two TOP tasks.

The model-based subscale scores were calculated based on the rater parameters and the subscale variance parameters. The resulting scores were compared to observed sum scores, which were calculated by simply adding the two observed rater scores together. The distributions of the model-based subscale scores and the observed sum scores are presented in the first and second rows, respectively, of Figure 4-3.

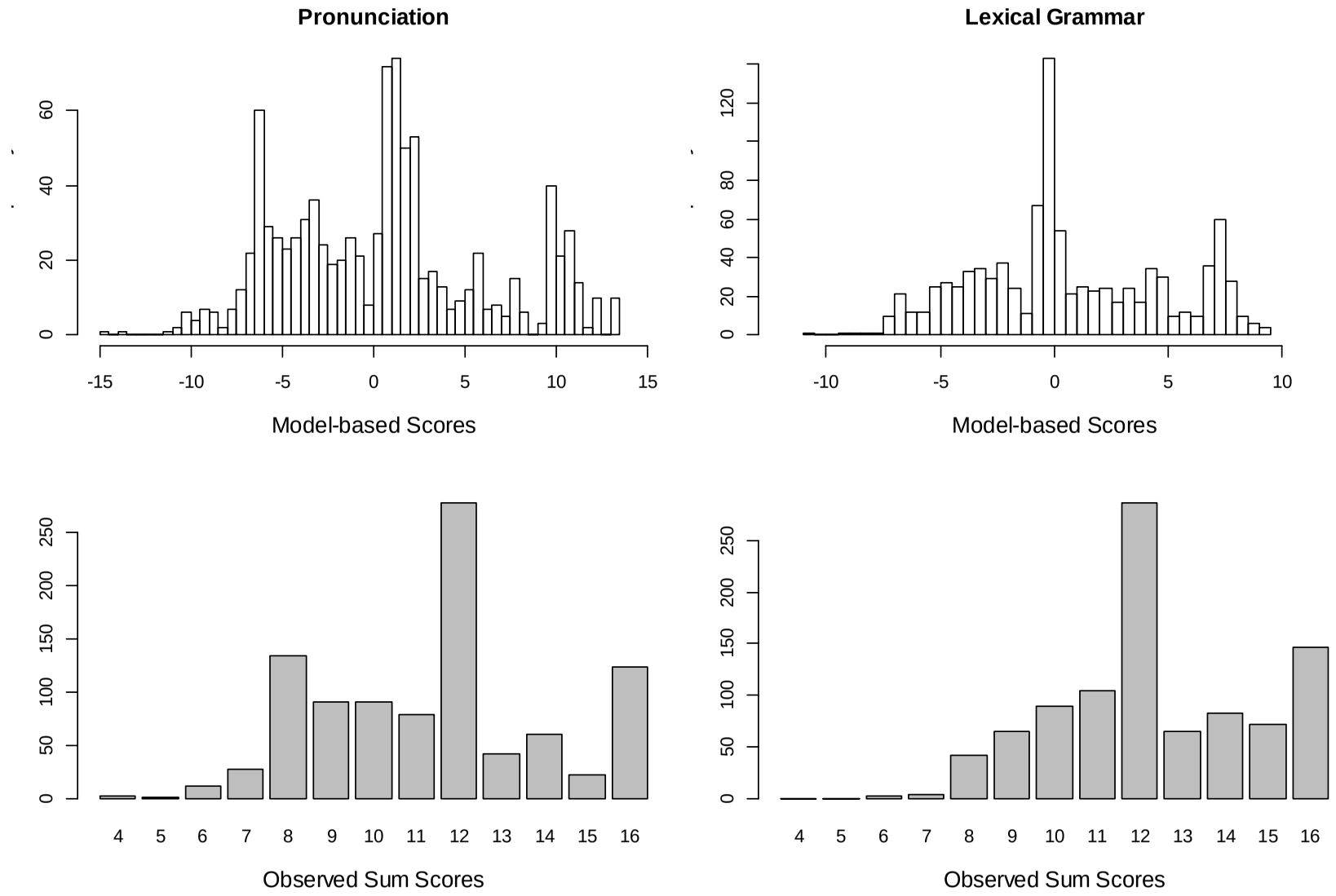


Figure 4-3. The distributions of the model-based subscale scores (in the first row) and the observed sum scores (in the second row).

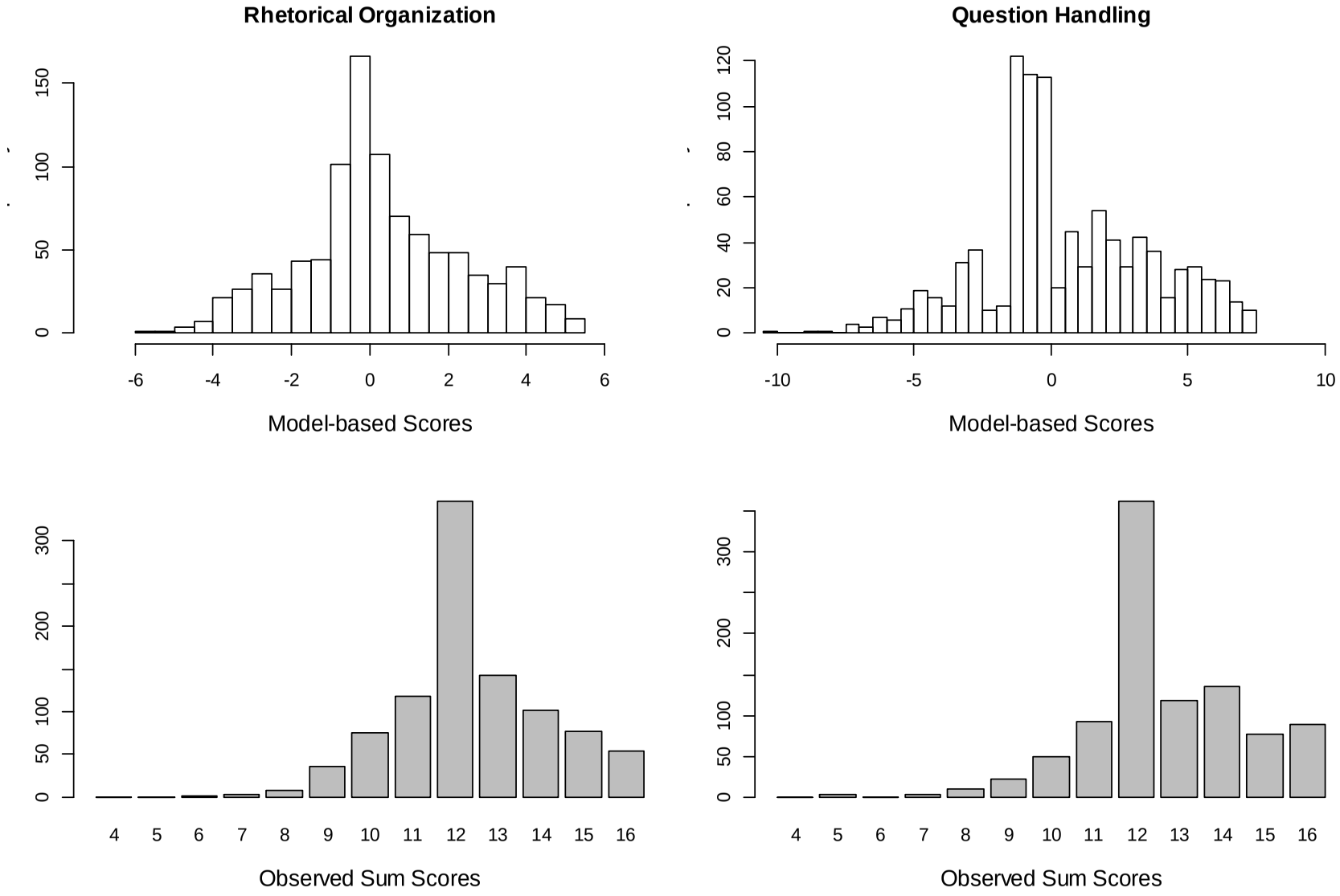


Figure 4-3. continued

Figure 4-3 shows that the model-based subscale scores successfully captured the distributional characteristics of the observed sum scores. In particular, multiple modes in the sum score distributions were retained by the corresponding model-based score distributions. Furthermore, by virtue of being continuous, the model-based subscale scores ensured more fine-tuned distinction among test takers than the observed sum scores. The continuous nature of the model-based subscale scores was particularly well-suited for a clustering technique to explore underlying group structures in terms of subscale score patterns. Indeed, the fine-tuned distinction available with continuous variables was one of the main motivations for the model-based subscale score estimation in this study.

Bivariate relationships among the four subscales were examined using Pearson product-moment correlation coefficients. The correlation coefficients among the model-based subscale scores are given in the upper diagonals of Figure 4-4. Figure 4-4 also presents the corresponding scatter plots in the lower diagonals.

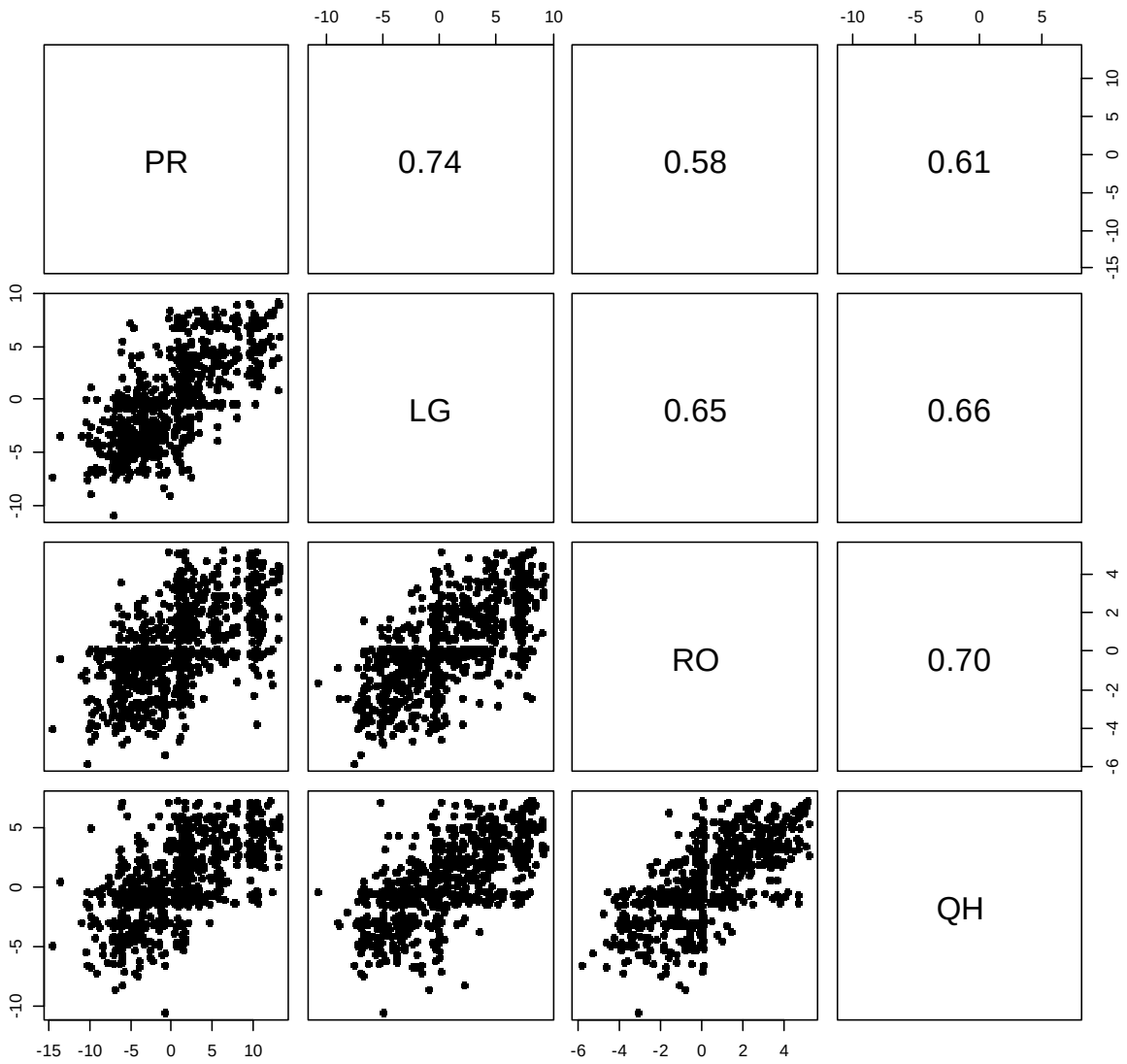


Figure 4-4. Scatter plots and correlations among the TOP subscales

A few outliers can be found in each plot, but the proportion of the outliers was not large enough to affect the overall correlations. Furthermore, the scatter plots show no sign of curvilinear relationships among the subscale scores. All subscales were highly correlated with one another. Pronunciation and Rhetorical Organization were the least correlated subscales, but even their correlation was as high as 0.58. The highest correlation, which was approximately 0.74, was observed between Pronunciation and Lexical Grammar.

4.2.3 Summary of the section

This section reported the procedures and the results of the model-based subscale score estimation. The model-based scores were estimated using the second-order GRM with MCMC as the main estimation method. Five chains were constructed for each of the four TOP subscales. Each chain was run for 15,000 iterations from the starting values that were obtained based on ML estimates and standard errors. Multiple convergence diagnostic indices suggested that the chains converged after 5,000 iterations. Consequently, the first 5,000 draws were discarded as the burn-in samples and the remaining draws were used to summarize the joint posterior distribution of model parameters. The goodness of fit of the second-order GRM was evaluated in relation to the goodness of the fit of a similar model, namely the HRM (Patz et al., 2002). The model fit comparison between the second-order GRM and the HRM showed that the former model fit the data better than the latter model across all subscales, albeit with more free parameters. The distributions of the estimated model-based subscale scores were compared to the observed sum score distributions. The comparison showed that the two sets of distributions were highly comparable across all subscales. The key difference between them was the continuous nature of the estimated model-based scores. Obtaining continuous scores was the main motivation of the use of the second-order GRM in that these provided more fine-tuned distinctions among test takers. The model-based scores were highly correlated across all subscales.

4.3 Clustering of TOP subscale score patterns

The second-order GRM fitted to the TOP dataset yielded four model-based subscale scores, one per subscale, for each of the 960 test takers. The resulting model-based scores were plugged

in for use as the data for the subsequent mixture analysis. That is, the estimated model-based subscale score for test taker *i* on subscale *h* was used as the (*i, h*)th element of the mixture model data matrix **X** presented in Chapter 3.

The plug-in procedure adopted in this study (i.e., using the estimates of the second-order GRM as the data for the mixture analysis) is not the most desirable approach from a purely theoretical point of view. The main caveat of a plug-in approach is that it ignores the uncertainty involved in plugged-in estimates and treats the estimates as given values. A direct approach that can acknowledge the uncertainty inherent in estimates is more desirable. In this study, such a direct approach would amount to a mixture of second-order GRMs. However, mixtures of IRT models have been largely limited to the Rasch model family (see, e.g., Cho, Cohen, & Kim, 2013; Rost, 1990; Rost & von Davier, 1995) due to technical difficulties. A mixture model using the second-order GRM as a component distribution is not currently feasible. The decision to rely on the plug-in procedure, instead of adopting the direct approach, was made based on this practical consideration.

This section reports the results from the mixture model analysis for investigating subscale score profile groups. The results are presented in the following order. First, the employed mixture models are described in detail. Model selection procedures are reported next. Lastly, the group structures suggested by the selected model are interpreted.

4.3.1 Mixture component distributions

Mixture component distributions were selected based on the empirical distributions of the model-based subscale scores and the structure of the TOP. A visual inspection of the model-based score distributions in Figure 4-3 suggested that each subscale distribution could be

modeled as mixtures of normal distributions. Normal component distributions also have practical advantages such as ease of computation and interpretation. The dimensionality of component distributions was determined to reflect the test design. In particular, since the TOP consists of four subscales, it was natural to consider four dimensional distributions as mixture components. The examination of the correlation structure among the subscales presented in Figure 4-4 did not suggest any need to collapse the existing dimensions, which also supported the choice of four-dimensional component distributions. Consequently, four-dimensional normal distributions were selected as mixture component distributions.

It is possible to enhance computations involved in model estimation by imposing restrictions on the parameters of component distributions. However, such restrictions should be made with caution. It is well known that restrictions on component covariance matrices could negatively affect the performance of mixture models even when those restrictions are not completely unreasonable (McLachlan & Peel, 2000). Since this study adopted a standard mixture model with normal component distributions, the associated computational burden was of little concern. Furthermore, the subscale covariance structure in Figure 4-4 did not provide any justification for component parameter restrictions. For example, the considerable correlations among all subscales prohibited restricting the component covariance matrices to be diagonal matrices. Consequently, no restriction was made on the parameters of component distributions. That is, the mean vectors and the covariance matrices of the four-dimensional normal component distributions were all freely estimated.

4.3.2 Model selection

Multiple mixture models were fitted to the model-based subscale score. Since the type of component distributions was fixed (i.e., four dimensional normal distributions), the focus of model fitting was to find the number of underlying subscale profile groups that could adequately account for the data. Both statistical and substantive criteria were considered in the process of model selection. The statistical goodness of fit was evaluated using BIC. BIC has been successfully applied to mixture modeling for density estimation (see, e.g., Roeder & Wasserman, 1997; Solka, Wegman, Priebe, Poston, & Rogers, 1998). Furthermore, simulation results suggest that BIC performs better in finding the number of components than other indices including AIC (see, e.g., Cho et al., 2013; Dasgpta & Raftery, 1998; McLachlan & Peel, 2000). The generalized variances of component distributions were also monitored. In particular, solutions involving a component distribution with extremely small generalized variance were carefully investigated to avoid spurious solutions. Substantive interpretability was another important consideration. Given the exploratory nature of this study, the objective was to find a clear solution which could lead to better understanding of the TOP dataset. Consequently, solutions were evaluated in terms of their substantive interpretability.

Figure 4-5 presents the BIC values of the best fitting models conditional on the number of components, which ranges from one to ten. Models with more than ten component groups yielded BIC values less satisfactory than the models included in the figure.

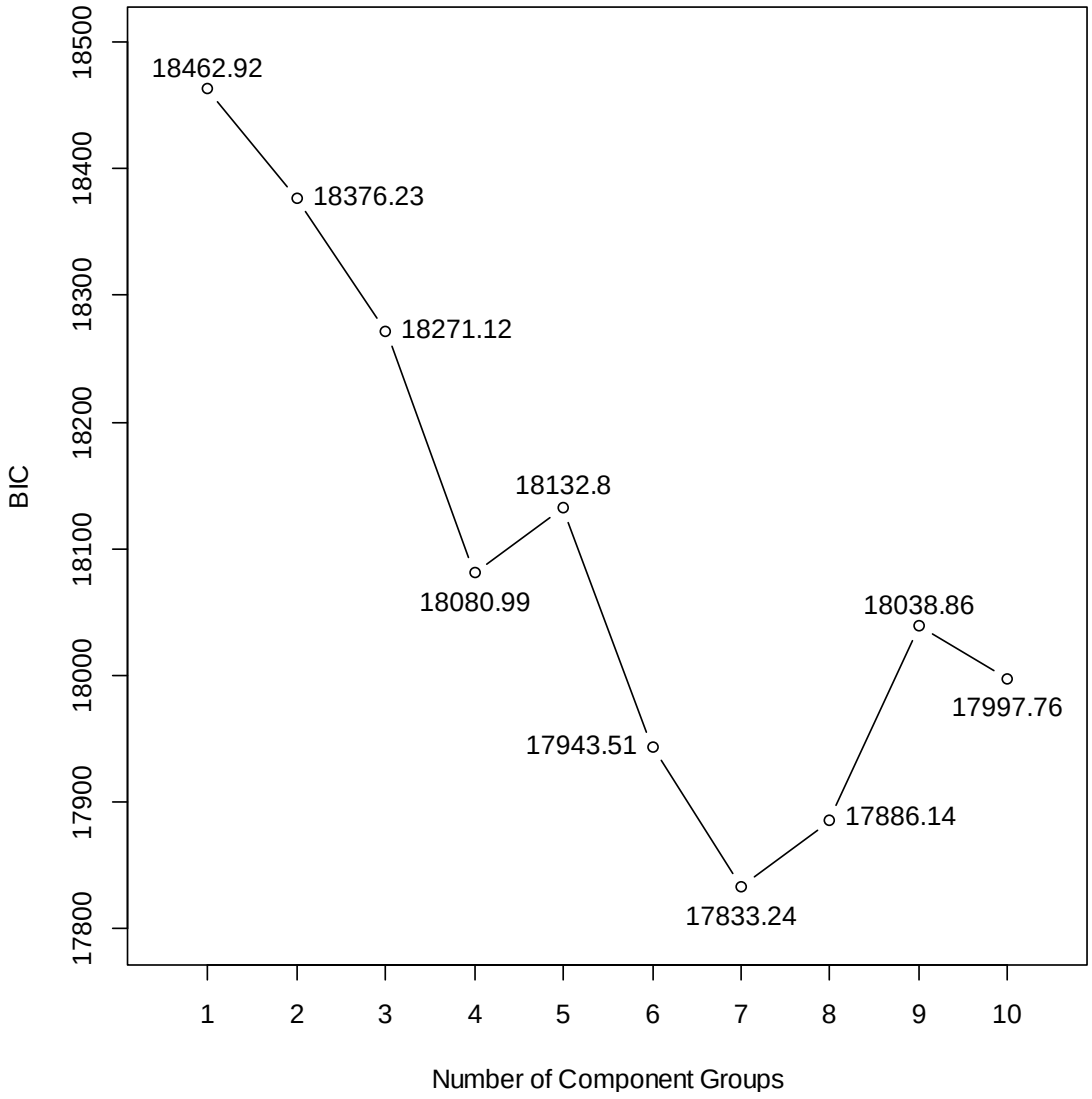


Figure 4-5. BIC of the best solutions given the number of component groups (from one to 10)

Figure 4-5 suggests that there were three modes within the range, and that the best fitting model corresponded to the seven group solution. Solutions corresponding to the three modes, namely the four, seven, and ten group solutions were examined in terms of the generalized variances of component distributions as well as interpretability. The ten group solution contained several spurious groups with very small generalized variances, which made it difficult to interpret. Both the four group solution and the seven group solution appeared legitimate in terms of the

generalized variances and interpretability. However, the seven group solution was preferred because of the following two reasons. The seven group solution fit the data much better than the four group solution, as can be seen in Figure 4-5. Furthermore, the four group solution yielded larger uncertainties in classifying the TOP test takers into groups. Consequently, the seven group solution was selected as the final model.

The component parameters of the final model included the mean vector and the covariance matrix of each component distribution, as well as the proportion of each component in the population. Table 4-11 presents the estimates of component-specific mean vectors and proportions. The component-specific covariance matrices are given in Appendix F.

Table 4-11

Mixture Model Parameter Estimates for Component Distributions

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Proportion (%)	10.07	19.22	23.01	10.20	16.73	15.29	5.49
PR Mean	10.72	5.55	-3.65	-5.99	-1.33	1.47	1.15
LG Mean	7.38	4.02	-2.04	-3.99	-0.61	1.33	-0.44
RO Mean	2.60	1.78	-1.16	-1.27	-0.17	1.14	-0.52
QH Mean	3.64	3.36	-0.71	-3.56	-0.63	1.48	-1.05

Table 4-11 shows that the groups were largely comparable in terms of their proportion. The largest group, namely Group 3, comprised slightly less than a quarter of the entire sample, while Group 7 was the smallest group and accounted for approximately five percent. The mean vectors of Groups 1 through 4 indicate a clear hierarchy among the four groups. In particular, Group 1 consisted of test takers of highest model-based scores on all subscales, followed by Group 2. Group 4 test takers in general received the lowest scores on all subscales.

In addition, the selected final model estimated the probability of a test taker's being classified into each of the seven groups. For an illustration, the estimated classification probabilities for the first five test takers are provided in Table 4-12.

Table 4-12

An Excerpt of the Estimated Classification Probabilities

	G1 Prob.	G2 Prob.	G3 Prob.	G4 Prob.	G5 Prob.	G6 Prob.	G7 Prob.
TT #1	0.00	1.00	0.00	0.00	0.00	0.00	0.00
TT #2	0.00	1.00	0.00	0.00	0.00	0.00	0.00
TT #3	0.00	0.02	0.96	0.01	0.00	0.00	0.00
TT #4	0.00	0.00	0.36	0.64	0.00	0.00	0.00
TT #5	0.00	1.00	0.00	0.00	0.00	0.00	0.00
…	…	…	…	…	…	…	…

Note. TT = Test Taker; G = Group; Prob. = Probability of being classified into that group

All 960 test takers were assigned a component group membership according to the estimated classification probabilities. Specifically, a test taker was classified into the group that had the highest classification probability. For example, Test Takers 1, 2, and 5 in Table 4-12 were classified into Group 2, while Test Takers 3 and 4 were classified into Groups 3 and 4, respectively. Almost all classifications were made with a high degree of confidence. In other words, the uncertainty involved with the classification was in general very small. The classification uncertainty can be calculated as the difference between one and the highest classification probability. For example, the classification uncertainty of Test Taker 3 in Table 4-12 was 0.04 (i.e., 1 – 0.96). Figure 4-6 presents the cumulative percentages of the 960 test takers according to the classification uncertainty.

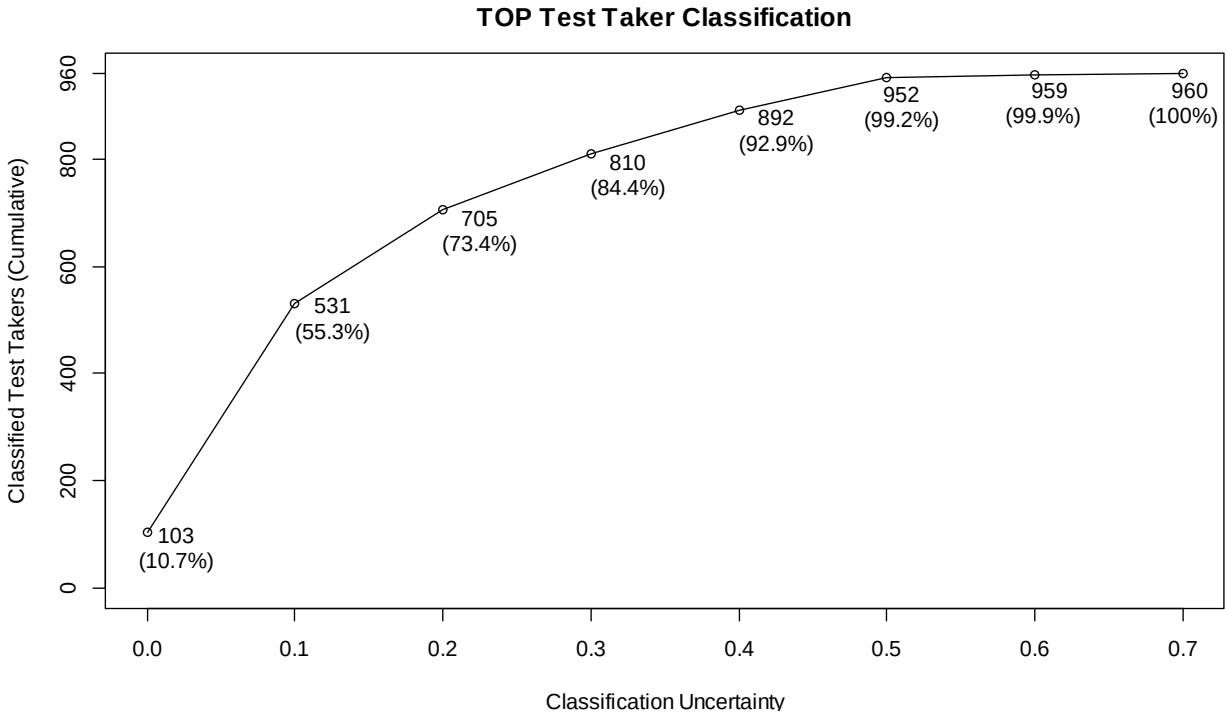


Figure 4-6. The cumulative percentages of the 960 test takers according to the classification uncertainty

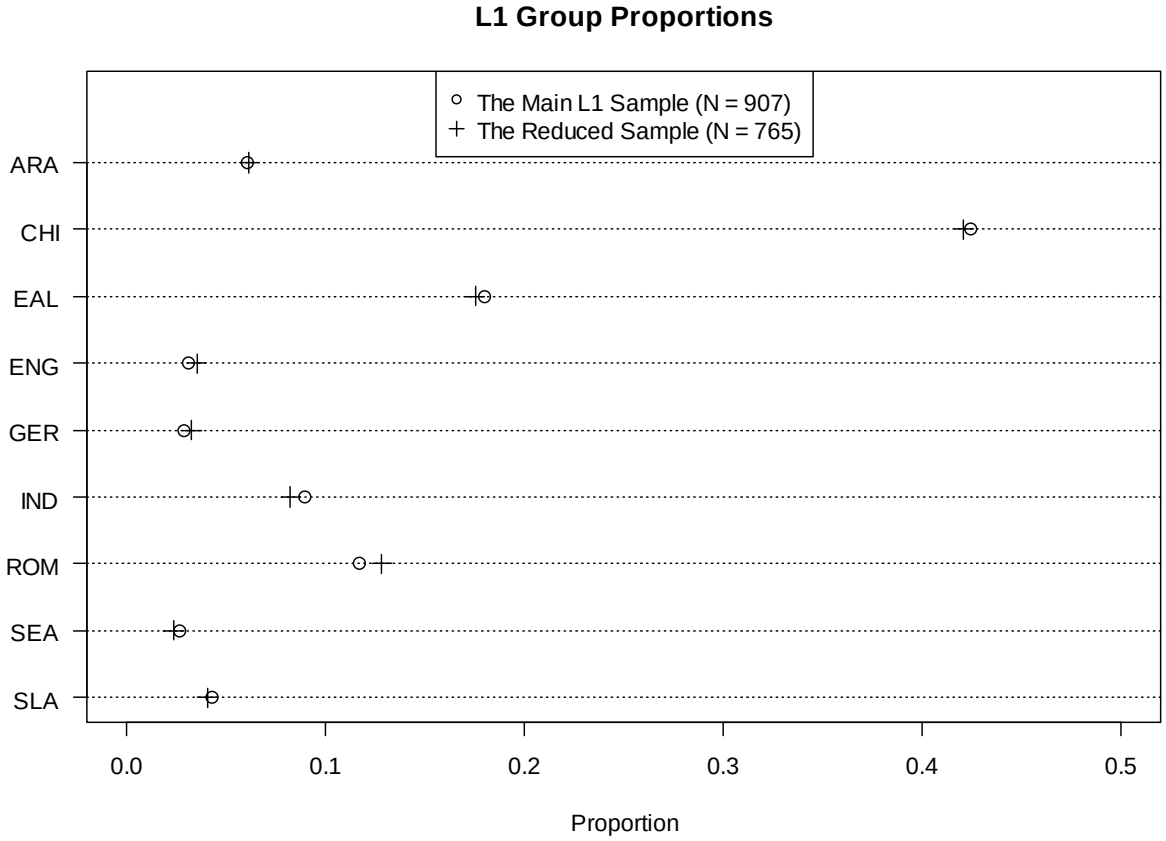
Figure 4-6 shows that the final model classified more than half of the test takers with more than 90 percent probability (or, equivalently, less than 10 percent uncertainty). Furthermore, 810 out of the 960 test takers were given a group membership with classification uncertainties lower than 30 percent. Out of 960, only eight classifications involved classification uncertainties larger than 50 percent.

4.3.3 Interpretation of the mixture model results

The final model classified the 960 TOP test takers into seven groups based on their model-based subscale scores. In order to better understand the characteristics of the groups and to benefit from the modeling results, the members of each group were examined in terms of their background variables pertinent to their academic oral English proficiency. As described in

Chapter 3, the linguistic background (i.e., L1) of the 960 test takers was known. In addition, the decision category of each test taker was retrieved from the TOP database.

To further facilitate the interpretation of the final model, it was decided to focus on test takers who were assigned a component group membership with a high degree of confidence. Consequently, test takers whose classification uncertainty exceeded 30 percent were not considered in interpreting the profile group structure suggested by the final model. The choice of the 30 percent cutoff was driven by practical considerations. The 30 percent cutoff provided a reasonable degree of confidence in the classification results, while retaining approximately 85 percent (810 out of 960) of the entire sample. Furthermore, test takers whose L1 was not specified in the dataset or did not belong to the nine major language groups were excluded from the interpretation. These two filtering procedures left a total of 765 test takers who were assigned a group membership with a high-degree of confidence and whose L1 belonged to one of the nine major language groups. While these two procedures resulted in approximately a 20 percent reduction in the sample size, the remaining sample of 765 test takers did retain major characteristics of the entire TOP dataset. Figure 4-7 gives the proportions of the major language groups and the TOP decision categories in both the entire TOP dataset and the reduced sample of 765.



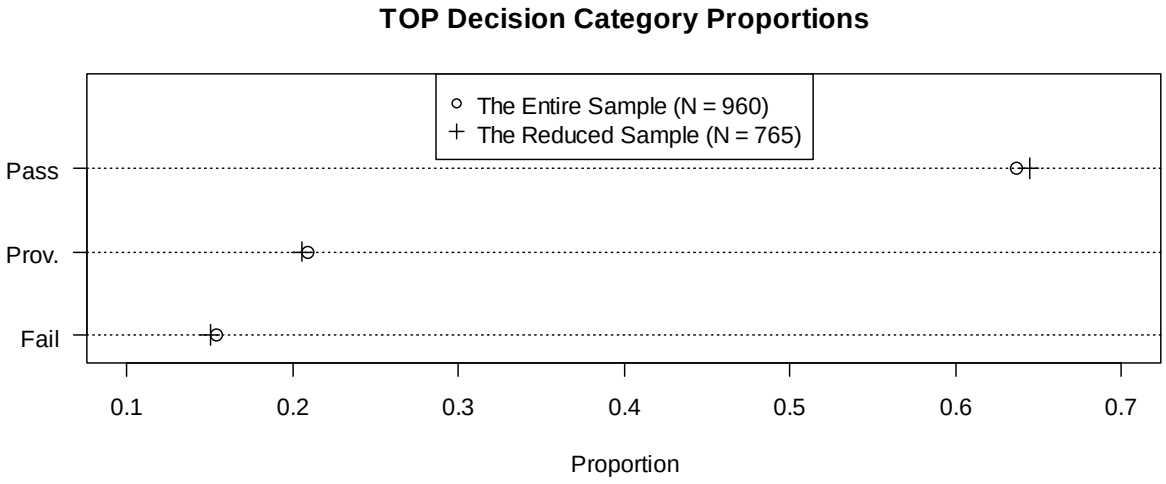


Figure 4-7. The proportions of the first language groups and TOP decision categories in the main L1/entire sample and the reduced sample

The two sets of proportions showed little difference, suggesting that the reduced sample was representative of the entire dataset in terms of the test taker L1 and decision category. This provided a justification for focusing the interpretation of the mixture modeling results on the reduced sample. For simplicity, the reduced sample will be referred to as the baseline sample.

In the remainder of this section, the group structure suggested by the final model is presented in detail. For each of the seven groups, group members' model-based subscale score patterns are presented first. The composition of each group in terms of group members' L1 as well as their decision categories is reported next. Finally, each group is labeled based on its group members' characteristic subscale score patterns and background information.

4.3.3.1 Interpreting mixture model results: Group 1

The first group suggested by the final model consisted of test takers with very high model-based scores on Pronunciation and Lexical Grammar. There were variations in the Rhetorical Organization and Question Handling scores. Figure 4-8 provides bivariate scatter plots that demonstrate the distributional characteristics of Group 1 members' model-based subscale scores. Figure 4-8 also shows the TOP decision category of each test taker in Group 1. The characteristic feature of this group was the concentration of the Pronunciation and Lexical Grammar scores around the highest possible scores, as can be seen in Figure 4-8 (a). Figures 4-8 (b) to (f) illustrate that the variances of Rhetorical Organization and Question Handling accounted for almost all of the within-group variance. However, even the model-based scores on the two latter subscales did not vary widely in that those variations were mostly limited to above-average areas.

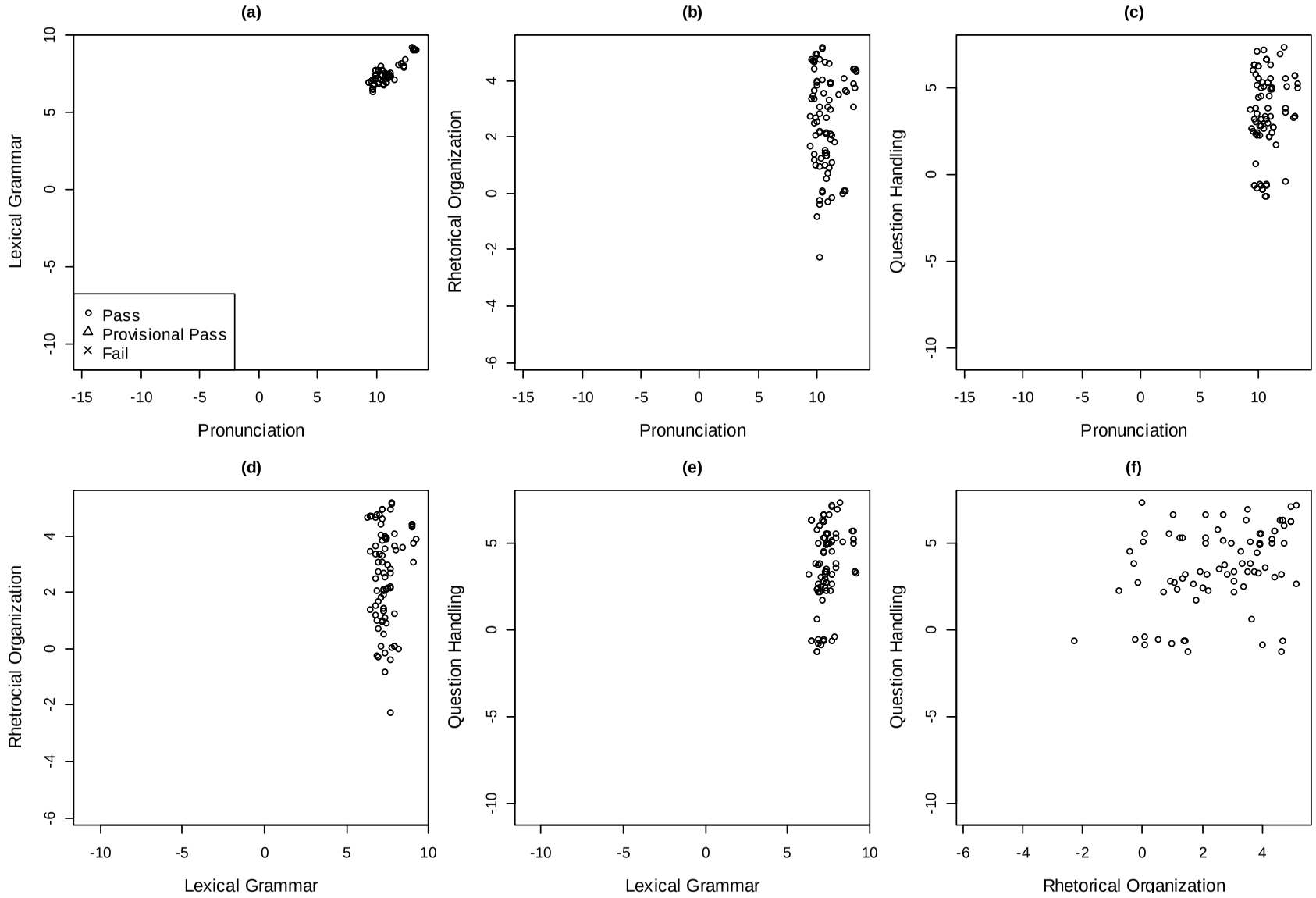


Figure 4-8. Bivariate relationships between the model-based subscale scores (Group 1)

The distributional characteristics of the model-based subscale scores illustrated in Figure 4-8 provided a basis for a clear understanding of Group 1. The characteristic subscale score pattern observed in Group 1 was consistent with a pattern expected from a typical English native speaker. The TOP scoring rubric was designed such that a typical native speaker of English should be able to receive perfect scores on Pronunciation and Lexical Grammar, but not necessarily on Rhetorical Organization and Question Handling. However, a typical English native speaker is likely to score highly on the two latter subscales. All Group 1 members passed the test, which also supports interpreting Group 1 as the group of near native speakers of English.

The linguistic background of Group 1 members was in line with the near-native interpretation. Figure 4-9 presents the proportion of each L1 group in Group 1 as well as the corresponding baseline proportions.

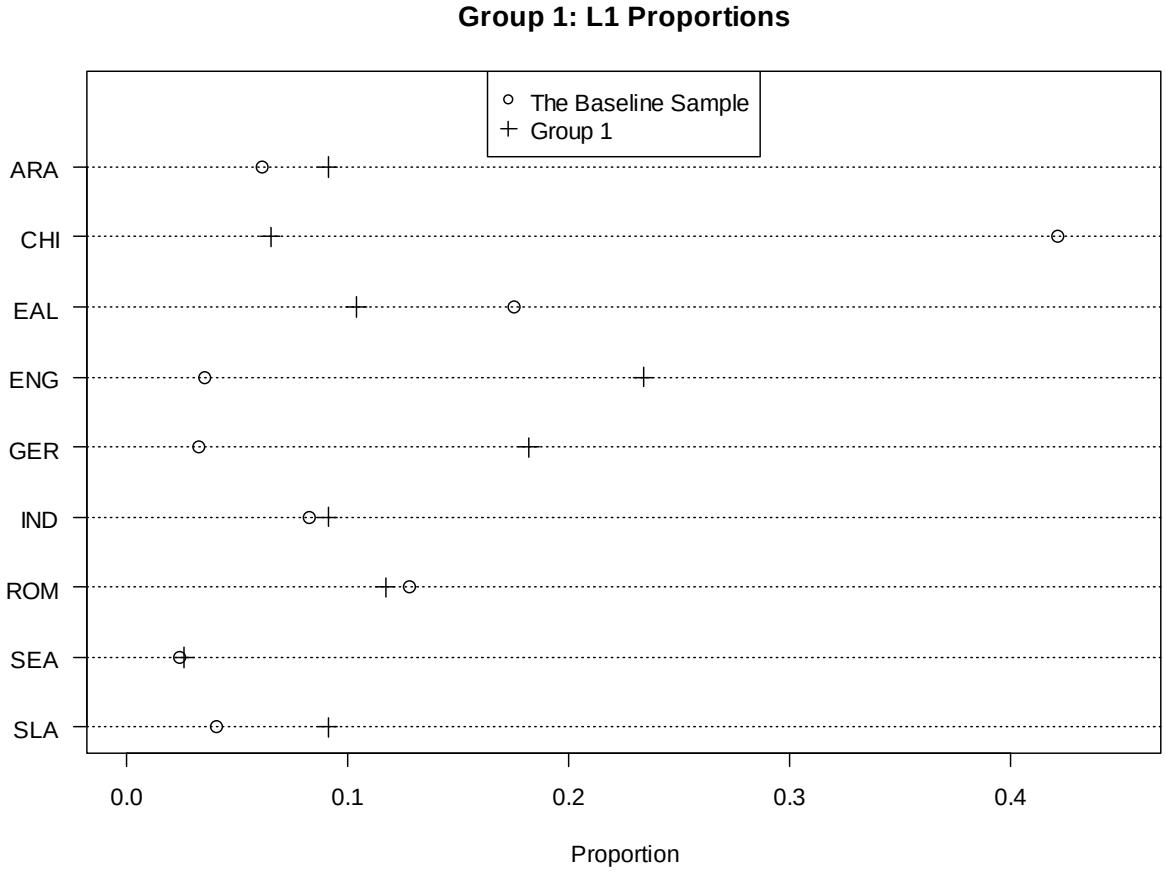


Figure 4-9. L1 proportions in Group 1

It is evident from Figure 4-9 that the proportion of English speakers in Group 1 was considerably higher than the corresponding baseline proportion. In fact, English speakers were the largest L1 group in Group 1 despite its very small baseline proportion. Approximately 70 percent of all English speaking test takers (18 out of 27) belonged to this group. Germanic language speakers comprised the second largest L1 group in Group 1, followed by Romance language speakers. The proportion of Germanic language speakers in Group 1 was also much higher than their baseline proportion. On the other hand, the proportion of Chinese and East Asian language speakers was much lower in Group 1 than in the baseline sample.

In sum, test takers who were classified into Group 1 possessed a profile consisting of almost perfect scores on Pronunciation and Lexical Grammar and above average scores on Rhetorical Organization and Question Handling. This pattern was consistent with what would be expected from a typical English native speaker. The linguistic background of Group 1 members agreed with the subscale score profile in that the largest L1 group consisted of English speaking test takers. Speakers of Germanic and Romance languages also accounted for large proportions in Group 1, which was expected considering their linguistic similarities to English. The subscale score pattern and the linguistic background of group members made it clear that Group 1 members possessed a near-native level of academic oral English proficiency. Consequently, Group 1 was labeled as the Near-native group.

4.3.3.2 Interpreting mixture model results: Group 2

The members of Group 2 turned out to carry a similar subscale score pattern to Group 1 members, with one important difference. The model-based scores of Group 2 members varied across all four subscales, while only Rhetorical Organization and Question Handling contributed to the within-group variance of Group 1. Figure 4-10 (a) shows that Group 2 members also varied in terms of their Pronunciation and Lexical Grammar scores. The model-based scores of almost all Group 2 members were located around the top-right quadrant of Figure 4-10 (a), which indicates that most Group 2 members scored above average on the first two subscales. Figures 4-10 (b) through (f) illustrate the same pattern in all bivariate distributions. Therefore, a typical Group 2 test taker could be regarded as having above-average scores on all four subscales.

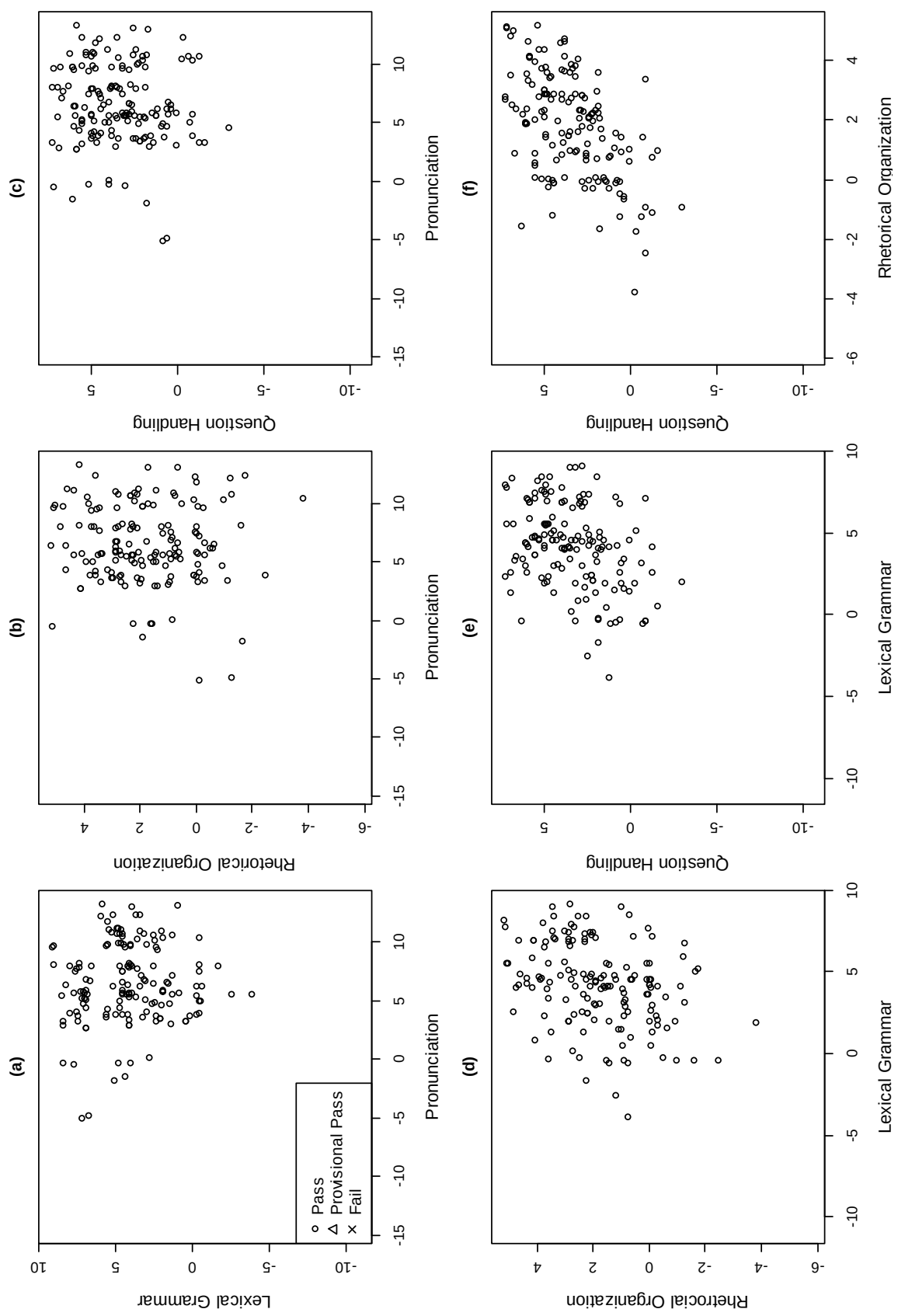


Figure 4-10. Bivariate relationships between the model-based subscale scores (Group 2)

The subscale score pattern observed in Group 2 corresponded to what would be expected from a typical test taker with advanced academic oral English proficiency. A test taker who is not a native speaker of English but highly proficient in English should be able to score well on all four subscales and pass the test without difficulty. Figure 4-10 shows that Group 2 members met both expectations; the vast majority of Group 2 members scored above-average on every subscale, and all Group 2 members passed the test. However, even a highly proficient learner is expected to make some pronunciation errors and carry a distinguishable accent (see, e.g., Bongaerts, 1999; Fayer & Krasinski, 1987; Yeni-Komshian, Flege, & Liu, 2000). In addition, occasional grammatical errors and unnatural word choices are commonly found in the speeches of proficient but non-native speakers of a language (see, e.g., Han, 1998; Ju, 2000). It was, therefore, not surprising to see variations in the Pronunciation and Lexical Grammar scores among highly proficient non-native English speakers.

Figure 4-11 presents the proportion of each language group in Group 2 together with the corresponding baseline proportions.

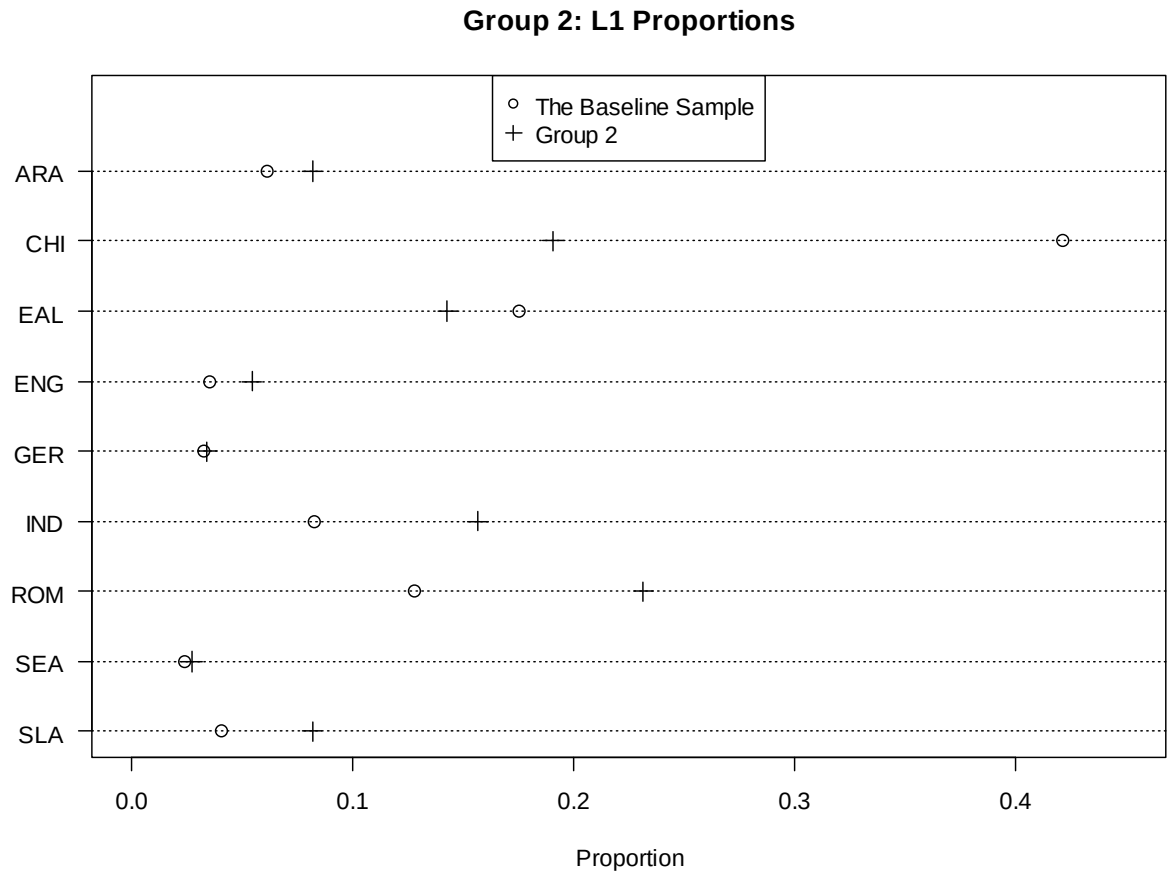


Figure 4-11. L1 proportions in Group 2

Group 2 was overall similar to Group 1 in terms of the linguistic background of the members. The proportions of Arabic, Chinese, East Asian, and Slavic language speakers in Group 2 deviated from the corresponding baseline proportions in the same direction as in Group 1. However, there were important differences between Group 2 and Group 1. First, the proportion of English speaking test takers was much smaller in Group 2. There were only a few English speaking test takers in this group. It is noteworthy that none of them were from the so-called

inner-circle countries (Kachru, 1985), in which English is used as the first language such as Canada and United Kingdom. All of the English speakers in Group 2 came from countries in which English is used as an official (second) language, such as India, Philippine, and Singapore. In contrast, many Group 1 English speakers were from inner-circle countries. The existence of English speaking test takers in Group 2 might appear counterintuitive given the interpretation of Group 1 as the near-native group. However, the difference in home countries between Group 2 and Group 1 English speaking test takers provided a reasonable explanation. In particular, region-specific accents as well as grammatical and lexical differences in English used in outer-circle countries can account for the variations in Pronunciation and Lexical Grammar observed in Group 1. Romance language speakers comprised the largest L1 group, and their proportion in Group 2 was much larger than their baseline proportion. Similarly, Indian language speakers accounted for a sizable portion of Group 2, and their proportion was much larger in Group 2 than in the baseline sample. English shares many commonalities with Romance languages and is used as an official language in India. Therefore, the relatively large proportion of Romance and Indian language speakers in Group 2 was consistent with the subscale score pattern involving high model-based scores across all four subscales.

The model-based subscale scores suggested that test takers in Group 2 possessed advanced academic oral English proficiency across all subscales. Furthermore, every Group 2 member passed the test. The linguistic background of Group 2 members presented supporting evidence for the inference based on the subscale score pattern and the TOP decision category. A large number of Group 2 members were either speakers of languages that share similarities with English or came from countries in which English is used as an official language. Considering the

subscale score pattern, the TOP decision category, and the linguistic background characterizing the members of Group 2, it was reasonable to label Group 2 as the High Proficiency group.

4.3.3.3 Interpreting mixture model results: Group 3

Group 3 consisted of test takers with a different subscale score pattern than the two previous group members. Figure 4-12 gives the bivariate distributions of the model-based scores among the four subscales in Group 3. Variations were observed in all six scatter plots. The variations in Pronunciation and Lexical Grammar were largely concentrated around the below-average region. Rhetorical Organization and Question Handling scores varied slightly more. The bivariate relationships among the subscales were not always positive, as can be seen in Figure 4-12 (c). As expected from the widely varying scores on all subscales, all three decision categories were present in Group 3. The majority of Group 3 provisionally passed the test. The proportions of the fail and pass categories were comparable, each of which accounting for approximately 20 percent of Group 3.

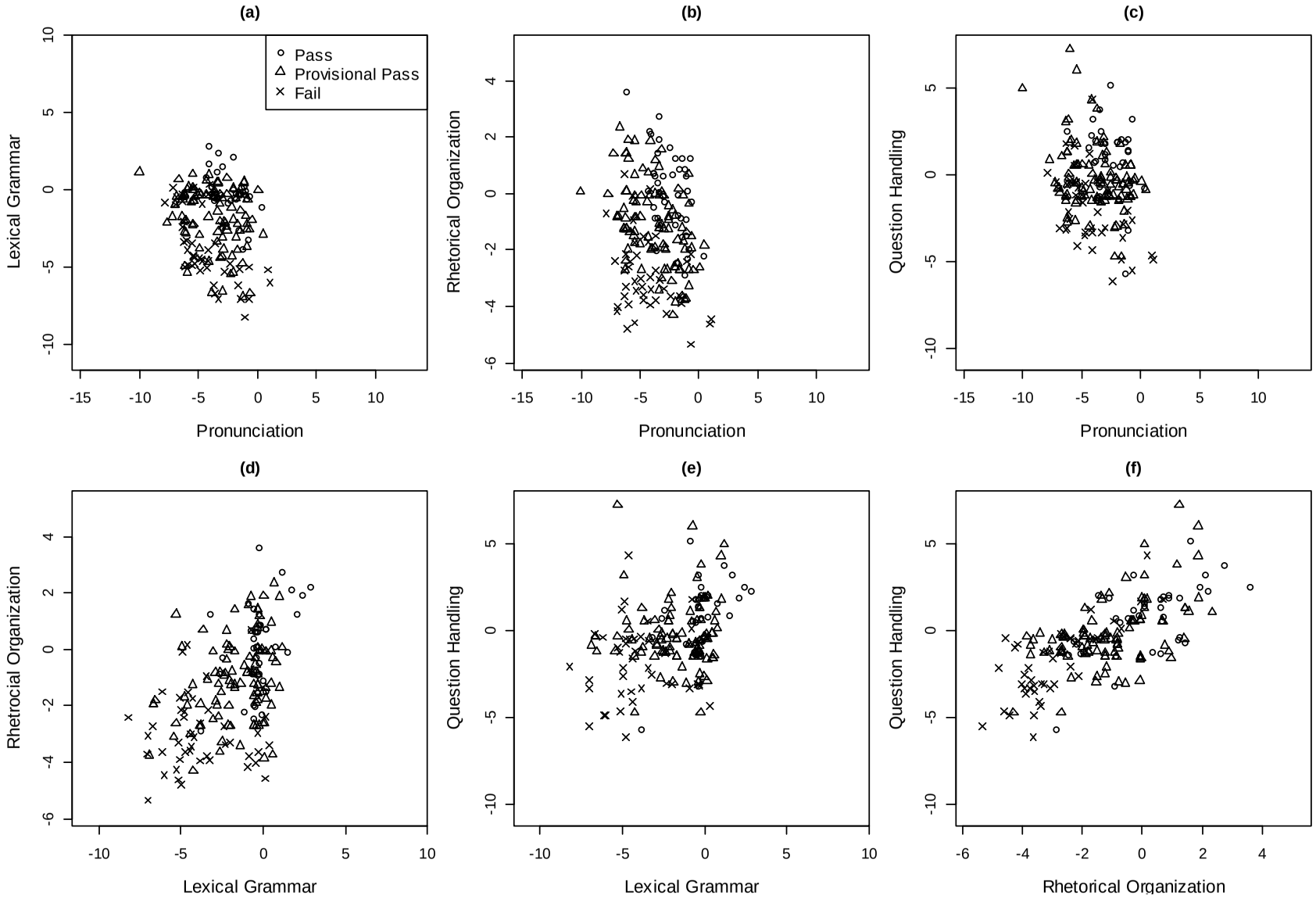


Figure 4-12. Bivariate relationships between the model-based subscale scores (Group 3)

The subscale score pattern characterizing Group 3 could be regarded as a middle ground between all-high profiles observed in Groups 1 and 2 and an all-low profile consisting of low scores across all subscales. It was clear that the test takers in Group 3 did not perform as well as Group 1 or 2 members. However, many Group 3 members received around- and even above-average scores on more than one subscale. In addition, the majority of Group 3 members belonged to the provisional pass category, which functions as a middle category between pass and fail. The intermediate nature of Group 3 is illustrated in Figures 4-12 (a), (c), and (e), which show the concentration of Group 3 members around the central regions.

The subscale score pattern and the decision categories of Group 3 members can be interpreted as supporting evidence for the current university policy for provisionally passed test takers, which mandates them to take any ESL oral-skills course. Figure 4-12 shows that a slight improvement on any of the four subscales would have led provisionally passed test takers to pass the test. Assuming the positive contribution of ESL oral-skills courses to a test taker's academic oral English proficiency, it is reasonable to expect that provisionally passed test takers would benefit from any ESL oral-skills course available to them.

Figure 4-13 gives the proportions of different L1 groups in Group 3 as well as the corresponding baseline proportions.

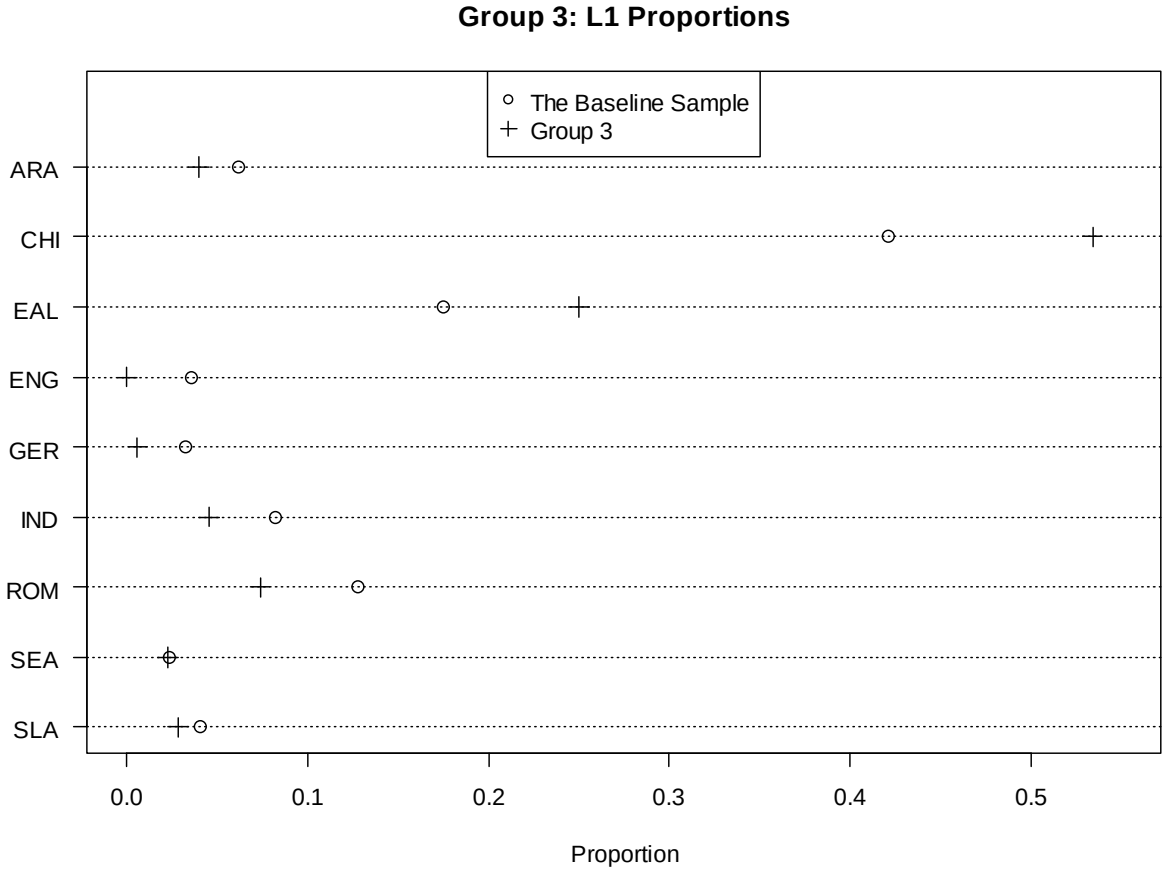


Figure 4-13. L1 proportions in Group 3

Group 3 consisted of test takers whose linguistic background differed considerably from the members of the two previous groups. In both Groups 1 and 2, Chinese and East Asian language speakers comprised a small proportion despite their large share in the baseline sample. In contrast, Group 3 consisted almost exclusively of the speakers of those three languages. Chinese and East Asian language speakers together accounted for approximately 80 percent of Group 3. Their proportion in Group 3 exceeded the corresponding baseline proportion. Group 3 only contained a few speakers of Arabic, European, and Indian languages, and no English speaker.

Test takers who were classified into Group 3 shared a subscale score pattern that could be characterized as intermediate. Their subscale scores were mostly located at below- or around-

average, with considerable variations in every subscale dimension. The majority of Group 3 members belonged to the provisional pass category, which lies between the pass and fail categories. Group 3 members were quite different from Group 1 and 2 members in terms of their linguistic background. Compared to the near-native and overall high-proficiency possessed by the members of Groups 1 and 2, respectively, Group 3 members could be regarded as having an all-intermediate profile. Consequently, Group 3 was labeled as the Intermediate Proficiency group.

4.3.3.4 Interpreting mixture model results: Group 4

Figure 4-14 shows the distributions of Group 4 members' model-based scores as well as the bivariate relationships among the subscales. Variations were present in all six plots in Figure 4-14, but mostly confined to the lower-left quadrants. This indicates that the test takers classified into Group 4 mostly received below-average scores on all subscales. Given the low model-based scores, it is not surprising that most test takers (61 out of 78) in Group 4 belonged to the fail category. There was only one test taker who passed. Considering the low scores on all subscales, that student appeared to be erroneously assigned the pass category due to rater effects. Test takers who provisionally passed the test comprised approximately 20 percent of Group 4. The difference between Group 4 and Group 2 was noteworthy. Group 4 test takers turned out to share a subscale score pattern that was almost the exact opposite of the pattern shared by Group 2 members. The contrast between Group 4 and Group 2 becomes clear when Figures 4-10 and 4-14 are compared. The figures present pairs of scatter plots that are almost mirror images of each other. Not all bivariate relationships were positive, as can be seen in Figures 4-14 (b) and (e). In

particular, it appeared that Question Handling was negatively correlated with Pronunciation and Lexical Grammar.

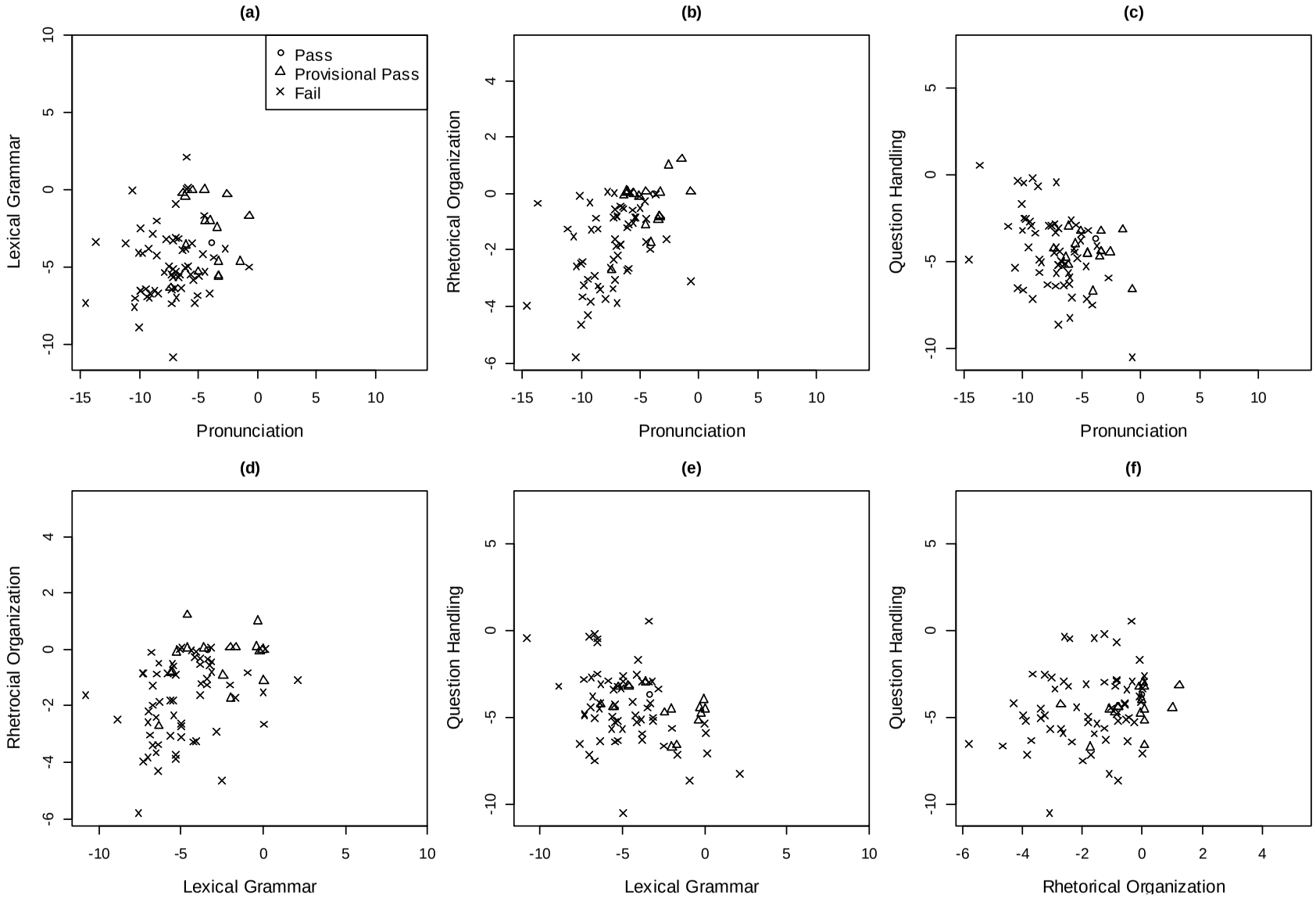


Figure 4-14. Bivariate relationships between the model-based subscale scores (Group 4)

The subscale score pattern illustrated in Figure 4-14 largely corresponded to a profile expected from a test taker with a low level of academic oral English proficiency. Few in Group 4 achieved above-average scores on any of the subscales. The majority were given lower model-based scores across all subscales than Group 3 test takers, who were regarded as intermediate test takers. Furthermore, the vast majority of Group 4 members could not either pass or provisionally pass the test. The negative relationships between subscales demonstrated in Figures 4-14 (c) and (e) provided an interesting insight. In particular, these negative relationships might reflect an impact of test preparation among low-proficiency test takers. It has been suggested that interpersonal skills such as Question Handling are in general easier to improve with preparation than language-specific skills such as Pronunciation and Lexical Grammar (Halleck & Moder, 1995).

The subscale score pattern shared by Group 4 members can be interpreted as supporting evidence for the current policy for test takers who failed the test. The TOP decision rule states that test takers who belong to the fail category are not allowed to work as a TA until they take the test again and are assigned the pass or provisional pass category. It is clear from Figure 4-14 that, but for only one exception, even the highest scoring Group 4 members failed to pass the test. This suggested that most test takers in Group 4 would not be able to pass the test even with moderate improvements. In this light, Group 4 was different from Group 3, in which within-group improvements in any dimension would be likely to lead one to pass the test. Therefore, the current TOP policy, which does not allow failed test takers to start working as a TA regardless of their enrollment in ESL oral-skills courses, can be viewed as consistent with the empirical classification results of this study.

The proportion of each L1 group in Group 4 and in the baseline sample is given in Figure 4-15.

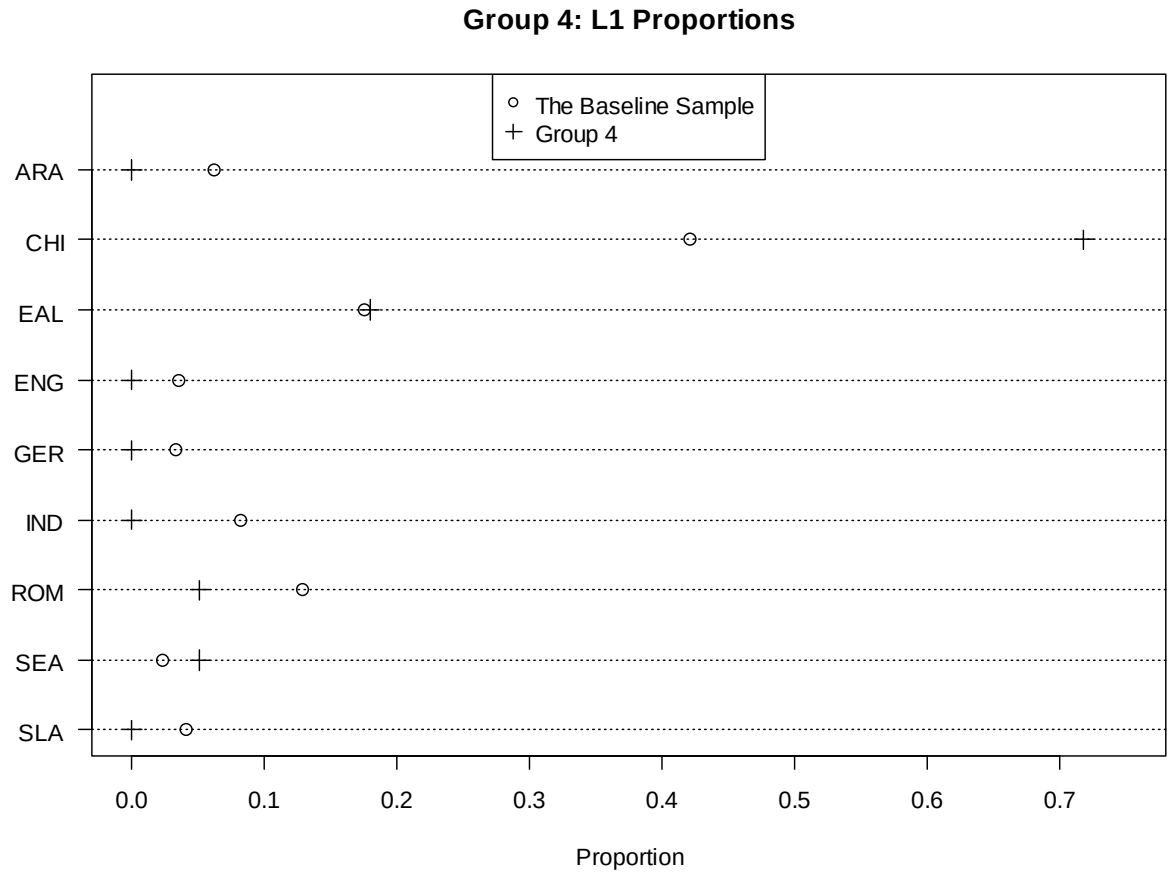


Figure 4-15. L1 proportions in Group 4

Figure 4-15 shows that Group 4 largely consisted of two L1 groups – Chinese and East Asian languages. Chinese and East Asian language speakers together comprised approximately 90 percent of Group 4. While Group 4 members demonstrated largely comparable linguistic background to Group 3 members, the predominance of the Chinese and East Asian language speakers was more pronounced in Group 4. Chinese language speakers accounted for a considerably larger proportion in Group 4 than in the baseline sample. The contrast between Group 4 and Group 2 was also observed in the linguistic background. As expected, no English

speaker was classified into Group 4. Group 4 also did not include speakers of Arabic, Germanic, Indian, or Slavic languages, which accounted for a sizable portion of Group 2. Only 10 percent of Group 4 members spoke Romance languages as their L1, while the same L1 group comprised approximately 25 percent of Group 2.

Most test takers in Group 4 received below-average scores on all subscales. This pattern was in complete contrast to the subscale score pattern observed in Group 2, which was interpreted as the high-proficiency group. The contrast between Group 4 and Group 2 continued in terms of linguistic background. Group 4 consisted almost exclusively of Chinese and East Asian language speakers, who comprised a much smaller proportion in Group 2. Group 4 members in general received lower model-based scores than Group 3 members across all subscales. The vast majority of Group 4 members failed the test. Considering the subscale score pattern and the predominant decision category, it was decided to label Group 4 as the Low Proficiency group.

4.3.3.5 Interpreting mixture model results: Group 5

Test takers who belonged to Groups 1 through 4 turned out to possess flat profiles. That is, most of their model-based scores were either above-, around-, or below-average on all subscales. Group 5 differed from the previous four groups in that its members were widely scattered in some dimensions while highly concentrated with little variance in other dimensions. Figure 4-16 presents the model-based subscale score distributions and bivariate relationships. Figure 4-16 (f) shows that Group 5 members were homogeneous in terms of Rhetorical Organization and Question Handling. In contrast, they differed widely in terms of their Pronunciation and Lexical Grammar scores, as can be seen in Figure 4-16 (a). Group 1 was similar to Group 5 in that it also had two subscale dimensions, namely Pronunciation and Lexical Grammar, along which test

taker scores seldom varied. However, there were two important differences between Group 1 and Group 5. First, the Pronunciation and Lexical Grammar scores of Group 1 members were concentrated around the highest possible scores on both subscales. In Group 5, on the other hand, the Rhetorical Organization and Question Handling scores were concentrated around the average. In addition, the score variations of Group 1 test takers in Rhetorical Organization and Question Handling subscale dimensions were largely confined to the upper-right quadrants of Figure 4-8 (f), whereas Group 5 test takers were distributed across the whole range of Pronunciation and Lexical Grammar subscales, as can be seen in Figure 4-16 (f).

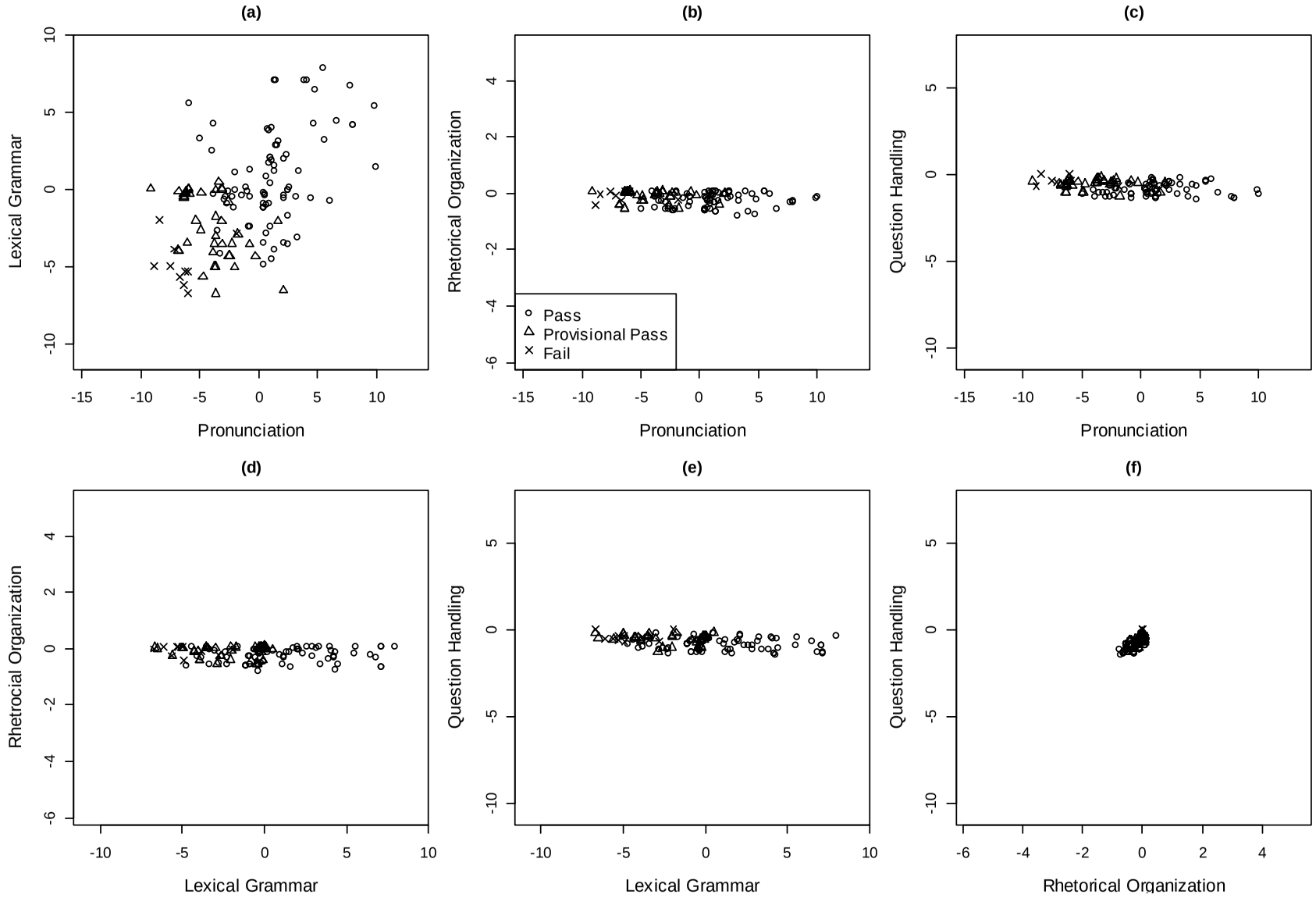


Figure 4-16. Bivariate relationships between the model-based subscale scores (Group 5)

All three TOP decision categories were present in Group 5. Approximately 60 percent of Group 5 test takers passed the test, while provisionally passed and failed test takers comprised approximately 30 and 10 percent, respectively. Given the little variance in Rhetorical Organization and Question Handling, the decision category of Group 5 test takers was largely determined by their Pronunciation and Lexical Grammar scores. Figure 4-16 (a) illustrates the crucial role of Pronunciation and Lexical Grammar in assigning the decision categories. In particular, most test takers in the upper-right quadrant of Figure 4-16 (a) passed the test, whereas those who were distributed in the lower-left quadrant did not. Figures 4-16 (b) through (e) illustrate a similar impact. Those four figures show that the decision category changes roughly around zero (the average) of the x-axes, which represent Pronunciation in (b) and (c), and Lexical Grammar in (d) and (e).

The variances of Pronunciation and Lexical Grammar scores were considerable despite the almost complete lack of variance of the other two subscale scores. While all subscales were highly correlated in the entire sample, Rhetorical Organization and Question Handling were irrelevant to Pronunciation and Lexical Grammar in Group 5. Language proficiency is often regarded as a multi-componential construct in the second language learning and testing literature (e.g., Bachman & Palmer, 2010; Foster, Tonkyn, & Wigglesworth, 2000; Iwashita, Brown, McNamara, & O'Hagan, 2008; Poehner, 2008), and many frameworks suggest that organizational and interpersonal skills pertinent to Rhetorical Organization and Question Handling are distinct from more language-specific skills such as Pronunciation and Lexical Grammar (e.g., Bachman & Palmer, 2010; Canale & Swain, 1980). Group 5 shows that, for a subset of second language speakers, Pronunciation and Lexical Grammar can vary widely even when Rhetorical Organization and Question Handling are held constant at a certain level. This

could be interpreted as supporting evidence for the multi-componential nature of second language oral proficiency.

From a TOP decision point of view, Group 5 consisted of two subgroups: test takers who passed the test and who did not. As previously described, the difference in the decision was completely driven by Pronunciation and Lexical Grammar. Figure 4-16 (a) clearly shows that, for Group 5 members who did not pass the test, improvements were needed in Pronunciation and Lexical Grammar. Furthermore, the around-average Rhetorical Organization and Question Handling scores shared by Group 5 test takers were comparable to those of many Group 2 members, all of whom passed the test. The decisive role of Pronunciation and Lexical Grammar and the around-average scores in the other two subscales suggested that Group 5 members who did not pass the test would benefit most from interventions focusing on their Pronunciation and Lexical Grammar. In this light, Group 5 provides evidence against the current policy that allows provisionally passed test takers to take any of the available ESL oral-skills courses. At least to test takers who were classified into Group 5, ESL courses focusing on one's pronunciation, oral grammar, and vocabulary uses would be more pertinent than those focusing on organization and/or dealing with questions.

Figure 4-17 presents the proportion of each L1 group in Group 5 together with the corresponding baseline proportion.

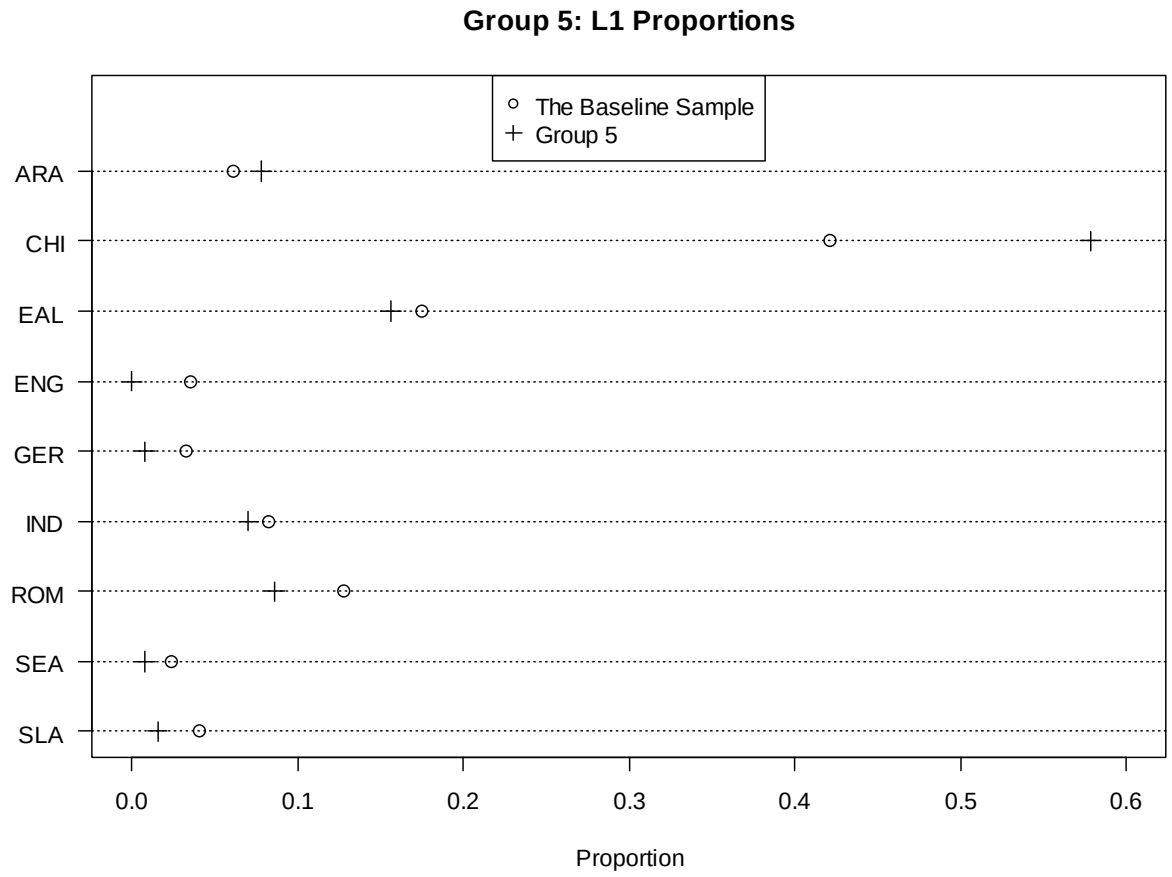


Figure 4-17. L1 proportions in Group 5

The linguistic background of Group 5 members was characterized by the predominance of East Asian language speakers. Test takers who spoke Chinese and East Asian languages accounted for approximately 80 percent of Group 5. The proportion of Chinese language speakers exceeded the baseline proportion. Speakers of Arabic, Indian, and Romance languages together accounted for the remaining 20 percent. As described earlier, Group 5 could further be divided into two subgroups depending on the TOP decision category, and the compositions of the two subgroups

in terms of their linguistic background were compared. The members of the two subgroups came from different linguistic background. Speakers of Arabic, Indian, and Romance languages mostly belonged to the "pass" subgroup. In contrast, the "non-pass" subgroup consisted almost exclusive of Chinese and East Asian language speakers. The linguistic background of the non-pass group members was in fact highly comparable to that of Group 4 members.

Group 5 was interesting in that its members did not possess a flat profile. Test takers classified into Group 5 were very homogenous in terms of their Rhetorical Organization and Question Handling scores, but differed widely along the Pronunciation and Lexical Grammar subscale dimensions. All three decision categories were present, and Pronunciation and Lexical Grammar were crucial in assigning the decision categories. The reasonably satisfying scores on Rhetorical Organization and Question Handling subscales suggested that Group 5 test takers who failed to pass the test would benefit the most from an ESL course focusing on Pronunciation and Lexical Grammar, which is not in line with the current TOP policy. Considering the non-flat subscale score pattern characterizing Group 5, it was decided to keep its label as descriptive as possible. In particular, Group 5 was labeled as the Varying Pronunciation/Grammar group, given its large variance in the two subscale dimensions. Group 5 mostly consisted of Chinese and East Asian language speakers. When only those who did not pass the test were considered, those two L1 groups accounted for almost 90 percent, which resulted in a highly comparable L1 group composition to Group 4.

4.3.3.6 Interpreting mixture model results: Group 6

Group 6 members shared similar scores on the Pronunciation subscale, while varying considerably along the other three subscale dimensions. The subscale score patterns of Group 6

test takers, therefore, were not flat. Figure 4-18 presents the model-based score distributions of Group 6 test takers using bivariate scatter plots. Bivariate relationships not involving Pronunciation were all positive. Figures 4-18 (a), (b), and (c) illustrate the lack of variance along the x-axes, all of which represented Pronunciation. The Pronunciation scores of Group 6 members were concentrated slightly above the average. On the other hand, each of the Lexical Grammar, Rhetorical Organization, and Question Handling scores was distributed across the whole range except around the lowest extremes. Despite the large variances of the latter three subscale scores, however, all but two test takers in Group 6 passed the test. This appears to be the impact of the current TOP score weighting system, which assigns a higher weight to Pronunciation than the others. By virtue of having above-average scores on Pronunciation, most test takers in Group 6 were able to pass the test even with below-average scores on the other three subscales.

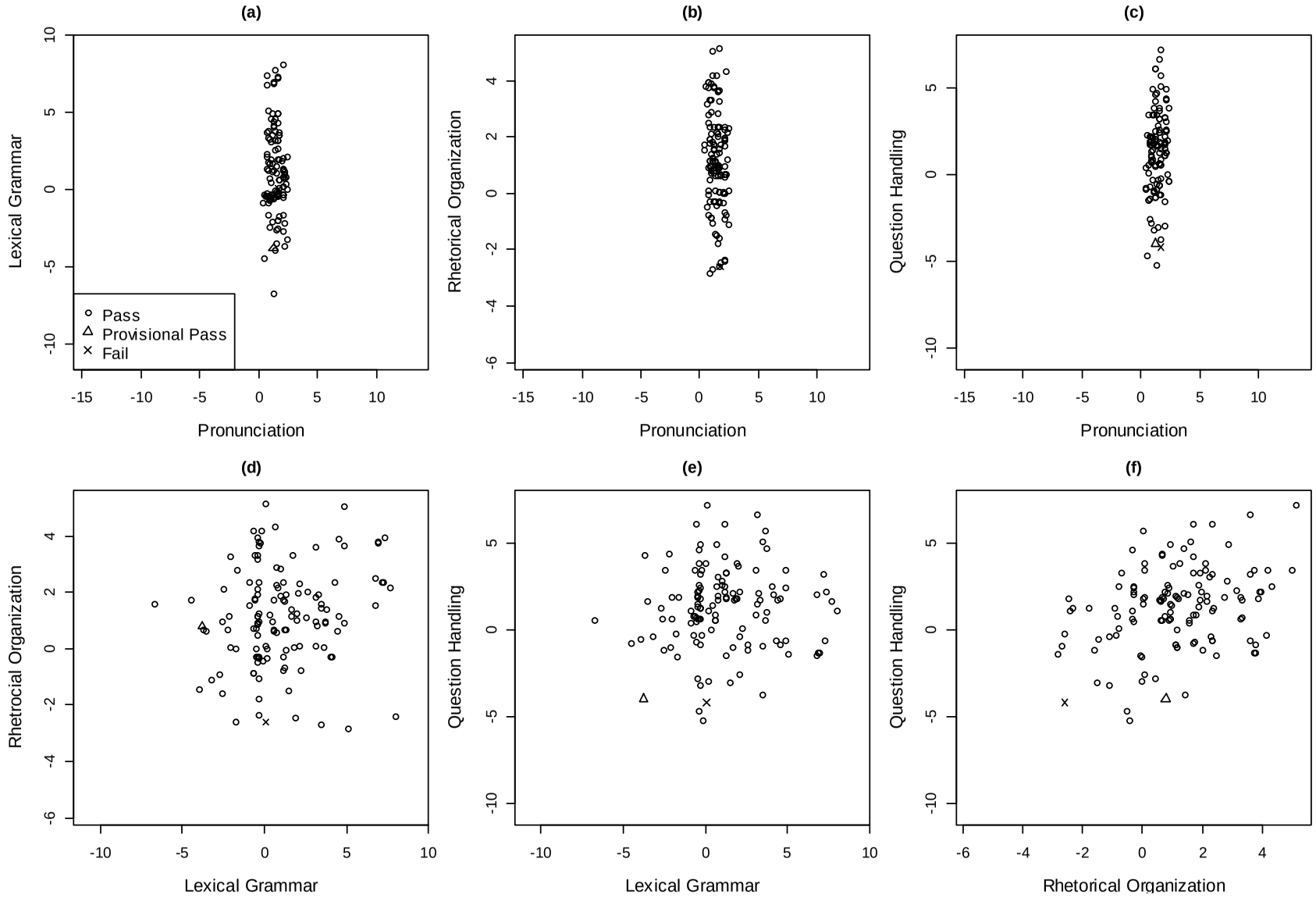


Figure 4-18. Bivariate relationships between the model-based subscale scores (Group 6)

Pronunciation and Lexical Grammar scores tended to move together in the same direction in Groups 1 through 5. Group 6 presented an exception to this tendency in that its members varied widely in terms of Lexical Grammar while their Pronunciation scores were homogeneous. This provided an interesting insight into the relationship between Pronunciation and Lexical Grammar across different subscale profiles. The Pronunciation scores of Group 6 members were slightly above-average, which were higher than those of Groups 3 and 4 members but lower than those of Groups 1 and 2 members. When Pronunciation was near-native, Lexical Grammar was also native like, as was the case in Group 1. Group 2 provided evidence for the positive relationship between Pronunciation and Lexical Grammar among high-proficiency test takers. The same held among intermediate- and low-proficiency test takers in Groups 3 and 4, respectively. However, when Pronunciation scores were slightly above-average, which was the case in Group 6, it appears that the relationship could break down. This could indicate a short window in some language learners' second-language development during which learners arrive at a certain level in terms of pronunciation whereas their oral grammar still can vary across individuals. Given the exploratory and local nature of this study, however, it should be noted that this is not the only possible interpretation of the observed phenomenon. Future studies are needed to evaluate the validity of this interpretation.

Group 6 provided another important insight for a future standard setting study. Figures 4-18 shows that many test takers in Group 6 passed the test even with below-average scores on at least two out of the three subscales other than Pronunciation. As previously mentioned, this was largely due to the higher weight Pronunciation carries. Given the current weighting scheme and the cutoff score for the pass category, once a test taker gets an above average Pronunciation score, it is very rare for the test taker to fail to pass the test. It is not clear whether this was

considered when the weighting scheme and the cutoff scores were first determined. Whether test takers with a probable profile consisting of an above-average Pronunciation score and below-average scores on all other subscales would be regarded as proficient enough to qualify as a TA is an important question that needs to be addressed in a standard setting study. Test takers who were classified into Group 6 can be used as a focused sample set for such a standard setting project.

Figure 4-19 gives the proportion of each L1group in Group 6 together with the corresponding baseline proportion.

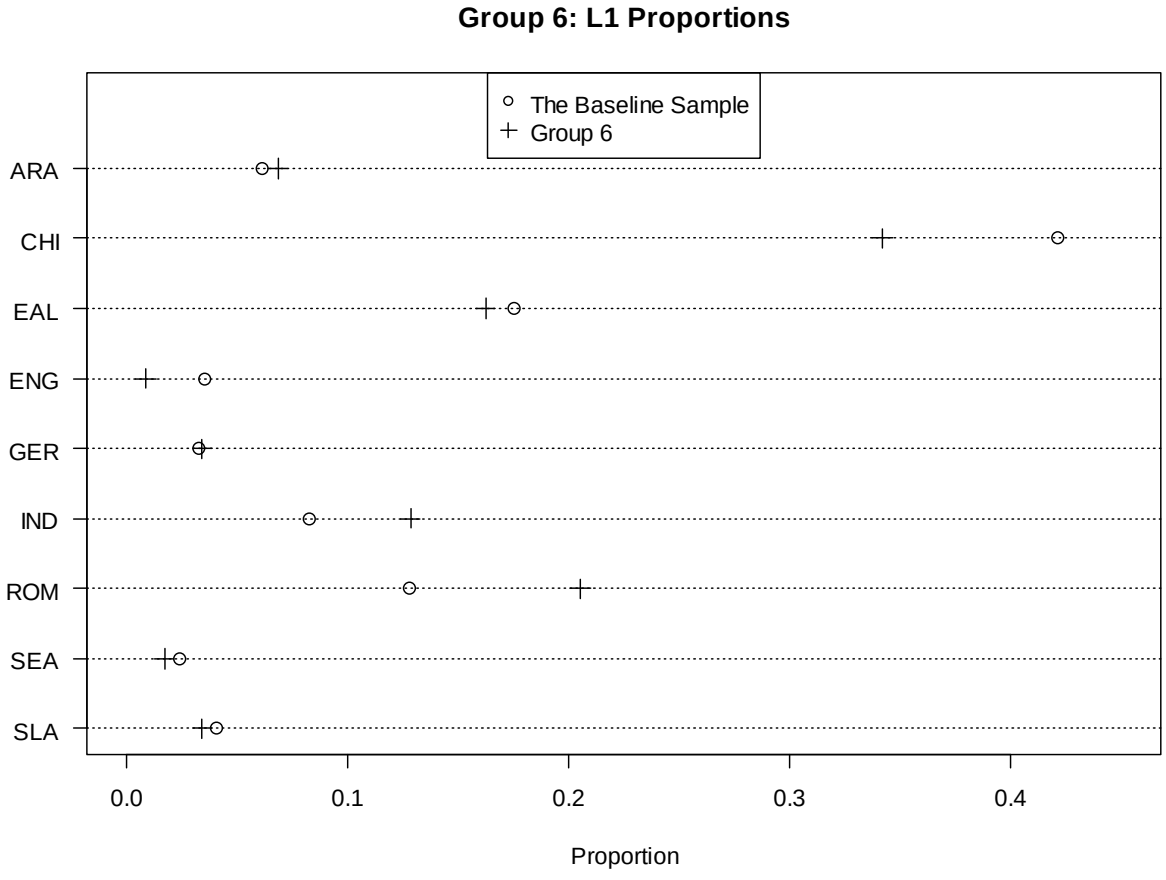


Figure 4-19. L1 proportions in Group 6

The linguistic background of Group 6 members did not differ much from that of the baseline sample. The proportions of Arabic, East Asian, German, Slavic, and South East Asian language speakers were almost identical to the corresponding baseline proportions. Speakers of Chinese languages comprised the largest L1 group, accounting for approximately 35 percent of Group 6. However, their proportion here was slightly lower than the baseline proportion. Romance language speakers comprised the second largest group, and their proportion in Group 6 was higher than their baseline proportion. Indian language speakers also accounted for a larger proportion in Group 6 than in the baseline sample.

Test takers who were classified into Group 6 presented another case of a non-flat profile. While their Pronunciation scores were highly concentrated around the average, all other subscale scores varied considerably. All but two test takers in Group 6 passed the test, including many who received below average scores on other subscales, as a result of the higher weight assigned to Pronunciation. In sum, Group 6 could be characterized by the concentration of the Pronunciation scores around the average, and therefore was labeled as the Average Pronunciation group. The linguistic background of Group 6 members was largely comparable to that of the baseline sample.

4.3.3.7 Interpreting mixture model results: Group 7

Group 7 members presented still another case of a non-flat score profile. Six scatterplots in Figure 4-20 give the bivariate distributions of Group 7 members' model-based subscale scores. Figures 4-20 (a), (c), and (e) show that Group 7 members were very homogeneous in terms of Pronunciation, Lexical Grammar, and Question Handling. In particular, their scores on the three subscales were concentrated around the corresponding averages. Rhetorical Organization was the

sole source of within-group variance, as can be seen in Figures 4-20 (b), (d), and (f). The variance of the Rhetorical Organization scores was relatively small with most test takers distributed around the average. Extreme scores were not observed in Group 7. All Group 7 members passed the test.

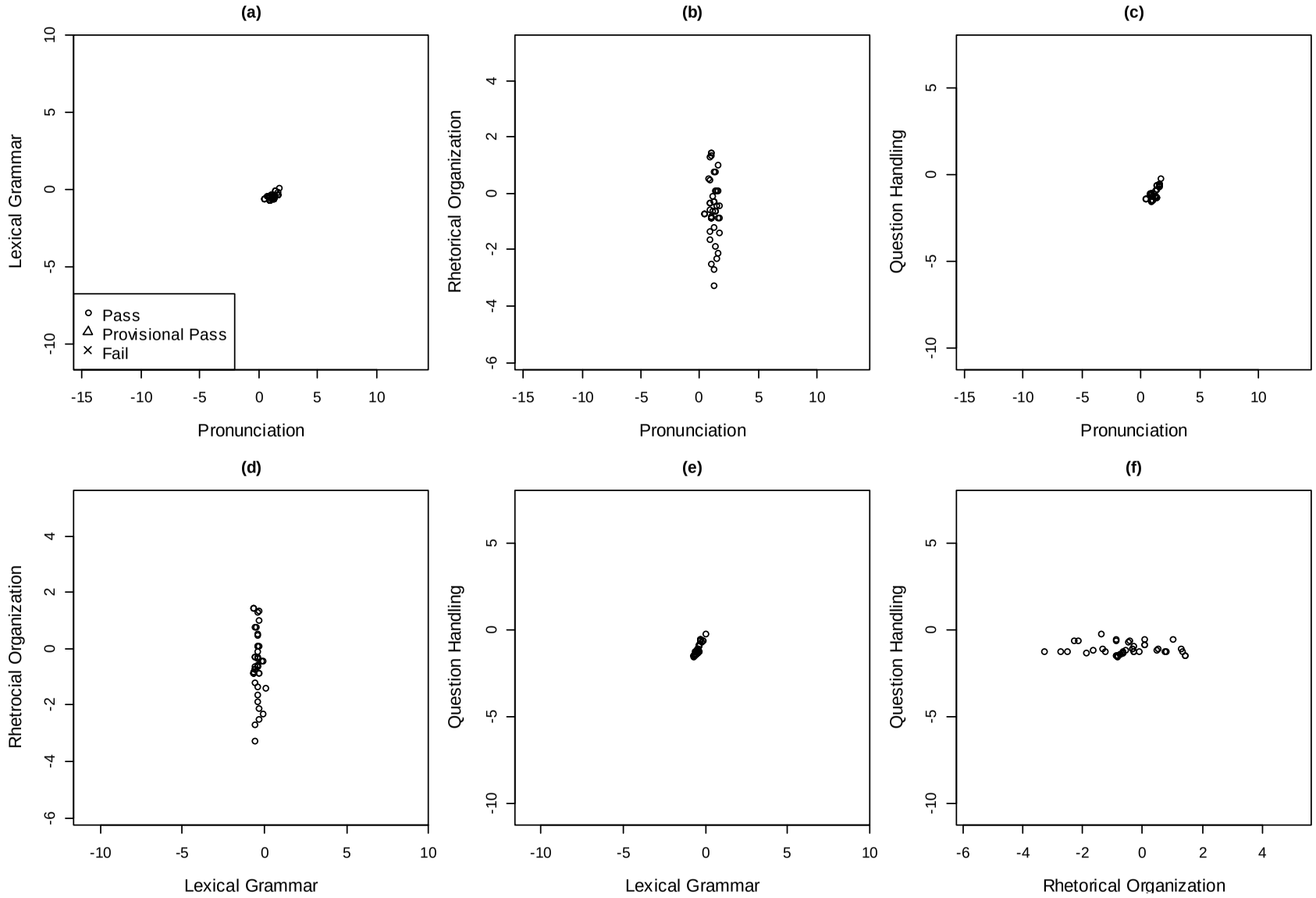


Figure 4-20. Bivariate relationships between the model-based subscale scores (Group 7)

The characteristic feature of Group 7 was the variance of the Rhetorical Organization scores, in contrast to the lack of variances on the other subscales. This might have to do with a slight difference in the nature of Rhetorical Organization compared to the other subscales. In particular, the prior preparation of a test taker is directly relevant to Rhetorical Organization, but not necessarily to the other subscales. Being native-like does not guarantee a perfect score in Rhetorical Organization, as was the case in Group 1. In order to receive a high score in Rhetorical Organization, a test taker needs to give a presentation in a coherent and clear manner, which can be enhanced with a thorough preparation. Therefore, the characteristic subscale score profile of Group 7 might be interpreted as reflecting the impact of prior preparation on Rhetorical Organization among test takers whose proficiency is around average in terms of the other three subscales. The homogeneous decision category in Group 7 could further be interpreted as suggesting that preparation did not appear to have an impact on the TOP decision for Group 7 members. However, considering the exploratory and local nature of this study, this is by no means the only possible interpretation, and future studies are needed to evaluate its validity.

Figure 4-21 gives the proportion of each L1 group in Group 7, as well as the corresponding baseline proportions.

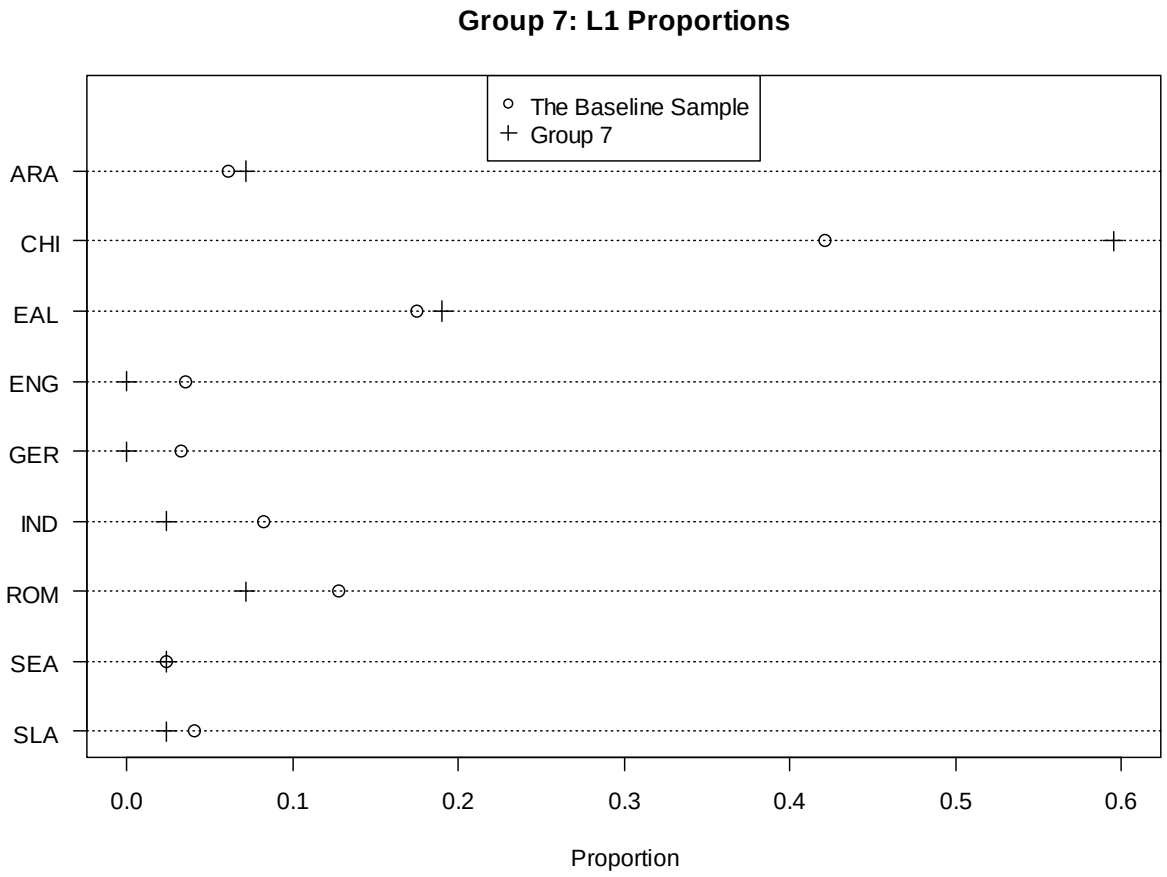


Figure 4-21. L1 proportions in Group 7

Speakers of Chinese languages comprised the largest L1 group, accounting for approximately 60 percent of Group 7. The Chinese language speakers also accounted for a higher proportion in Group 7 than in the baseline. The proportions of other L1 groups were in general smaller in Group 7 than the corresponding baseline proportions, but mostly by only a small margin. East Asian language speakers comprised the second largest L1 group, followed by Arabic and Romance language speakers.

Group 7 test takers could be characterized by the variance of the Rhetorical Organization scores and the homogeneity in terms of Pronunciation, Lexical Grammar, and Question Handling. The former might be interpreted as the impact of preparation considering the need for preparation to receive a good score on Rhetorical Organization. Given the characteristic Rhetorical Organization variance, it was decided to label Group 7 as the Varying Organization group. As expected from the above-average scores on all subscales but Rhetorical Organization, every member of this group passed the test. The proportion of each L1 group was similar to the corresponding baseline proportion, with the exception of Chinese language speakers who comprised the majority.

4.3.4 Summary of the section

This section reported the clustering of the model-based scores for an investigation of subscale profile groups. FM models with four dimensional normal components were fit to the model-based subscale scores estimated in the previous section. The BIC (Schwarz, 1978) and substantive interpretability were considered in model selection. The final model suggested seven profile groups in the TOP dataset, and assigned most test takers a group membership with high confidence. All seven groups were interpreted and labeled based on characteristic subscale score patterns, as well as group members' linguistic background and TOP decision categories. Table 4-13 summarizes of the label and composition of each profile group.

Table 4-13

Labels and Member Background Information of the Seven Profile Groups

Group	Label	Decision Categories	High Prop. L1
1	Near-native	All pass	ENG, GER
2	High Prof.	All pass	IND, ROM
3	Intermediate Prof.	Mostly prov. pass	CHI, EAL
4	Low Prof.	Mostly fail & prov. pass	CHI, EAL
5	Varying PR/LG	Approx. half pass	CHI, EAL
6	Average PR	Mostly pass	CHI, EAL, ROM, IND
7	Varying RO	All pass	ARA, CHI, EAL

4.4 Summary of the chapter

This chapter reported the findings of analyses addressing the first research question, "What different subscale score profiles of academic English oral proficiency can be meaningfully produced by the TOP?" This chapter consisted of three main analysis steps. The first step investigated the parameter recovery of the second-order GRM through a small-scale simulation study. The results of the simulation served as a justification for the use of the second-order GRM in estimating the model-based subscale scores of the 960 TOP test takers, which comprised the second analysis step. The last analysis step explored profile groups underlying the estimated model-based subscale scores.

The parameter recovery of the second-order GRM was evaluated in three simulation settings that were designed to represent typical language performance assessment contexts with varying levels of challenge. True values and synthetic datasets were generated using the second-order GRM as the true model, and parameters were estimated via MCMC. The starting values for the MCMC estimation were obtained based on ML estimates and standard errors. The simulation results showed that the second-order GRM successfully recovered all parameters across all three

settings. The results also demonstrated the advantage of the second-order GRM over a simple average scoring method in estimating the model-based subscale scores.

Given the positive results of the simulation study, the second-order GRM was fitted to the TOP dataset to estimate the model-based subscale scores. MCMC was utilized as the main estimation method, and ML estimates and standard errors were used to obtain starting values that were reasonable and well-dispersed. The second-order GRM showed a better fit than the HRM (Patz et al., 2002), and was favored by both the AIC and BIC for all subscales except Pronunciation, despite the penalty for added model complexity. The second-order GRM yielded four model-based subscale scores for each of the 960 test takers. The resulting model-based subscale scores were then compared to the observed sum scores. The comparison showed that the distributions of the two sets of scores were comparable, while the second-order GRM estimates provided a continuous scale which yielded additional fine-tuned distinctions between test takers owing to its continuous nature.

The model-based subscale scores from the second analysis step were used as the data for the FM analysis that examined subscale profile groups in the TOP dataset. The final model identified seven profile groups and classified the TOP test takers into the seven groups. Test takers with near-native academic oral English proficiency comprised Group 1. Groups 2, 3, and 4 consisted of test takers with high-, intermediate-, and low-proficiency, respectively, in academic oral English. Group 5 test takers varied widely in terms of Pronunciation and Lexical Grammar. Group 6 consisted of test takers whose Pronunciation scores were concentrated in the average. Lastly, Group 7 test takers only varied in terms of Rhetorical Organization.

Chapter 5: Features of Oral Discourse across Subscale Score Profiles

This chapter presents the findings of analyses addressing the second research question, "What features of oral discourse characterize test takers who have different subscale score profiles?" In particular, this chapter discusses the results from a comparison between the TOP test taker performances and a reference corpus consisting of academic lectures. Furthermore, the distribution of discourse organizing lexical bundles across subscale score profile groups is examined in this chapter. An integral element of the both analyses was a small corpus consisting of a subset of TOP test taker performances. The procedures used in corpus construction and the resulting corpus are reported here. This chapter is organized in the following order. The corpus construction procedures and the characteristics of the resulting corpus are presented first. The resulting corpus is then compared to the reference corpus, and the results from the comparison are discussed. Lastly, the distributional and formal characteristics of discourse organizing lexical bundles are presented.

5.1 The TOP corpus

The second research question was addressed based on a small oral corpus consisting of the lecture task performances of 82 TOP test takers (hereafter the TOP corpus). This section provides a detailed description of the corpus building procedures and the characteristics of the TOP corpus. Selection criteria for the 82 test takers are discussed first, followed by a description of the transcription procedures. Lastly, the characteristics of the TOP corpus are presented.

5.1.1 Sample selection

The main motivation for constructing the TOP corpus was to enhance the identification and search of important discourse features in test taker performances. It would have been ideal if the search could be based on the performances of all test takers. However, given the spoken nature of the TOP performances, the construction of a corpus based on them inevitably involved the transcription of spoken words into a machine-readable text format, which is an extremely time-consuming process. Transcribing the lecture task performances of all 960 test takers would require a prohibitive amount of transcription time which was not available at the time of analysis. Therefore, it was decided to select a subset of the TOP test takers based on important characteristics to obtain a small but representative corpus. Both group-level and individual-level characteristics were considered in the selection procedure.

To ensure comparability among the profile groups, it was decided to include the same number of test takers from each of the seven groups. There were, however, two exceptions, both of which had to do with the TOP decision categories. Chapter 4 reported mixed decision categories in Groups 3 (the Intermediate Proficiency group) and 5 (the Varying Pronunciation/Grammar group), and it was suggested that those two groups could be further divided into two subgroups, namely the pass and non-pass subgroups. Therefore, it was desirable to equally represent the pass and non-pass subgroups by selecting the same number of test takers from both. It was decided to select eight test takers from each of the pass/non-pass subgroups in the two profile groups, and ten test takers from each of the other five profile groups. While the specific numbers of test takers (i.e., ten and sixteen) in each group were somewhat arbitrarily

decided, they were believed to achieve a reasonable balance between the practicality of transcription procedures and the representativeness of the resulting corpus.

Primary selection criteria at the individual level included the classification uncertainty and L1 of test takers. Given the goal of obtaining a representative subset of each profile group, it was clear that test takers with low classification uncertainty should be preferred. In addition, considering the vast difference in the L1 composition among the profile groups reported in Chapter 4, efforts were made to retain the L1 proportions of each profile group. However, the relatively large number of different L1 groups (i.e., nine) made the selection procedure very difficult. Consequently, it was decided to reduce the number of L1 groups by collapsing languages that demonstrated similar characteristics in terms of profile group membership. In particular, speakers of Chinese languages and East Asian languages were combined into the Chinese, Japanese, and Korean group (CJK). Germanic languages and English were collapsed into another group (GEE). Lastly, speakers of Arabic, South East Asian, and Slavic languages were grouped into a large category that was conveniently named Other (OTH), given their small and stable proportions in all profile groups.

The pool of selectable test takers became relatively small once the classification uncertainty and L1 were considered. When different choices were available, test takers' academic background was considered as a secondary selection criterion. The 765 test takers in the baseline sample came from six large academic divisions: Arts & Humanities (AHM), Engineering (ENN), Life & Health Sciences (LHE), Physical Sciences (PHY), Social Sciences (SOC), and Interdisciplinary Programs (OTH).[7] The objective of employing the secondary selection criterion

[7] These divisions were based on the administrative structure of UCLA.

was to retain the baseline proportions of the six academic divisions in the TOP corpus. The decision to use the academic divisions as a secondary criterion instead of a primary one was made based on empirical grounds. While research suggests that different academic disciplines amount to differences in classroom language use (e.g., Biber, 2006), the proportions of academic disciplines remained fairly stable across the seven profile groups. This was not surprising, however, since TOP test takers are strongly encouraged to prepare an introductory level lecture and the topic selection is not limited to their own academic discipline. Consequently, it was not uncommon to see the same topic presented by, say, a computer science student and an economics student.

The application of the aforementioned selection criteria identified 82 test takers who constituted a representative subset of the baseline sample. For simplicity, these 82 test takers will be collectively referred to as the corpus sample in the remainder of this study. All selected test takers were assigned a profile group membership with extremely low uncertainty. The average classification uncertainty in the corpus sample was as low as 0.03. Furthermore, the L1 composition of each profile group was successfully retained in the corpus sample. Table 5-1 gives the proportion of each L1 group in the corpus sample and the baseline sample.

Table 5-1

L1 Percent Proportions in the Corpus Sample

Profile Group	CJK	GEE	IND	ROM	OTH
Near-native	20 (17)	40 (42)	0 (9)	20 (12)	20 (21)
High Prof.	30 (33)	0 (9)	20 (16)	30 (23)	20 (19)
Intermediate Prof.	75 (78)	0 (1)	13 (5)	13 (7)	0 (9)
Low Prof.	80 (90)	0 (0)	0 (0)	10 (5)	10 (5)
Varying PR/LG	82 (73)	0 (1)	6 (7)	6 (9)	6 (10)
Average PR	50 (50)	0 (4)	20 (13)	20 (21)	10 (12)
Varying RO	80 (79)	0 (0)	0 (2)	10 (7)	10 (13)

Note. The figures in the parentheses indicate the corresponding percent proportions in the baseline sample of 765.

In addition, the proportion of each academic division in the corpus sample corresponded well to the baseline proportion, as can be seen in Table 5-2.

Table 5-2

Academic Division Proportions in the Corpus and the Baseline Samples

Academic Divisions	Corpus Sample Proportion (%)	Baseline Sample Proportion (%)
AHM	11.0	12.5
ENN	42.7	48.8
LHE	6.1	7.5
PHY	18.3	15.3
SOC	18.3	13.9
OTH	3.7	2.1

5.1.2 Transcription of the selected test taker performances

As noted in Chapter 3, the selected test takers' performances on the TOP lecture task were transcribed using the notations and conventions used in the CA literature (Sacks et al., 1974). The CA transcription method was chosen because its attention to details of speech events, such as speaker identification and overlapped speech, was expected to be useful in understanding the structure of the resulting transcripts. Test takers' use of whiteboard space was also noted in the

transcripts to help interpret reference markers such as *here* and *this*. The following excerpt from a transcribed test taker performance gives an example of the resulting transcripts:

```
Q1:     For control and order:?
TT:     Okay:? Okay, for control and order? Okay, so that's good. Any
        other reasons?
Q1:     ((long pause)) Good question. [((laughs))
TT:                                   [((laughs)) (2.0) Yeah, I mean it
        is a good question. You're right, you know, because .h that's
        what um political philosophy helps us to do:, like it helps us
        ask the very basic questions. Like .h why do we have the system
        that we have, .h you know it helps us kind of like tear down all
```

Note. Q1 = Questioner #1; TT = Test Taker.

While the CA transcription method was helpful in capturing the details of the test takers' performances and their interaction with the TOP questioners, the resulting transcripts were not particularly suited for the purpose of corpus construction. This was mainly due to the heavy formatting and nonconventional spellings of the CA transcription method. Consequently, the resulting transcripts were converted into a simple text file format. The converted transcripts adopted the conventional spelling to allow word-level searches for the subsequent corpus analysis. The contributions of the TOP questioners in the converted transcripts were tagged to distinguish their language use from test taker performances. Also tagged in the converted transcripts included pauses and whiteboard uses. In addition, each converted transcript contained a header indicating test taker L1, academic background, and the TOP decision category. The header information was used to construct a sub-corpus for each profile group.

5.1.3 Description of the TOP corpus

The converted text versions of the 82 transcripts constituted the TOP corpus. This small corpus consisted of 80,553 words spoken by test takers. As previously noted, the contribution of the TOP questioners was not considered as the main body of the corpus, and therefore, excluded

from the word count. There were seven sub-corpora within the TOP corpus, one for each subscale score profile group. Table 5-3 gives the word count of each profile group sub-corpus.

Table 5-3

Word Counts of Subscale Score Profile Group Sub-Corpora

Near-native	High Prof.	Inter. Prof.	Low Prof.	VarPRLG	AvgPR	VarRO
12,013 (10)	11,143 (10)	15,790 (16)	7,434 (10)	15,572 (16)	9,951 (10)	8,650 (10)

Note. The figures in the parentheses indicate the number of test takers in the corresponding profile group.

Table 5-3 shows that there were slight differences in the size of the sub-corpora. Average word count per test taker ranged from 743 to 1201. The order of the average word counts roughly corresponded to the order of the mean vectors reported in Table 4-11 in Chapter 4. That is, profile groups that had higher average model-based subscale scores tended to contain more words in the corresponding sub-corpora.

Most frequent words in the TOP corpus as a whole and in each sub-corpus were also examined. Table 5-4 gives the ten most frequently used words in each profile group sub-corpus and in the entire TOP corpus.

Table 5-4

The 10 Most Frequently Used Words in Each Profile Group Sub-corpus and the TOP corpus

Near-native	High Prof.	Inter. Prof.	Low Prof.	VarPRLG	AvgPR	VarRO	TOP
the	the	the	the	the	the	the	the
you	to	and	is	and	uh	uh	and
um	you	uh	of	is	so	is	is
to	and	so	a	to	to	and	uh
and	of	is	and	so	and	of	to
of	so	you	uh	of	is	this	of
that	it	to	it	you	of	a	so
uh	is	it	this	uh	a	to	you
is	a	this	so	this	s	you	a
a	s	a	in	it	this	so	this

Table 5-4 shows that all sub-corpora were highly comparable in terms of the frequently used words. The vast majority of the ten most frequent words in the entire corpus were shared by all sub-corpora, while relative standings of the frequent words only slightly varied across the sub-corpora.

5.2 The TOP corpus and MICASE

Comparing a newly developed corpus to a well-established corpus with a similar objective can help understand the characteristics of the new corpus and often provide a useful starting point for further analysis (Evison, 2010). In order to explore characteristic features of TOP test takers' language use, the TOP corpus was compared to a subset of Michigan Corpus of Academic Spoken English (MICASE) (Simpson et al., 2002), which is a large academic oral language corpus constructed and maintained by the University of Michigan, Ann Arbor (UMAA). This section reports the findings from a statistical comparison of word frequencies between the two corpora. This section begins with a brief introduction of MICASE, followed by a description

of the subset of MICASE that was used as the reference corpus. Lastly, the results of the comparison between the TOP corpus and the reference corpus are reported.

5.2.1 MICASE and its lecture subset

MICASE (Simpson et al., 2002) is an academic oral language corpus constructed by a research team at UMAA. It consists of approximately 1.8 million spoken words from various academic settings, including lectures, colloquia, discussion sections, dissertation defenses, and office hours, to name a few. As a large collection of spoken words from many academic settings, MICASE is one of the most representative academic oral language corpora publicly available, and has been used in a number of corpus linguistics studies.[8]

The large size and the academic nature of MICASE rendered it a natural reference corpus against which the TOP corpus could be compared. However, while the variety of academic settings included in MICASE is integral to its representativeness, it also made a direct comparison between the TOP corpus and MICASE difficult. In particular, MICASE included many settings that were not relevant to the TOP corpus, which consisted solely of simulated lectures that were highly monologic. Consequently, it was decided to extract a sub-corpus of MICASE that was most comparable to the TOP corpus and use that subset as the reference corpus.

The subset of MICASE was obtained using an interactive browsing tool implemented on the official MICASE website. Since the TOP corpus consisted of simulated lecture task performances, MICASE transcripts from academic settings other than lectures were excluded

[8] See the MICASE publication webpage (http://micase.elicorpora.info/micase-publications-and-presentations) for the complete list of studies using MICASE since 1999.

from the subset. Highly interactive lectures were further filtered out to match the predominantly monologic nature of the TOP corpus. Lastly, lectures given by non-native speakers of English were excluded. This last step allowed potential differences between the resulting MICASE subset and the TOP corpus to be attributed to the differences in English oral language proficiency of the speakers when appropriate. The resulting MICASE subset included the transcripts of 34 lectures from a variety of academic disciplines, and consisted of 177,119 words. For convenience, this subset of MICASE will be referred to as the MICASE lecture sub-corpus.

5.2.2 Word frequency comparison between the TOP corpus and the MICASE subset

The word lists of the TOP corpus and the MICASE lecture sub-corpus were compared to identify words that were considerably more or less frequently used in one corpus than in the other. The comparison was aided by the likelihood ratio test statistic,[9] as suggested by Dunning (1993). The use of the likelihood ratio test statistic in this context requires a distributional assumption on the occurrence of a word in a given text. In particular, it is assumed that the occurrence of a word in a given text independently follows a binomial distribution with the occurrence (or success) parameter p specific to the word and the trial parameter n equal to the number of words in the given text. The likelihood ratio statistic then compares the likelihood of having the same occurrence parameter p for a given word in two compared texts (which serves as the null model) to the likelihood of having two separate occurrences parameters, one for each text (which serves as the alternative model). The well-known asymptotic property of the likelihood ratio statistic (see, e.g., Bickel & Doksum, 1978) allows the resulting test statistic to be interpreted against the chi-square distribution with one degree of freedom. In the corpus

[9] AntConc presents this statistic under the name of log likelihood.

linguistics literature, the statistical comparison of word lists between two corpora based on the likelihood ratio test statistic is called the keyword analysis (e.g., Scott, 2013), and words that are associated with a significant result (i.e., *p* value less than .05) are called keywords.

Table 5-5 gives the list of 40 keywords[10] from the comparison between the TOP corpus and MICASE lecture sub-corpus. The first 20 keywords are positive keywords in that they were used significantly more frequently in the TOP corpus, while the remaining 20 keywords are negative keywords which were used significantly less frequently in the TOP corpus.

[10] The list of all significant keywords included more than 1,100 words, making it difficult to present the list in its entirety. The choice of 40 keywords, 20 for each positive and negative, was to focus on the most pronounced differences between the two corpora.

Table 5-5

The TOP Corpus vs. The MICASE Lecture Sub-corpus: Top 20 Positive and Negative Keywords

Positive Keywords		Negative Keywords	
Keyword	Likelihood Ratio	Keyword	Likelihood Ratio
yeah	487.38	that	186.84
so	479.04	species	158.49
like	361.67	was	156.11
uh	266.18	had	133.79
is	257.00	were	131.11
eh	252.07	women	115.59
um	225.99	he	95.81
voltage	183.72	sort	91.02
this	181.02	family	88.59
energy	174.31	things	86.88
can	170.48	of	82.02
will	154.26	been	66.86
current	153.26	those	61.45
x	118.20	on	61.29
output	114.33	who	60.37
transfer	100.33	at	57.90
heat	99.47	did	52.64
signal	99.47	really	49.50
confess	98.05	up	49.11
electrical	95.76	in	49.00

Note. All keywords in this table were significant at $\alpha < .05$.

Table 5-5 shows several interesting differences between the two corpora. The TOP corpus and the MICASE lecture sub-corpus differed mostly in terms of the use of discourse markers such as *yeah*, *so*, *like*, *that*, and *things*. Differences between the two corpora were also observed in the use of disfluency markers, including *uh*, *eh*, and *um*, all three of which were more frequently occurred in the TOP corpus. *Sort* provided another interesting contrast between the two corpora. In both corpora, *sort* was almost always observed as a part of *sort of*, which shares a formal and functional similarity with *kind of*. The TOP test takers clearly preferred *kind of* over *sort of* despite the similarity; the former saw almost 20 times more frequent uses than the latter in the

TOP corpus. On the other hand, the two expressions were used with almost identical frequencies in the MICASE lecture sub-corpus. Little is known about the difference between these two similar expressions, and therefore, it would be premature to interpret this contrast. However, such a difference in the usage frequencies of the two might be attributed to the difference in academic oral English proficiency of the contributors of the two corpora.

The positive content keywords showed the impacts of a few popular topics among the TOP test takers. Content words related to the field of Electrical Engineering, in particular, were ranked higher in the positive keyword list.[11] The vast majority of *he* in the MICASE lecture sub-corpus were used to refer to a scholar who was relevant to class discussions. The TOP test takers rarely introduced a specific theory, and therefore, it was not surprising to see *he* in the negative keyword list.

Having *will* in the positive keyword list and past tense *be* verbs in the negative keyword list constituted an interesting contrast. This can be interpreted in multiple ways. A contextual explanation has to do with the difference in the nature of the TOP lecture task and the real-classroom lectures in the MICASE lecture sub-corpus. While TOP test takers typically prepared a short introductory lecture appropriate for the first day of a course, none of the MICASE lectures were recorded on the first day. In this light, it was not surprising to observe more past tense usage in the MICASE lecture sub-corpus in which the instructors would have more need to refer back to an earlier part of the course. Another explanation, which involves the difference in within-speech organization, will be presented in detail in section 5.3.

[11] Since the introduction of the TOP in 2004, Electrical Engineering has been the biggest user of the test, accounting for more than 20 percent of all test takers.

As previously noted, the TOP corpus consisted of seven sub-corpora, which represented the oral language use of the seven subscale score profile groups. The word list of each sub-corpus was compared to the word list of the MICASE lecture sub-corpus to examine the word usage difference across different profile groups. Table 5-6 provides 10 most positive- and 10 most negative-keywords for each of the seven keyword comparisons.

Table 5-6

The Subscale Score Profile Group Sub-corpora vs. The MICASE Lecture Sub-corpus: Top 10 Positive and Negative Keywords

	Near-native	High Prof.	Inter. Prof.	Low Prof.	VarPRLG	AvgPR	VarRO
Positive Keywords	like	gdp	yeah	yeah	confess	so	team
	law	energy	like	decibel	so	china	energy
	islamic	layer	mantra	is	prisoner	laser	signal
	yeah	so	so	electrical	yeah	wave	output
	sunnah	like	voltage	x	data	voltage	traffic
	expense	transfer	device	region	is	car	displacement
	indo-european	lactate	current	barrier	like	antenna	bags
	japan	research	electric	friendship	can	beam	shoes
	quran	football	india	voltage	jail	input	direction
	know	mussorgsky	rat	current	x	yeah	is
Negative Keywords	he	sort	of	that	was	he	he
	again	things	that	was	he	that	was
	had	who	were	would	that	were	people
	his	that	at	at	were	was	that
	and	had	who	up	with	on	they
	here	and	in	time	things	who	with
	sort	i	was	had	his	had	would
	of	those	his	to	of	they	really
	up	they	been	well	really	out	could
	get	know	population	their	on	sort	things

Note. All keywords in this table were significant at $\alpha < .05$.

Table 5-6 shows that most of the positive keywords were content words. On the other hand, all negative keywords were function words. In particular, past-tense *be* verbs and pronouns were constantly less frequently occurred in the TOP corpus. General nouns (Halliday & Hasan, 1976) such as *things*, *way*, and *people* were also not as commonly used in the TOP corpus as in the MICASE lecture sub-corpus. It is well known that both pronouns and general nouns serve referential functions that are integral for micro-level discourse organization. The relative lack of pronouns and general nouns in the TOP corpus might indicate the relative weakness in micro-level organization of the TOP test taker performances compared to the real classroom discourse given by the MICASE instructors. However, such an interpretation is only one of many possible ones, and its validity should be evaluated in a more focused study. The Near-native group was the only exception in that its negative keywords did not include either the past-tense be verbs or pronouns, except for *he* and *his*. This could be interpreted as supporting evidence for the near-native interpretation of this group, while such an interpretation requires a more rigorous study to be validated.

In sum, the keyword comparison between the TOP corpus and the MICASE lecture sub-corpus showed a consistent pattern. In particular, the TOP test takers tended not to use certain function words as much as the MICASE instructors. Some of the less frequently used function words, such as *he*, *his*, and past tense *be* verbs, might be interpreted as the results of contextual differences between the two corpora. The subsequent keyword analyses for each of the TOP profile group sub-corpora suggested that the observed differences between the TOP corpus and the MICASE lecture sub-corpus were largely consistent across different profile groups, with the only exception being the Near-native profile group.

5.3 Discourse organizing lexical bundles in the TOP corpus

This section describes the procedures for identifying and analyzing discourse organizing lexical bundles in the TOP corpus and reports findings from the analysis. As noted in Chapter 3, this part of the study focused on metadiscourse and textual reference (MTR) bundles, which explicitly refer to "prior or upcoming discourse" (Simpson-Vlach & Ellis, 2010, p. 507), such as *I was saying* and *as I said*. This section begins with a description of how MTR bundles in the TOP corpus were identified. The identified MTR bundles are further classified depending on the timeframe of their reference point, as well as their formal appropriateness. The distribution of the classified MTR bundles across the subscale score profile groups is reported and interpreted.

5.3.1 The identification and coding of MTR bundles

The existing corpus studies on lexical bundles have relied heavily upon automatic *n*-gram finders for bundle identification. However, the nature of the TOP corpus made it crucial to complement automatic searches with manual reading of the transcripts. The TOP corpus consisted of words that were spoken by non-native speakers of English, with only a few exceptions in the Near-native group. It was not uncommon to find lexical bundles that functioned properly but that deviated from the normal grammatical form (e.g., *was talking it* for *was talking about it*). Furthermore, a number of bundles were separated by disfluency markers such as *uh* and *um* (e.g., *was eh talking uh um about* for *was talking about*). Consequently, using an automatic search for lexical bundles as a whole (e.g., using *was talking about* as a search query) would have not been effective in identifying MTR bundles in the TOP corpus.

The manual search began with the MTR bundles in the academic formula list suggested by Simpson-Vlach and Ellis (2010) and was expanded by adding items suggested in Biber et al.

(2004) and Nattinger and DeCarrico (1992). Instead of using an entire bundle as a search query, each element of a bundle was searched separately to include ill-formed and/or separated bundles. For example, search queries such as *talk*, *talked*, *talking*, and *about* were used to look for instances of *was talking about*. This search method inevitably yielded multiple hits for the same instance of a lexical bundle. For example, *was talking about* will appear in the search result for queries *talking* and *about*. In order to avoid multiple counts of a single bundle use, the resulting concordance lines from all searches were stored in a relational database developed for the purpose of this study. The database was equipped with test taker ID and neighboring words to provide a convenient means to examine search results without the danger of multiple counts. The stored concordance lines were manually examined by two readers[12] to determine whether a searched word belonged to a MTR bundle. There were several disagreements between the two readers' independent judgments, but these were resolved in an extra discussion session.

Despite its time-consuming and labor-intensive nature, this manual identification procedure had another advantage over the automatic n-gram search approach in that repetitions of the same bundle were counted only once. For example, repetitions such as *I was talking about uh um was talking about* were counted as a single instance of the bundle use. A purely automatic search would have counted such repetitions as two instances.

The aforementioned manual search procedure identified a total of 82 instances of MTR bundle use. The entire 82 MTR bundles with neighboring contexts are provided in Appendix G. The identified instances were then coded according to their timeframe. That is, MTR bundles referring back to a previous part of discourse were classified into the past frame, whereas MTR

[12] Both readers had backgrounds in Applied Linguistics and experience as TOP raters.

bundles referring to a forthcoming part of discourse were categorized as the future frame. This dichotomous timeframe coding was made in a straightforward fashion based on the context and the neighboring words. The two readers' independent judgments agreed on every case.

Since most of the TOP test takers were nonnative speakers of English, it was believed to be helpful to also examine the grammatical form of the MTR bundles included in the TOP corpus. Therefore, during the timeframe coding, it was also noted whether an MTR bundle was well-formed or not. For example, *was talking about* was coded as a well-formed instance, whereas *was talk about* was noted as an incorrect usage. This additional coding procedure was conducted in a straightforward manner, and the two readers again agreed on every instance.

5.3.2 Distributional characteristics of the MTR bundles

The number of the MTR bundles in each subscale score profile group was examined first. Table 5-7 provides the frequency counts of the MTR bundles in each profile group.

Table 5-7

Frequency Counts of MTR Bundles across the Subscale Score Profile Group Sub-corpora

Near-native	High Prof.	Inter. Prof.	Low Prof.	VarPRLG	AvgPR	VarRO
11	7	8 | 13	13	2 | 8	14	6

Note. For the Intermediate Proficiency and Varying PR/LG groups, the figures on the left- and right-side of the vertical bar give the frequency counts of the MTR bundles in the non-pass and pass subgroups, respectively.

Table 5-7 shows a fair amount of between-group variation in the MTR bundle frequencies. However, the frequency counts did not appear to be correlated with the overall model-based subscale scores. The Average Pronunciation group used the largest number of MTR bundles,

whereas the Near-native group and the High Proficiency group corpora contained fewer MTR bundles than many other groups.

The results of the timeframe coding showed that the MTR bundles in the TOP corpus referred to both past and future frames with almost identical frequencies. Out of 82 identified MTR bundles, 40 were designated as future reference bundles (e.g., *let's move to*, *will explain later*), and the remaining 42 as past reference bundles (e.g., *going back to*, *was telling you*). The majority of the MTR bundles were well formed. Only 22 out of 82 were noted as incorrect usages (e.g., *let's get a start* for *let's get started*). Figure 5-1 shows the distribution of each MTR bundle across the seven profile group sub-corpora with the timeframe and formal appropriateness information.

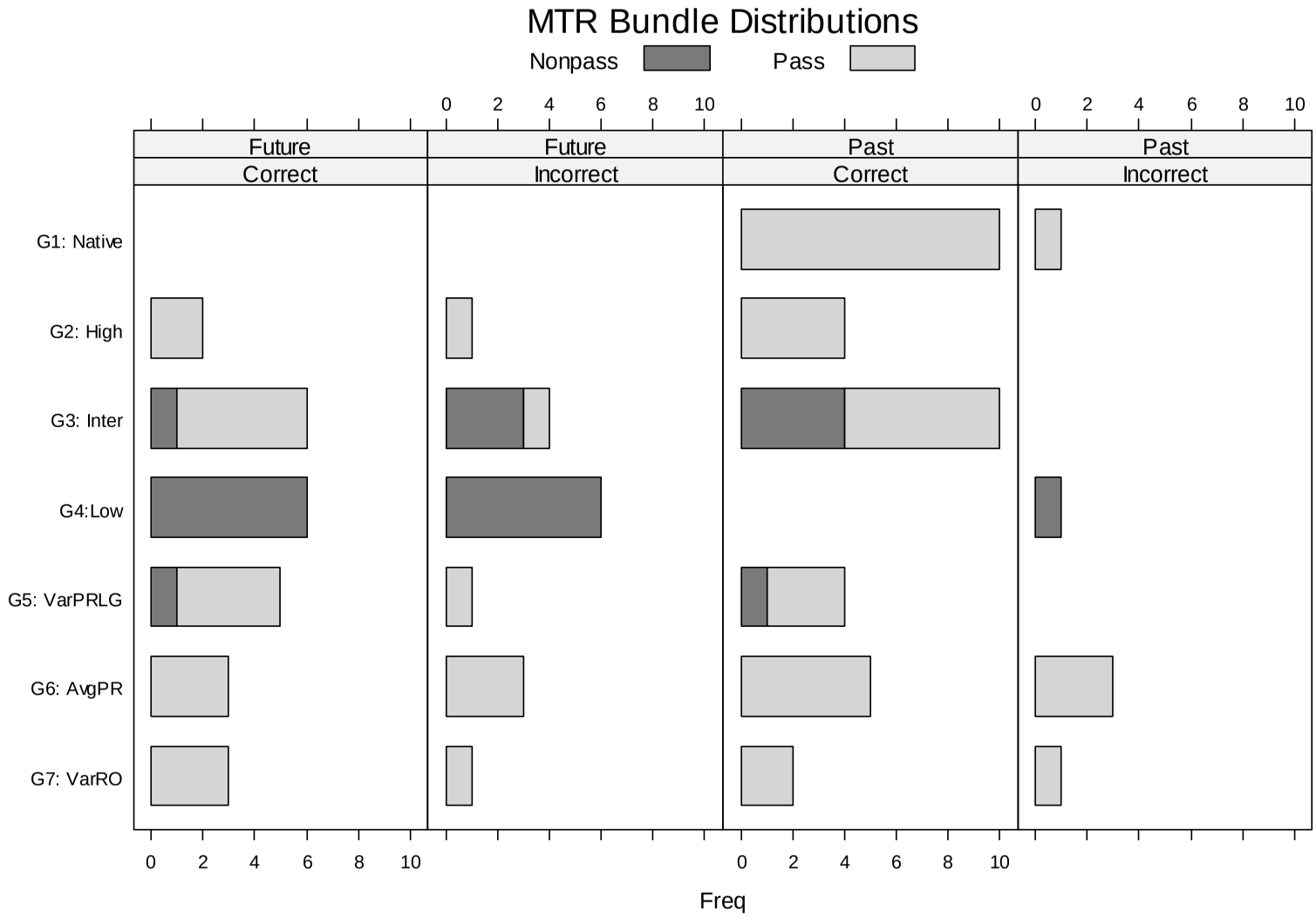


Figure 5-1. The frequency counts of MTR bundles with the timeframe and formal appropriateness coding information. Pass and Non-pass indicate the pass and non-pass (i.e., provisional pass and fail) decision categories, respectively.

Figure 5-1 shows that the MTR bundles were unevenly distributed across the subscale score profile groups. It is noteworthy that the Near-native group presented a complete contrast to the Low Proficiency group in terms of the MTR bundles' timeframe. While all MTR bundles in the Near-native group were employed to refer to a previous part of discourse, all but one MTR bundles in the Low Proficiency group turned out to be a future reference marker. Given the score profile differences between the two groups, this contrast could be interpreted as a reflection of the difference in academic oral English proficiency. Past-looking MTR bundles typically involve a past-tense verb, which might give an additional layer of difficulty to test takers with low academic oral English proficiency. The majority of the past timeframe bundles in the High Proficiency group might be taken as supporting evidence for this interpretation, although the small sample size would prohibit a strong interpretation.

The above interpretation did not provide a reason for the lack of forward-looking MTR bundles in the Near-native group. An inspection of the concordance lines and CA-style transcripts suggested a plausible explanation. A large number of MTR bundles referring to a forthcoming discourse part occurred when a test taker attempted to delay an answer after being asked a question, as can be seen in the following excerpt:

```
Q1:     Does the distance matter? Will it affect that(.) at all?
TT:     Mmhmm. In fact, uh:: (0.3) [p is always positive.
Q1:                                [Does it have to be-
Q1:     okay.
TT:     but q may be negative. (0.3) we will come to this soon. hhh
Q1:     hhh okay.
```

Note. Boldface italic parts indicate the bundle of interest.

It turned out that the ten members of the Near-native group in the TOP corpus answered all questions when the questions were asked, and therefore, the most common context for future reference MTR bundles did not occur.

Another potential problem with interpreting MTR timeframe occurrences as an indicator of oral proficiency was the large number of past-reference MTR bundles in the Intermediate Proficiency group. In fact, there were as many well-formed past-looking MTR bundles in the Intermediate Proficiency group as in the Near-native group. Even the non-pass members of the Intermediate Proficiency group used the past timeframe MTR bundles as often as did the High Proficiency group members. Again, the concordance lines and CA-transcripts turned out to be helpful in accounting for this deviation. Upon inspection of the contexts around the past-looking MTR bundles, it appeared that past-reference MTR bundles could be further distinguished into two types. Those that made an explicit reference to a specific part of the previous discourse were classified as Type 1, as illustrated in the following excerpt:

TT: ...first place, you know? Um:: you know, so, you know, one thing you were saying, ***to go back to the why question***, is you know we

Note. Boldface italic parts indicate the bundle of interest.

Past-looking MTR bundles that did not specify the point being referenced were classified as Type 2. A prototypical bundle in this category was *as I said*. An example of a Type 2 MTR bundle use is given in the following excerpt:

TT: … And I'll be talking about that in the date up on final presentation. (hh) so. ((clears throat)) so ***as I said*** like uh: the speed- the speed of the storage devices and the how fast the

Note. Boldface italic parts indicate the bundle of interest.

All 42 instances of past reference MTR bundles were thus classified into the two types. Table 5-8 shows the distribution of the two types in each profile group.

Table 5-8

Further Classification of Past Reference MTR Bundles

	Near-native	High Prof.	Inter. Prof.	Low Prof.	VarPRLG	AvgPR	VarRO	Totals
Type 1	7	2	1	1	0	2	0	13
Type 2	4	2	10	0	4	6	3	29

Table 5-8 presents two important points. First, by far, the majority of the past reference MTR bundles in the TOP corpus belonged to Type 2. That is, the test takers in the TOP corpus mostly used past-looking MTR bundles without an explicit reference to a point of the previous discourse. This might be due to the additional grammatical difficulty of marking an explicit reference point, which is discussed in the next paragraph. Second, the Near-native group alone saw more Type 1 MTR bundles than all the other groups combined. The numbers of past MTR bundles classified into Type 1 across the groups corresponded roughly to the order of the model-based subscale scores. This suggests that considering the two different types of past reference MTR bundles provides a better explanation of their relationship with model-based subscale score profiles. However, given the extremely small sample size, a strong interpretation of this relationship would be premature.

Classifying past reference MTR bundles into two types revealed another interesting phenomenon. While the majority of past-looking MTR bundles were Type 2, instances of incorrect grammar usage in past-looking MTR bundles occurred mostly as a result of failed attempts to use those that were Type 1, as can be seen in the following excerpt:

TT: So- (1.0) ***now that we are seen*** the relationships, we can- we can see different patterns of relationships. A- This is important

Note. Boldface italic parts indicate the bundle of interest.

This makes intuitive sense, in that the complexity of a bundle generally increases as its reference becomes more specific, and therefore, the likelihood of a grammatical error also increases. From a pedagogical point of view, this provides empirical evidence that some bundles are more difficult to learn that others. Simple MTR bundles such as *as I said* were seldom used incorrectly, and were often used by test takers with relatively low academic oral English proficiency. An inspection of MTR bundles with grammatical errors indicated that bundles containing *be* verbs and auxiliary verbs included many errors. Prepositions were also a frequent cause of errors. The small sample size made it difficult to generalize this finding, but a future study based on a much larger sample would be fruitful.

In sum, the investigation of the distributions of MTR bundles across the subscale score profile groups showed that the use of MTR bundles might be a distinguishing feature for TOP test takers who were classified into different profile groups. Future reference MTR bundles commonly occurred when a test taker tried to delay answering a question. The ability to use past-looking MTR bundles that had a specific reference point appeared to be correlated with the model-based subscale scores of TOP test takers. Test taker errors were more common with bundles that had specific reference points, while simple bundles were successfully used even by test takers with relatively low English proficiency.

5.4 Summary of the chapter

This chapter reported the findings of analyses addressing the second research question, "What are features of oral discourse that characterize test takers who have different subscale score profiles?" This chapter consisted of three main sections. The first section described the procedures of constructing the TOP corpus. The overall characteristics of the TOP test takers'

language use were then investigated by comparing the TOP corpus to a lecture sub-corpus selected from the MICASE corpus. The last section described the use of discourse organizing lexical bundles in the TOP corpus to explore relationships between the use of such bundles and subscale score profiles.

The TOP corpus was constructed based on the lecture task performances of 82 test takers. In order to obtain a representative corpus, the test takers included in the corpus construction were selected based on multiple criteria, including profile group membership, classification uncertainty, L1, academic background, and TOP decision categories. The performances of the 82 test takers were transcribed following the CA transcription method to capture the details of their performances and interactional events. The CA-style transcripts were then converted to unformatted text files to constitute the TOP corpus, which consisted of approximately 80,000 spoken words.

The TOP corpus was compared to the MICASE lecture sub-corpus to explore the characteristic features of TOP test takers' language use. In particular, frequently observed words in the two corpora were contrasted. The results showed that the TOP test takers tended to use fewer function words than the instructors included the MICASE lecture sub-corpus. Examples of such less frequently used function words included discourse markers, past-tense *be* verbs, pronouns, and general nouns. A series of profile group specific comparisons was made by examining differences between the word list of each profile group sub-corpus and that of the MICASE lecture sub-corpus. The results of these group-specific comparisons demonstrated that the relative lack of function words in the TOP test taker performances was a consistent phenomenon across all profile groups, with the only exception of the Near-native group.

Characteristic features of each profile group's discourse were investigated through an identification and examination of MTR bundles. Based on the list of MTR bundles suggested in the literature, a comprehensive search and manual concordance reading procedure were combined to identify instances of MTR bundles in the TOP corpus. Identified MTR bundles were then coded according to their timeframe and grammatical correctness. The distribution of the coded MTR bundles across the score profile groups revealed several interesting findings. MTR bundles referring to a previous point in the discourse were most frequently used by test takers who belonged to the Near-native group, whereas the vast majority of MTR bundles used by the Low Proficiency group members involved a future reference point. A further distinction among past reference MTR bundles suggested that the use of past MTR bundles with an explicit reference point might be related to test takers' academic oral English proficiency.

Chapter 6: Discussion and conclusion

This chapter summarizes the results reported in the previous chapters, discusses their implications, and provides suggestions for future studies. This chapter begins with a summary of the findings from analyses addressing the two research questions of this study. The implications of this study are discussed next. The chapter concludes with an acknowledgement of the limitations of this study and suggestions for future studies.

6.1 Summary of the findings

This study aimed to obtain information from the Test of Oral Proficiency (TOP) subscale scores that can be valuable to TOP stakeholders. The TOP subscale scores capture different aspects of test takers' academic oral English proficiency, and therefore, have potential for providing valuable feedback to TOP test takers, TOP administrators, and to ESL program administrators and instructors. Despite this potential, however, the TOP subscale scores are not reported or used. It was believed that an investigation of TOP subscale score profiles would yield more detailed information about test takers' performance in terms of the four TOP subscales, and to provide useful information that is not available in the current score report.

This study attempted to explore the profiles of the TOP test takers based on their subscale score patterns and to understand characteristic discourse features of test takers who belong to different profile groups. In particular, this study was guided by the following research questions:

1. What different subscale score profiles of academic English oral proficiency can be meaningfully produced by the TOP?

2. What features of oral discourse characterize test takers who have different subscale score profiles?

This study consisted of two data analysis stages. In the first stage, which addressed the first research question, the model-based subscale scores of 960 TOP test takers were estimated using the second-order graded response model (GRM) to account for structural dependencies and rater effects contained within the observed rater scores. Finite mixture (FM) models were then fitted to the resulting model-based subscale scores to identify groups of TOP test takers who exhibited homogeneous subscale score profiles. The subscale scores of the 960 test takers constituted the main source of data in this stage.

The second stage was designed to address the second research question, and therefore, focused on oral discourse features characterizing the performances of test takers who had different subscale score profiles. Videotaped performances of 82 test takers were transcribed, and the transcripts constituted a small oral corpus of the test taker performances. The resulting corpus was compared to the MICASE lecture sub-corpus in terms of frequently used words. In addition, the distribution of metadiscourse and textual reference (MTR) (Simpson-Vlach and Ellis, 2010) bundle usage across the profile groups was investigated.

6.1.1 Research question 1

The first research question, "What different subscale score profiles of academic English oral proficiency can be meaningfully produced by the TOP?", was addressed by a series of closely related analysis steps. The first step evaluated the parameter recovery of the second-order GRM. In the next step, the second-order GRM was fitted to the observed subscale scores of the 960

TOP test takers to estimate their model-based subscale scores. Lastly, FM models were employed to explore a group structure underlying the estimated model-based subscale scores.

The parameter recovery of the second-order GRM was evaluated through a small scale simulation study. There were three simulation settings, which were designed to present data structures from typical language performance assessments with varying levels of complexity. The true values of model parameters were generated using the second-order GRM as the true model. Model parameters were then estimated using Markov Chain Monte Carlo (MCMC) simulations, with the starting values obtained from maximum likelihood (ML) point estimates and standard errors. The results showed that all model parameters were successfully recovered in all three settings. In addition, the model estimates of test taker proficiency were more highly correlated with the true values than the standardized sum scores across all three settings, which demonstrated the added value of using the second-order GRM over a simple average scoring method in estimating the model-based subscale scores.

The results of the simulation study provided a justification for the use of the second-order GRM to estimate the model-based subscale scores of the 960 TOP test takers. MCMC was used as the main estimation method, with its starting values obtained from ML estimation results. The goodness of fit of the second-order GRM was evaluated in relation to the goodness of the fit of the hierarchical rater model (HRM) (Patz et al., 2002). The model fit comparison results showed that the second-order GRM fit the data better than the HRM. The model-based subscale scores estimated using the second-order GRM were compared to observed sum scores. The comparison showed that the two sets of scores were distributed in a comparable fashion across all subscales. However, the continuous nature of the model-based subscale scores allowed more fine-tuned

distinctions among takers, and therefore, provided data that were better suited for the subsequent FM analysis.

The model-based subscale scores from the second analysis step were used as the data for the FM analysis that explored profiles of subscale score patterns. FM models with four dimensional normal components were fit to the model-based subscale scores. The final model was selected based on the Bayesian Information Criterion (BIC) (Schwarz, 1978) and substantive interpretability. The selected model suggested seven profile groups. All seven groups were interpreted and labeled based on the characteristic subscale score patterns, linguistic background, and TOP decision categories of group members. The first group consisted of test takers with near-native academic oral English proficiency, and therefore, was labeled as the Near-native group. Groups 2, 3, and 4 comprised test takers with high-, intermediate-, and low- academic oral English proficiency, respectively. Group 5 was labeled as the Varying Pronunciation/Grammar group, since its members varied only in terms of Pronunciation and Lexical Grammar. Group 6 consisted of test takers whose Pronunciation scores were concentrated around the average, and therefore, was labeled as the Average Pronunciation group. Group 7 consisted of test takers who varied only in terms of Rhetorical Organization, and therefore, was labeled as the Varying Organization group.

6.1.2 Research Question 2

The second research question, "What features of oral discourse characterize test takers who have different subscale score profiles?", was addressed based on the TOP corpus, a small oral corpus consisting of TOP test taker performances. The TOP corpus was compared to a lecture sub-corpus selected from the MICASE corpus (Simpson et al., 2002) to investigate the overall

characteristics of the TOP test takers' oral language use. In addition, characteristic features of each profile group's oral discourse were explored through an identification and examination of discourse organizing lexical bundles.

The TOP corpus consisted of the lecture task performances of 82 TOP test takers. The 82 test takers included in the TOP corpus were selected as a representative subset of the entire 960 test takers. The selection criteria included profile group membership, the estimated uncertainty of group membership, L1, academic background, and TOP decision categories. The performances of the selected 82 test takers were transcribed according to the CA transcription convention (Sacks et al., 1974). Unformatted text files were then created from the CA transcripts for quick and straightforward searches of words and expressions in the resulting corpus. The unformatted transcripts constituted the TOP corpus, whose total word count was approximately 80,000.

The lists of most frequently occurring words in the TOP corpus and the MICASE lecture sub-corpus were contrasted to find potential differences in the language use patterns of the TOP test takers and the instructors in the MICASE sub-corpus. The results showed the relative lack of function word usage in the TOP corpus, indicating that the TOP test takers tended to use fewer function words than the MICASE instructors. The subsequent group-level comparisons demonstrated the following: the TOP test takers use relatively fewer function words consistently across all profile groups, with the only exception of the Near-native group.

Potential differences in discourse organization across different profile groups were investigated by identifying and examining the use of MTR bundles. The identification of MTR bundles in the TOP corpus began with the list of MTR bundles suggested in the literature, and was complemented by manual reading of the CA transcripts and concordance lines. Identified

MTR bundles were further examined in terms of their timeframe and grammatical appropriateness. The distribution of MTR bundles with different timeframes across the seven profile groups demonstrated a relationship between the use of past reference MTR bundles and the subscale score profiles. In particular, MTR bundles with an explicit past reference were mostly used by test takers who belonged to the Near-native and the High Proficiency groups, whereas the majority of MTR bundle use among the Low Proficiency group members referred to an upcoming point in discourse.

6.2 Implications of the findings

6.2.1 Implications for the TOP

This study was situated in the local context of the TOP. Consequently, the findings of this study have direct implications for the TOP and its stakeholders. First and foremost, the identification of distinct profiles based on TOP subscale score patterns provides information that can be used to evaluate the current standards and policy. As noted in Chapter 4, the High, Intermediate, and Low Proficiency groups corresponded to the three TOP decision categories. This can be considered as supporting evidence for the current TOP standards. The Intermediate Proficiency group provides further supporting evidence for the current policy in that a small improvement on any of the four subscales would place its typical members into the pass category. However, the Varying Pronunciation/Grammar group presents evidence against the same policy in that its members would benefit more from improvements on two specific subscales, namely Pronunciation and Lexical Grammar. It is noteworthy that the current TOP weighting system favors test takers who score highly on the Pronunciation subscale. However, the extent to which a high score on Pronunciation could compensate relatively low scores on the other scales was not

thoroughly investigated when the weights were determined. The members of the Average Pronunciation group provide a focused sample set to evaluate the current weighting system in that almost all of them passed despite relatively large variations on the other three subscales.

In addition, the findings of this study can provide valuable information to enhance ITA training. As described in Chapter 4, the final model assigned a profile group membership to most of the 960 test takers with a high level of confidence. The profile group information can be provided to each test taker as detailed feedback on his or her performance. Test takers who did not pass the test will be able to make informed decisions about appropriate ESL courses to take based on the profile information. Even test takers who passed the test can use the feedback to further improve their academic oral English proficiency. Furthermore, such detailed feedback will be beneficial for ESL instructors, who can use the information to plan and customize their classes. ESL courses can be arranged and developed based on the proportion of test takers who belonged to a specific profile group. For example, the results of the FM analysis showed that a large proportion of test takers who did not pass the test would benefit most from ESL courses focusing on Pronunciation or Lexical Grammar.

One of the practical difficulties for the TOP program has been the lack of an effective means to estimate the relative severity and discrimination of an individual rater. The TOP employs an incomplete block design for rater assignment, with two raters per test taker. Therefore, the scores from one rater can only be directly compared to those of the other rater in a given pair. While variations in rater severity and discrimination were not part of this study, and it did not specifically investigate variations across raters, the second-order GRM used in this study can address this issue directly by providing an estimate of rater severity and discrimination with the

corresponding standard errors. The rater severity and discrimination estimates can provide information about how lenient or severe a rater was relative to the other raters, and how large random fluctuations in a rater's score assignment were, respectively. It is straightforward to use the estimates of rater severity and discrimination parameters as feedback for individual raters in rater training and norming.

The relationship between the past reference MTR bundle use and the academic oral English proficiency presents another practical implication for TOP. This relationship can be used to improve the scoring rubric for Rhetorical Organization and to design effective test tasks. Raters have had difficulty in evaluating test takers' Rhetorical Organization, in large part because of the relative ambiguity of the current scoring rubric for that subscale, which is provided in Appendix H. Appropriate use of the past-reference MTR bundles could be included in the scoring rubric to make it more concrete. The relationship could also be utilized to improve the current TOP tasks or to develop a new task aimed at eliciting the use of past reference MTR bundles. However, it should be noted that the relationship between the past reference MTR bundle use and the academic oral English proficiency was examined based a small sample, and therefore, it should be validated in future studies before being used to implement changes in the TOP.

6.2.2 Methodological implications

This study generalized the GRM (Samejima, 1969) in an attempt to address dependencies and rater effects inherent in the TOP rater scores. The resulting model, namely the second-order GRM, can be applied to any performance assessment contexts that involve multiple tasks rated by multiple raters. Since the second-order GRM draws upon the models for a single task rated by multiple raters (Baldwin et al., 2009; Johnson, 1996), it can be easily understood as the

continuation of the existing models. In addition, the transformation of the second-order GRM into a restricted testlet model described in Chapter 3 allows efficient ML estimation based on the bifactor dimension reduction (Gibbons & Hedecker, 1992). The resulting ML point and standard error estimates can be used on their own as final estimates, or utilized to obtain sound starting values for MCMC estimation. The transformation also reveals a conceptual relationship between items in a testlet-based test and raters in a multiple-task, multiple rater assessment, and provides a straightforward solution to address incomplete block designs with a large amount of missing data.

This study employed a combination of the second-order GRM and FM models to investigate subscale score profiles. While this combination arose within the context of the TOP, it can be applied to any language performance assessment context to explore group structures in terms of subscale profiles. Language performance assessments typically involve a small number of tasks that are designed to capture a large number of different skills, which makes it difficult to apply the standard cognitive diagnostic assessment (CDA) models to obtain skill profiles. The combination of the second-order GRM and a multivariate normal mixture model provides a practical alternative to the CDA models in that this combination is also capable of generating profile information based on subscale score patterns.

The identification of MTR bundles in this study relied upon manual reading of the CA transcripts as well as the search results from the TOP corpus. This study effectively utilized two different versions of transcripts to combine the advantages that each of them provided. While the CA transcripts were better suited for retaining the details of language use in test taker performances, constructing a corpus that allows straightforward search for a given word or

phrase with them presented a challenge. This study tackled this challenge by creating another version of the CA transcripts that were simple and unformatted text files. By constructing the TOP with the unformatted text files, the resulting corpus could accommodate a variety of search queries. The CA-transcripts were tied to each text file in a relational database, which allowed a more context-rich reading of the original CA transcript when necessary. Given the importance of context-rich reading in studies based on relatively small corpora, the specific combination of the CA transcripts and their unformatted versions appears to be a promising approach to constructing a new corpus, especially when the resulting corpus is small.

6.3 Limitations and suggestions for future studies

6.3.1 Limitations

This study was situated in a local setting and was exploratory in nature. Therefore, some findings may not be generalizable beyond the local setting, and any attempts to generalize the findings must be validated based on further research. The seven profile groups identified in Chapter 4 are specific to the TOP context. While the methodologies employed to identify those profile groups can be applied to many other, similar settings, the specific structure of the profile groups and their characteristics are not intended to generalize. Furthermore, it should be noted that the findings from the corpus-based analysis were based on a small sample. While extensive efforts were made to assure that the TOP corpus was representative of the TOP test taker population, results based on small samples are inherently subject to random and systematic fluctuations. Consequently, further investigations based on a more comprehensive collection of the TOP test taker performances would be useful to validate the corpus-based findings in this study. In addition, this study was certainly limited in that it focused only on the MTR bundles

and their timeframe and grammatical correctness. The MTR bundles investigated in this study are only one of many possible types of lexical bundles that function as discourse organizers in academic settings.

While the combination of second-order GRM and FM models presents an effective and practical solution to explore different profiles in terms of subscale score patterns, there were a few choices in its implementation that were not theoretically ideal. As mentioned in Chapter 4, the model-based subscale scores were estimated by fitting the second-order GRM to the scores on each subscale, resulting in four parallel univariate models. A more desirable approach would involve the multivariate generalization of the second-order GRM, which was briefly described in Chapter 4. In addition, the lack of an absolute fit index for the second-order GRM made it difficult to gauge the goodness of fit of the model. The estimated model-based subscale scores were used as the data for the FM model. It is noteworthy that this procedure entailed ignoring the uncertainty around the model-based subscale scores. While this choice was a practical necessity given the difficulty of fitting mixtures of the second-order GRM, it would be desirable to find a way to properly incorporate the uncertainty in estimating the FM model.

6.3.2 Suggestions for future studies

The aforementioned limitations offer suggestions for future research efforts. It appears that investigating the use of MTR bundles and its relationship with academic oral English proficiency presents a promising line of research. There are large oral corpora that capture academic English in university settings, upon which such investigations could be conducted. The timeframe coding scheme used in this study could be applied to a larger corpus in a straightforward manner. Focusing on other elements of academic oral discourse organization could lead to another line of

promising research. Based on a number of lexical bundles that serve referential and discourse organizing functions suggested in the literature, and an empirical investigation of their relationships to academic oral proficiency would be fruitful. The organization of academic discourse in terms of pragmatic functions also appears to deserve further investigation; there are several schemes for coding pragmatic functions in academic discourse (Maynard & Leicher, 2007; Nesi, Ahmad & Ibrahim, 2009). At the same time, it should be noted that the difficulty of applying such simplified coding schemes to complex language use in academic discourse might present a challenge.

The generalization of the second-order GRM to a multivariate model presents a natural extension of this study. The resulting model could provide a methodological tool for handling data from analytic scoring schemes that are frequently used in language performance assessments (Luoma, 2004; Weigle, 2002). Given the complexity of the model, however, the multivariate generalization should be attempted based on a solid understanding of rating processes. Unfortunately, little is known about raters' scoring processes based on analytic scoring schemes. Therefore, a qualitative investigation of the rating process would be the first step towards the multivariate generalization of the second-order GRM.

APPENDICES

Appendix A: R syntax for the simulation data generation (Setting 1)

```
## simulation setup
nper <- 960 # Number of test takers (i = 1, ..., N)
ntask <- 5 # Number of tasks (j = 1, ..., J)
nrater <- 5 # Number of raters (r = 1, ..., R)
ncat <- 4 # Number of scoring categories (k = 1,...,K)

## True Values
mu.theta <- 0
sd.theta <- sqrt(8)
mu.a <- 0
sd.a <- 1
mu.u <- 0
sd.u <- 1
mu.tau <- 0
sd.tau <- 1
mu.gamma <- 0
sd.gamma <- 1

## calling msm for drawing samples from truncated normal distributions
library(msm)

set.seed(1223)
## Draw the true parameters
theta <- rnorm(nper, mu.theta, sd.theta)
theta <- matrix(theta, nrow = nper, ncol = 1)
a <- rtnorm(nrater, mu.a, sd.a, lower = 0)
a <- matrix(a, nrow = nrater, ncol = 1)
```

```
u <- rnorm(nper * ntask, mu.u, sd.u)
u <- matrix(u, nrow = nper, ncol = ntask)
tau <- rnorm(nrater - 1, mu.tau, sd.tau)
tau <- matrix(c(1, tau), nrow = nrater, ncol = 1)
gamma <- matrix(rep(0, nrater * (ncat - 1)), nrow = nrater, ncol = ncat - 1)
for (r in 1:nrater) {
      gtemp <- rnorm(ncat - 2, mu.gamma, sd.gamma)
      gamma[r,] <- sort(c(gtemp, -sum(gtemp)))
      }
prsco <- rep(0, ntask * ncat * nrater * nper)
dim(prsco) <- c(nper, ntask, nrater, ncat)

# Calculate model implied cumulative probabilities
for (i in 1:nper){
      for (j in 1:ntask){
            for (r in 1:nrater){
                  for (k in 1:(ncat-1)){
                        exp.in <- gamma[r,k] - a[r] * (theta[i] + u[i,j] -
tau[r])
                        prsco[i,j,r,k] <- exp(exp.in)/(1 + exp(exp.in))
                  }
                  prsco[i,j,r,ncat] <- 1
            }
      }
}
prob <- rep(0, ntask * ncat * nrater * nper)
dim(prob) <- c(nper, ntask, nrater, ncat)

# Calculate model implied probabilities per each scoring category
for (i in 1:nper){
```

```
      for (j in 1:ntask){
            for (r in 1:nrater){
                  prob[i,j,r,1] <- prsco[i,j,r,1]
                  for (k in 2:ncat){
                        prob[i,j,r,k] <- prsco[i,j,r,k] - prsco[i,j,r,k-1]
                  }
            }
      }
}
gendata <- rep(0, nper * ntask * nrater)
dim(gendata) <- c(nper, ntask, nrater)
# Generating observed rater scores
for (i in 1:nper){
      for (j in 1:ntask){
            for (r in 1:nrater){
                  gendata[i,j,r] <- sample(1:ncat, 1, prob = prob[i,j,r,])
            }
      }
}
```

Appendix B: flexMIRT and WinBUGS syntax for the simulation estimation (Setting 1)

flexMIRT

```
<Project>

Title = "Simulation /w 5 tasks";

Description = "One general factor, 5 tasks (testlets) - 2013/02/23";

<Options>

 Mode = Calibration;

 saveSCO = YES;

 Score = EAP;

 GOF = Complete;

 M2 = Ordinal;

 Processors = 4;

<Groups>

%UNI%

  File = "gendata_flexMirt_format.dat";

  Dimensions = 6;

  Primary = 1;

  Varnames = I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1,
I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4,
I5R5;

  N = 960;

  Code(I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1, I3R2,
I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4, I5R5)
= (1,2,3,4),(0,1,2,3);

  Ncats(I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1,
I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4,
I5R5) = 4;

  Model(I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1,
I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4,
I5R5) = Graded(4);

  BetaPriors(I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1,
I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4,
I5R5) = 1.5;
```

```
<Constraints>

  // Bifactor structure

  Fix  UNI,(I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1,
I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4,
I5R5),Slope;

  Free UNI,(I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5, I3R1,
I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3, I5R4,
I5R5),Slope(1);

  Free UNI,(I1R1, I1R2, I1R3, I1R4, I1R5),Slope(2);

  Free UNI,(I2R1, I2R2, I2R3, I2R4, I2R5),Slope(3);

  Free UNI,(I3R1, I3R2, I3R3, I3R4, I3R5),Slope(4);

  Free UNI,(I4R1, I4R2, I4R3, I4R4, I4R5),Slope(5);

  Free UNI,(I5R1, I5R2, I5R3, I5R4, I5R5),Slope(6);

  // intercepts equality restrictions within the same rater

  Equal UNI,(I1R1, I1R2, I1R3, I1R4, I1R5),Intercept:

        UNI,(I2R1, I2R2, I2R3, I2R4, I2R5),Intercept:

      UNI,(I3R1, I3R2, I3R3, I3R4, I3R5),Intercept:

      UNI,(I4R1, I4R2, I4R3, I4R4, I4R5),Intercept:

      UNI,(I5R1, I5R2, I5R3, I5R4, I5R5),Intercept;

  // slope proportionality constraint

  Equal UNI,(I1R1, I1R2, I1R3, I1R4, I1R5),Slope(1):

      UNI,(I2R1, I2R2, I2R3, I2R4, I2R5),Slope(1):

      UNI,(I3R1, I3R2, I3R3, I3R4, I3R5),Slope(1):

      UNI,(I4R1, I4R2, I4R3, I4R4, I4R5),Slope(1):

      UNI,(I5R1, I5R2, I5R3, I5R4, I5R5),Slope(1):

      UNI,(I1R1, I1R2, I1R3, I1R4, I1R5),Slope(2):

      UNI,(I2R1, I2R2, I2R3, I2R4, I2R5),Slope(3):

      UNI,(I3R1, I3R2, I3R3, I3R4, I3R5),Slope(4):

      UNI,(I4R1, I4R2, I4R3, I4R4, I4R5),Slope(5):

      UNI,(I5R1, I5R2, I5R3, I5R4, I5R5),Slope(6);

  Fix UNI,Mean(1);

  Fix UNI,Mean(2);
```

```
  Fix UNI,Mean(3);

  Fix UNI,Mean(4);

  Fix UNI,Mean(5);

  Fix UNI,Mean(6);

  Free UNI, Cov(1,1);

  Value UNI, Cov(2,2), 1.0;

  Value UNI, Cov(3,3), 1.0;

  Value UNI, Cov(4,4), 1.0;

  Value UNI, Cov(5,5), 1.0;

  Value UNI, Cov(6,6), 1.0;

  Prior UNI, (I1R1, I1R2, I1R3, I1R4, I1R5, I2R1, I2R2, I2R3, I2R4, I2R5,
I3R1, I3R2, I3R3, I3R4, I3R5, I4R1, I4R2, I4R3, I4R4, I4R5, I5R1, I5R2, I5R3,
I5R4, I5R5),Slope(1) : logNormal(0,0.25);

  Prior UNI, (I1R1, I1R2, I1R3, I1R4, I1R5),Slope(2) : logNormal(0,0.25);

  Prior UNI, (I2R1, I2R2, I2R3, I2R4, I2R5),Slope(3) : logNormal(0,0.25);

  Prior UNI, (I3R1, I3R2, I3R3, I3R4, I3R5),Slope(4) : logNormal(0,0.25);

  Prior UNI, (I4R1, I4R2, I4R3, I4R4, I4R5),Slope(5) : logNormal(0,0.25);

  Prior UNI, (I5R1, I5R2, I5R3, I5R4, I5R5),Slope(6) : logNormal(0,0.25);
```

WinBUGS

```
model{
      for (i in 1:N){
            for (j in 1:J){
                  for (r in 1:R){
                        ## Observed Data Generating Probability
                        Y[i,j,r] ~ dcat(prob[i,j,r,1:K])
                        ## Computing elements for the logit likelihood
                  au[i,j,r] <- a[r] * u[i,j]
                  aall[i,j,r] <- a[r] * theta[i] + au[i,j,r] - a[r] * tau[r]
                        for (k in 1:(K-1)){
                              ## Logit link function and the model
                              logit(P[i,j,r,k]) <- gamma[r,k] - aall[i,j,r]
                        }
                  P[i,j,r,K] <- 1.0
                  }
            u[i,j] ~ dnorm(0.0, 1.0)
            }
            ## Prior for theta: Normal(0,1)
            theta[i] ~ dnorm(0.0 , pr.theta)
            ## Differences between neighboring probabilities
            for (j in 1:J){
                  for (r in 1:R){
                        prob[i,j,r,1] <- P[i,j,r,1]
                        for (k in 2:K){
                              prob[i,j,r,k] <- P[i,j,r,k] - P[i,j,r,k-1]
                        }
                  }
            }
      }
```

```
      ## In this parameterization, lambda[j] is set to 1 for all j
      ## to estimate the variance of theta
      ## Prior for lambda: Normal(0,1) with truncated at 0 (only positive)
      # lambda[1] ~ dnorm(0.0, 1.0)I(0,)
      ## Equality constraints for lambdas (only if j < 3)
      ##for (j in 2:J){
      ##    lambda[j] <- lambda[1]
      #     }
## Prior for rater discrimination parameter: truncated normal (0,1) I(0,)
      for (r in 1:R){
            a[r] ~ dnorm(0,0.1)I(0,)
      }
## Prior for rater overall severity parameter: Normal(0,1)
## tau[1] is set to one for identification (arbitrary)
      tau[1] <- 1
      for (r in 2:R){
            tau[r] ~ dnorm(0,0.1)
      }
      for (r in 1:R){
            for (k in 1:(K-2)){
                  gamma.star[r,k] ~ dnorm(m.gamma,pr.gamma)
                  gamma[r,k] <- ranked(gamma.star[r,1:(K-2)],k)
            }
      # Gamma sum-to-zero Constraint for identification
      gamma[r, K-1] <- -sum( gamma[r, 1:(K-2)] )
      }
      pr.gamma <- pow(s.gamma, -2)
## Hyperprior for theta precision: Gamma(1,1)
      pr.theta ~ dgamma(1,1)
}
```

Appendix C: flexMIRT syntax for the ML estimation of model-based subscale scores (Lexical Grammar)

```
<Project>
Title = "TOP Second Order GRM: LG";
Description = "TOP Lexical Grammar";
<Options>
  Mode = Calibration;
  Quadrature = 21,6.0;
  Etol = 1e-5;
  MAXE = 15000;
  MAXM = 10000;
  GOF = Complete;
  M2 = Full;
  SaveSCO = Yes;
  Score = EAP;
  Processors = 4;
  FactorLoadings = Yes;
<Groups>
%LG%
  File = "TOP_LG_only_higher_order_3cat.dat";
  Missing = 9;
  Varnames = TTakerID, I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20,
             I1R21, I1R30, I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46,
             I1R48, I1R56, I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20,
             I2R21, I2R30, I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46,
             I2R48, I2R56;
  Select = I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20,
             I1R21, I1R30, I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46,
             I1R48, I1R56, I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20,
             I2R21, I2R30, I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46,
```

```
          I2R48, I2R56;

  CaseID = TTakerID;

  Dimensions = 3;

  Primary = 1;

  N = 960;

  Ncats(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20,

          I1R21, I1R30, I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46,

          I1R48, I1R56, I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20,

          I2R21, I2R30, I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46,

          I2R48, I2R56) = 3;

  Model(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20,

          I1R21, I1R30, I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46,

          I1R48, I1R56, I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20,

          I2R21, I2R30, I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46,

          I2R48, I2R56) = Graded(3);

  BetaPriors(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20,

          I1R21, I1R30, I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46,

          I1R48, I1R56, I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20,

          I2R21, I2R30, I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46,

          I2R48, I2R56) = 1.5;

<Constraints>

  // Bifactor structure

  Fix  LG,(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56, I2R02, I2R05,
I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30, I2R31, I2R32, I2R36, I2R40,
I2R44, I2R45, I2R46, I2R48, I2R56),Slope;

  Free LG,(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56, I2R02, I2R05,
I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30, I2R31, I2R32, I2R36, I2R40,
I2R44, I2R45, I2R46, I2R48, I2R56),Slope(1);

  Free LG,(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56),Slope(2);
```

```
  Free LG,(I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30,
I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46, I2R48, I2R56),Slope(3);

  // intercepts equality restrictions within the same rater

  Equal LG,(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56),Intercept:

        LG,(I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30,
I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46, I2R48, I2R56),Intercept;

  // slope proportionality constraint

  Equal LG,(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56),Slope(1):

      LG,(I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30,
I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46, I2R48, I2R56),Slope(1):

        LG,(I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56),Slope(2):

      LG,(I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30,
I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46, I2R48, I2R56),Slope(3);

  Fix LG,Mean(1);

  Fix LG,Mean(2);

  Fix LG,Mean(3);

  Free LG, Cov(1,1);

  Value LG, Cov(2,2), 1.0;

  Value LG, Cov(3,3), 1.0;

  Prior LG, (I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56, I2R02, I2R05,
I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30, I2R31, I2R32, I2R36, I2R40,
I2R44, I2R45, I2R46, I2R48, I2R56),Slope(1) : logNormal(0,0.25);

  Prior LG, (I1R02, I1R05, I1R11, I1R12, I1R18, I1R19, I1R20, I1R21, I1R30,
I1R31, I1R32, I1R36, I1R40, I1R44, I1R45, I1R46, I1R48, I1R56),Slope(2) :
logNormal(0,0.25);

  Prior LG, (I2R02, I2R05, I2R11, I2R12, I2R18, I2R19, I2R20, I2R21, I2R30,
I2R31, I2R32, I2R36, I2R40, I2R44, I2R45, I2R46, I2R48, I2R56),Slope(3) :
logNormal(0,0.25);
```

Appendix D: WinBUGS syntax for the MCMC estimation of model-based subscale scores (Lexical Grammar)

Note. Model likelihood is the same as the simulation setting in Appendix C, and therefore, not shown here.

```
for (r in 1:R){
      a[r] ~ dnorm(0,0.1)I(0,)
}
tau[1] <- 0
for (r in 2:R){
      tau[r] ~ dnorm(0,0.1)
}
for (r in 1:R){
      for (k in 1:(K-2)){
            gamma.star[r,k] ~ dnorm(m.gamma,pr.gamma)
            gamma[r,k] <- ranked(gamma.star[r,1:(K-2)],k)
      }
gamma[r, K-1] <- -sum( gamma[r, 1:(K-2)] )
}
pr.gamma <- pow(s.gamma, -2)
pr.theta ~ dgamma(1,1)
```

Appendix E: Rater parameter point estimates and standard errors

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
a_1	0.65 (0.13)	0.75 (0.16)	1.78 (0.61)	2.58 (0.79)
a_2	0.44 (0.07)	0.61 (0.1)	1.11 (0.22)	0.69 (0.13)
a_3	0.66 (0.18)	0.95 (0.26)	0.99 (0.32)	0.96 (0.31)
a_4	0.67 (0.13)	0.98 (0.2)	0.73 (0.16)	0.92 (0.21)
a_5	0.9 (0.16)	0.84 (0.16)	0.65 (0.15)	0.93 (0.2)
a_6	0.59 (0.13)	0.82 (0.17)	1.01 (0.25)	0.91 (0.24)
a_7	0.46 (0.09)	0.71 (0.16)	0.56 (0.13)	0.4 (0.09)
a_8	0.83 (0.17)	0.93 (0.18)	1.25 (0.28)	0.99 (0.22)
a_9	1 (0.18)	0.9 (0.13)	1.29 (0.23)	0.66 (0.11)
a_{10}	0.9 (0.14)	0.74 (0.1)	0.71 (0.1)	0.6 (0.09)
a_{11}	0.88 (0.18)	0.89 (0.17)	1.24 (0.3)	0.97 (0.22)
a_{12}	0.63 (0.15)	0.47 (0.11)	1.41 (0.37)	1.18 (0.3)
a_{13}	0.86 (0.17)	0.61 (0.13)	1.22 (0.3)	0.81 (0.2)
a_{14}	0.94 (0.18)	1.09 (0.22)	1.03 (0.23)	0.83 (0.19)
a_{15}	0.98 (0.22)	0.87 (0.2)	1.02 (0.28)	0.68 (0.19)
a_{16}	0.65 (0.13)	1 (0.2)	0.67 (0.14)	0.66 (0.12)
a_{17}	0.75 (0.18)	1.04 (0.23)	0.73 (0.21)	0.86 (0.24)
a_{18}	0.89 (0.19)	1.02 (0.24)	1.24 (0.29)	1.19 (0.27)
γ_{11}	-5.37 (0.75)	-4.15 (0.63)	-3.35 (0.96)	-3.55 (1)
γ_{12}	-1.49 (0.38)	-1.1 (0.31)	-1.33 (0.43)	-1.66 (0.55)
γ_{21}	-5.19 (0.41)	-5.87 (0.8)	-6.39 (0.96)	-5.09 (0.64)
γ_{22}	0.26 (0.19)	0.24 (0.38)	-0.84 (0.47)	-1.1 (0.31)
γ_{31}	-4.98 (0.94)	-5.07 (1.03)	-4.26 (1.02)	-4.04 (1)
γ_{32}	-0.05 (0.52)	-0.15 (0.56)	-0.95 (0.59)	-1.22 (0.67)
γ_{41}	-5.56 (0.7)	-6.01 (0.98)	-4.39 (0.79)	-4.91 (0.9)
γ_{42}	-0.62 (0.32)	-0.66 (0.5)	-0.29 (0.42)	-1.58 (0.55)
γ_{51}	-8.97 (1.01)	-6.71 (0.94)	-4.67 (0.78)	-5.74 (0.94)
γ_{52}	1.23 (0.44)	-0.02 (0.44)	-0.58 (0.42)	-1.5 (0.5)
γ_{61}	-5.13 (0.83)	-5.81 (0.89)	-5.04 (0.98)	-4.85 (1.02)
γ_{62}	0.04 (0.41)	-0.48 (0.48)	-1.78 (0.62)	-1.73 (0.64)
γ_{71}	-4.27 (0.46)	-5.29 (0.77)	-3.55 (0.48)	-3.17 (0.42)
γ_{72}	-0.32 (0.22)	-0.41 (0.3)	-0.53 (0.26)	-0.61 (0.23)
γ_{81}	-6.68 (0.85)	-6.27 (0.93)	-5.59 (0.98)	-5.6 (0.94)
γ_{82}	-0.67 (0.42)	-0.88 (0.5)	-1.12 (0.55)	-1.06 (0.52)
γ_{91}	-7.7 (0.84)	-6.91 (0.82)	-6.9 (0.88)	-5.12 (0.81)
γ_{92}	-0.36 (0.31)	0.4 (0.38)	-0.72 (0.43)	-1.02 (0.42)
$\gamma_{10,1}$	-7.51 (0.63)	-6.12 (0.61)	-4.78 (0.48)	-4.32 (0.42)
$\gamma_{10,2}$	-0.26 (0.22)	0.2 (0.28)	-0.16 (0.23)	-1.22 (0.22)
$\gamma_{11,1}$	-7.21 (0.99)	-6.11 (0.95)	-5.18 (0.96)	-4.9 (0.83)
$\gamma_{11,2}$	0.16 (0.43)	-0.53 (0.5)	-0.89 (0.54)	-1.26 (0.46)

$\gamma_{12,1}$	-4.79 (0.79)	-3.79 (0.75)	-5.09 (1.03)	-5.42 (1.07)
$\gamma_{12,2}$	-0.84 (0.42)	-0.81 (0.4)	-1.43 (0.65)	-1.65 (0.68)
$\gamma_{13,1}$	-7.9 (1.06)	-5.5 (0.89)	-5.65 (0.96)	-5.3 (0.93)
$\gamma_{13,2}$	0.5 (0.41)	-0.71 (0.46)	-0.63 (0.43)	-0.74 (0.53)
$\gamma_{14,1}$	-7.9 (1.02)	-6.16 (0.96)	-5.09 (0.83)	-5.34 (0.91)
$\gamma_{14,2}$	-0.35 (0.45)	-1.05 (0.51)	-0.19 (0.37)	-0.93 (0.52)
$\gamma_{15,1}$	-6.53 (1.04)	-5.68 (0.99)	-5.35 (1.01)	-4.71 (0.93)
$\gamma_{15,2}$	-1.26 (0.55)	-1.97 (0.66)	-0.87 (0.54)	-1.61 (0.56)
$\gamma_{16,1}$	-6.63 (0.91)	-6.61 (1.01)	-4.11 (0.79)	-4.25 (0.75)
$\gamma_{16,2}$	-0.21 (0.43)	-0.53 (0.49)	-0.38 (0.41)	-0.1 (0.39)
$\gamma_{17,1}$	-6.36 (1.02)	-6.17 (1.05)	-4.1 (0.87)	-4.62 (0.95)
$\gamma_{17,2}$	-1.31 (0.62)	-0.97 (0.61)	-0.75 (0.49)	-0.52 (0.52)
$\gamma_{18,1}$	-6.99 (1.07)	-5.45 (0.98)	-5.22 (1.01)	-6.02 (1.09)
$\gamma_{18,2}$	-0.25 (0.53)	-1.14 (0.61)	-1.79 (0.66)	-1.68 (0.64)
τ_2	-1.46 (0.5)	-4.47 (0.82)	-2.89 (0.52)	-3.82 (0.61)
τ_3	-3.61 (0.99)	-2.72 (0.79)	-2.62 (0.95)	-4.35 (1.11)
τ_4	-2.55 (0.58)	-3.83 (0.68)	-4.21 (0.9)	-4.69 (0.86)
τ_5	-3.17 (0.64)	-3.67 (0.66)	-3.9 (0.92)	-2.95 (0.6)
τ_6	-3.92 (0.96)	-3.91 (0.75)	-3.3 (0.78)	-4.36 (0.9)
τ_7	-1.34 (0.56)	-3.08 (0.53)	-3.54 (0.77)	-4.42 (0.97)
τ_8	-3.3 (0.65)	-4.2 (0.7)	-3.13 (0.59)	-4.24 (0.71)
τ_9	-2.67 (0.42)	-4.52 (0.62)	-2.64 (0.42)	-5.47 (0.9)
τ_{10}	-2.33 (0.37)	-4.39 (0.59)	-3.26 (0.51)	-4.67 (0.62)
τ_{11}	-3.34 (0.65)	-3.85 (0.68)	-3.5 (0.65)	-3.88 (0.6)
τ_{12}	-3.48 (0.83)	-6.05 (1.23)	-3.1 (0.66)	-4.31 (0.75)
τ_{13}	-1.88 (0.55)	-5.47 (0.94)	-2.47 (0.49)	-4.11 (0.88)
τ_{14}	-2.57 (0.54)	-3.39 (0.57)	-2.47 (0.51)	-4 (0.8)
τ_{15}	-2.91 (0.64)	-4.07 (0.84)	-2.69 (0.71)	-4.18 (1.06)
τ_{16}	-3.33 (0.8)	-3.87 (0.65)	-4.45 (0.97)	-5.11 (0.95)
τ_{17}	-3.03 (0.88)	-3.39 (0.7)	-3.96 (1.08)	-4.31 (0.93)
τ_{18}	-3.88 (0.76)	-4.51 (0.78)	-3.03 (0.68)	-3.16 (0.62)

Note. Standard errors are presented in parentheses.

Appendix F: Profile group subscale covariance matrices

Group 1

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	1.09			
Lexical Grammar	0.56	0.42		
Rhetorical Organization	0.06	0.09	2.92	
Question Handling	0.39	0.45	1.38	5.43

Group 2

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	14.76			
Lexical Grammar	0.97	7.15		
Rhetorical Organization	0.63	1.70	3.02	
Question Handling	1.01	2.43	2.12	4.72

Group 3

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	5.95			
Lexical Grammar	0.02	5.73		
Rhetorical Organization	-0.13	1.83	2.89	
Question Handling	-0.26	1.88	2.23	4.85

Group 4

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	7.26			
Lexical Grammar	1.72	5.80		
Rhetorical Organization	1.95	1.49	2.16	
Question Handling	-0.45	-0.61	1.14	4.84

Group 5

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	17.25			
Lexical Grammar	7.03	10.40		
Rhetorical Organization	-0.09	-0.04	0.06	
Question Handling	-0.37	-0.31	0.06	0.14

Group 6

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	0.42			
Lexical Grammar	0.03	10.03		
Rhetorical Organization	-0.09	1.98	3.15	
Question Handling	0.34	2.32	1.97	6.27

Group 7

	Pronunciation	Lexical Grammar	Rhetorical Organization	Question Handling
Pronunciation	0.11			
Lexical Grammar	0.04	0.03		
Rhetorical Organization	-0.03	-0.01	1.15	
Question Handling	0.08	0.06	0.00	0.13

Appendix G: MTR bundles in the TOP corpus

1	he drugs. So that's importance of the stem cell research. However we do have a lot of challenge *like you said*, transplant to back to the human is a lot of problem. For instance, how can we specifically divide diffe
2	in your family electricity to conduct the conduct electricity. It's such as metal. Ok? okay *let's come back to* my topic, this is the hold end, this is the cold end, with this temperature gradient, what
3	h is the Aether. Okay? And the Aether, well. *I'm gonna have to wait* a little bit. Uh a little bit later.
4	uddenly change into solving a linear equation and quadratic equation, well if the time allows, *I'll move on to* quadratic equation but I'll start with the linear equation. Well, first to solve the linear e
5	to the computer and the center. And I'll be talking about that in the date up on final presentation. so. so *as I said* like uh the speed the speed of the storage devices and the how fast the data from them is able to acce
6	of population flows, then that result in the population population. So that's that that *I'm goin to come up* with. So this your original population. Stock right? An then its your population stock experienc
7	all living and extinct organisms are related somehow. But he also suggested that yep? Phylogenetics, as I *as I said* is a field that is studying the historical relationships between organisms on earth. So Charles Dar
8	e this time dimension, we can state that let me *let me go to* that question in a minute. So now that we are seen the relationships, we can we can see different patterns of relationships. A This is important to understand t
9	s a very interesting question. So if we have this time dimension, we can state that let me let me go to that question in a minute. So *now that we are seen* the relationships, we can we can see differ
10	ay? Okay. Um let's talk about a quest uh talk about a story about the standard deviation, *before we get into* the uh the computation. The standard deviation is first come up by Carl Friedrich Gauss. Whi w
11	information, come from correction, is called negative evidence. So it's another important notion, so and *as I* if *just told you*, children aware in correctives. So mm and moreover, Uh negative evident is not providing
12	original object. yes. Mm hmm. In fact, uh p is always positive. but q may be negative. we *will come to this* soon. because step by step. so uh on that hand if the absolute value of m is less than o
13	hat the image of the candle will be like this. okay? Any question? So far so good? okay. *Let's move on*. and the distance from the object and the lens is denoted by, p. and the distance from the
14	I think you will get this equation, and we'll talk about this next time. okay so, so *can I just move on?* no question? Ok. let me just erase this a little. then I'm going to talk about the magnific
15	e story I'm telling you. you *will know the answer later*. okay. The Chinese version of the

	six syllables
16	o. So if prisoner one confess, I mean and can't don't confess, and same for prisoner two. okay. *We said* if nobody confess to both going to jail for being in possession for Mona Lisa, which is two years each.
17	ike this. Mhm? And a oh sorry. This is a this is the reason why I mention the mathematical speak. Because uh *as I mentioned before*, so transfer function is I said transfer function's really important to get the relationsh
18	cation because they don't need to. And I'm gonna go over it uh next. So I ha and *I was talking here about* link layer. So the network layer is next step, uh on the network layer it works um you need the net
19	whatever it does. So salleh move on? Um I went to hear or you want something? Mmhmm, okay. So now this j *as I said* that there are certain imperfections in this device, and that is like uh by changing the voltage ou
20	elp of that you just characterize the how much energy is your device con consuming. For example *I go back again* to the mobile phone, in that you have a battery, and it uh has a battery life. So if you
21	eir hands. From here to there. So this is how we define conduction. Any questions? Okay. *Next, we move on to* convection. This is another kind of mode of heat transfer. Now let us assume that we have the
22	m, which This type of transfer of heat is called as convection. Any questions? Okay. Then, *let us move to* the last part of uh mode of heat transfer, radiation. This uh to find an analogy for radiation w
23	esired to be transferred, this kind of thing is known as radiation. Now the kind uh and it were uh *I was sayi ng that* it is little bit tricky is because here a medium is involved, but whereas in actual reality if you
24	would be a little tricky, which I *will comment on later* on, but let us suppose that we had the same class
25	that's the plan C. Yeah. Yeah. Yeah. Thas that's that's that's why we have this reverse engineering. So uh *as I told you* if we need to optimize some functions. Because some some model, they don't fulfill or they can't f
26	ther use the mean or the median. However, when uh when the distribution of your data is not symmetric, then, *as I mentioned* before, the mean is very sensitive to uh each measurement, so it it is extremely sensitive to the
27	e they are connected. Yeah. Uh okay, what do you imagine is the principle we will *we will uh talk about* the next part. Uh so the power adapter of your laptop is one, example of our daily life. And ther
28	m sorry? The the energy can only be transformed euh as uh a as heat and uh work. Okay. *Let's go on*. Uh
29	onomic growth. Thus bring the changes in its higher education. So *that's my next thing* I wanna talk about just because of economic growth China's higher education started to change. So what's it like befo
30	k in class today, I wanna talk about the findings from this case study. Okay. So *next thing is* talk about findings. I analyzed the findings into two categories, the first one, we call it globally informe
31	you may feel any difficulty. You may not you may not feel any difficulty. From this class.

	So *let's get started*. I will dis I will discuss the definition, and benefit of using decibel. And if I have
32	garithm of the ones thousand, yes? Oh. It's different, I will I will discuss in later. Sweet. *So go to the* decibel. Uh decibel means, ten times log X. So, in this case, the decibel is ten, multip
33	h as f uh measure it's a lowest scale of the measured values such as voltage or current. *So next uh go to the next*. Pardon? Sound? So what is problem? Yes, of course. But mm to measure sound, uh
34	It's three. Because there are three zeroes here. So, this is the exponent number, and we we *should move on to* logarithm. So logarithm is the exponent number exponent of some number, so logarithm of the
35	es? Oh. It's different, I will *I will discuss in later*. Sweet. So go to the decibel. Uh decibel means,
36	are not familiar with, that is female patron of art in China. But first of all, uh *before I get to gett ing to* the topic, please name, uh please name one female ruler in the history that that you guys kno
37	re stable. Is is what the theory says. Um Yeah. Go ahead. I um I'm basically done. yeah. Ah No. Well, it *as I said* you know you could fill this whole uh room or I mean this whole floor with with people with literature
38	s be seeking to maximize their economic and military power, if need be at expense of other. So, and *going back to the* to the prison debate. Well there's need to you are going to form alliances with the st
39	nto lactate when there when there's no oxygen. It's an anaerobic um uh reaction. Lactate. So the Cori Cycle, *as we said*, the lactic acid is not good and we wanna get rid of it. The Cori Cycle is is gonna help you do that.
40	u're gonna run the Cori Cycle, you wanna run it when you are you have oxygen again. To be able to um or else, *as I said* I showed you over here if you keep on you don you can't keep running. It's you're using too much energ
41	So um *as I already said* this is lecture about ontology. Um I suppose uh although you're lots of undergrad that you cou
42	hat as a new statement, isn't it. something different. So we have to acknowledge a third realm. and that is *as I have said* um the realm of the abstract objects. Um now if these premises are true from as you know your i
43	nna do it through thoughts. Thoughts aren't are I'm sorry things in the outer world. *I've already told you* what you should have in lined. what do you think about things in the outer world. so this thing
44	If the case of three four five, then A is a friend of three out of five. Right? silent, now *we can go to* a first uh situation. A is a friend of three out of five. Then we can denote the three people,
45	n Buddhist studies. Eight kinds of languages. One. English. Two, French, three, German, and four, Japanese. *As I told you*, Buddhist studies was all where. Yes, in Europe. So there are many scholars whose native language
46	It's a very difficult question so *I will ask you later* This is India, what is the name of the island.
47	India, and now they are trying to move this group. And North Korea tested in two thousand

	six. But you know, *as I mentioned* the international uh organization and great power does not recognize these guys as uh nuclear po
48	Yeah right. Iran, not the North Korea*, I will go back to that*. South Korea, Not India. I will get back to that. Not Pakistan, Because there's some very you know Certain difference between those g
49	another group, latent nuke, Yeah. That's N, Yeah right. Iran, not the North Korea, *I will go back to that*. South Korea, Not India. I will get back to that. Not Pakistan, Because there's s
50	is charge, so the charge is changed by time. So uh in the in in norm case, it should be zero. *Because I said* uh one charge get in, and one charge get out. So it should be zero. It's uh it's it is called uh the co
51	se ones. Yes. Um so. Basically this how transformer works. Okay? Okay. So any questions? Transformer? Yeah. *As I said*, well, you can they are hidden in a for example a stage microphone, it's it can be as sm small as a th
52	w I'm going to just explain for you, what applications we can have with just these kind of straight beam. uh *as I explain*, do we can just have a verfukus point of light, so you can burn something. Let's say you can do e
53	onment. yeah. Oh ok. Ok. This section is actually very nice question. ah *I was just thinking to explain* it for you or not but right now I'll I'm going to explain. ah let's say I'm not good in painting
54	a lot of different terminology from linguistics, and that a little confusing. Especially *we look back to* IPA, because lot of symbols are used differently in IPA.
55	r? Okay. Yeah. So basically that's why we call it micro electrical mechanical system. And we just *I just said that* that two application is the sensor and actuator. And right now I give you the an example about sens
56	l *give you some* example about sensor and actuator *later*, and yeah. Let me talk about the the first word,
57	questions. Um, I mean, so just yeah. Yeah Yeah. The things Okay, let me give Yeah. *I was going to go into* that part. The idea is suppose you are
58	s going to to uh do what I just did, I'm sorry, I I I I ran ahead of myself, was which which *was uh telling you* what different uh like what we are going to do in this course, which is like read these papers,
59	st of all let's suppose that she um she has she wants to consume the following bundle of good. When *I said* when I say bundle of goods, is just a a number actually, a vector, but I'm not going to enter in the d
60	o we need it in the first place, you know? Um you know, so, you know, one thing you were saying, *to go back to the* why question, is you know we were talking about um I guess we were talking about order
61	y ask like why do we need it in the first place, you know? Um you know, so, you know, *one thing you were saying*, to go back to the why question, is you know we were talking about um I guess we were talking about o
62	aying, to go back to the why question, is you know *we were talking about* um I guess we were talking about order and stability, but you know also um ethics. The question of ethics comes in. Like um how do w
63	nformation about which pages have this word. The reason why you have this index, is to

	speed up the search *as you said before*. If we don't have this index, we need to look at all of these downloaded pages, And sear
64	Uh uh yeah that that that that's the definition of displacement. Any questions? Alright. Uh *let's move on to* velocity. Um the definition of velocity uh is actually based on displacement. It's it's the
65	how can we apply this idea to civil engineering design. For example, um if we wanna construct a building. So *as I mentioned before*, um there are a couple requirements to be satisfied towards sustainable design, so I'm gon
66	tell you what other requirements to be satisfied to be a sustainable design, so um alright um *let's get a start* on the lecture. So before um I give you the definition of sustainability, um I would like t
67	Okay let's uh let's go forward to uh that section. Um *as I said before* as I said before, I have to introduce a basic idea about solar cells to you. Uh generally, a solar cell is just
68	? Uh once you form this kind of s uh basic structure, you shine light very close to this junction. Junction, *as I said*, I just a uh just an intersection, just an interface between two semiconductor two semiconductors, the
69	ion over here. So the junction just stand from the interface between these two type of semiconductor. Uh *you mentioned* what does P stand for. P stand for the majority carriers in s these type of semiconductors are holes. And th
70	Okay let's uh let's go forward to uh that section. Um *as I said before* as I said before, I have to introduce a basic idea
71	lectron hole pairs will be generated nearby. Okay, uh if you want to know more in detail, I *have to go uh go a little bit digressive* to uh stru uh band structure of semiconductor. Uh generally in any an
72	creased twice. So which means same position, but increasing it twice, so this one is like this. Resistance. *As I said*, resistance is the slope of this figure. Do you think what, slope increase or decrease. Increase. Wh
73	antenna very small, it can support this. So. This formula governs the whole it it it isn't . *Now coming up to the* second reason, the the media transmissions being wireless and wireline. Now only channel
74	T to R. Uh this is a real world analogy of what modulation's about. So uh yeah. *Now* coming to the *moving to the* second part, which is why do we need modulation. Here, uh a a number of factors govern as
75	ight Yeah. If you said it's um typically an image of several distorted letters. Like Yeah I *will explain later*. It's a good question. um. Now, it's your job to uh type the correct series of the letters
76	m more thing. Because yeah. Yeah. It has uh more uh calculations. probly it makes um. I I *will explain later*. Yeah Okay. Um. Uh. Now let's see how it works. How the reCAPTCHA work. First, um the admini
77	ittle lower, and yo the voltage you'll apply down it will be a smaller. And uh if we apply um *let's go on* let's go on, if we apply a negative sign of voltage, maybe the the barrier will further increase
78	camera, cell phone, and uh laptop. So I think it is good to for you to know how it works. So *let's get started*. So I will divide this topic into two parts, the part one will be the material part, so w

79	so good? Okay, I think we have uh enough kn knowledge to fabricate the device now. So So so let's move on the device part. Yeah. So this is the basic device's structure of the of a MOSFET. So the gate
80	is? Uh very close, yeah. So No. Wood Wood is an insulator. Yeah. So? Before we *before talking about* a device part, I think we we need to know what the semiconductor is, because the MOSFET is made of
81	machine. Um just. So here're two important concepts, so before explain this, um so *I already explain this*. So performance is how fast it is, I will give you some quick example. Of algorithms, whi
82	actually um actually *I can answer that question later*. Explain decipher ee stuff. Um so okay so here'

Appendix H: Rhetorical Organization scoring rubric

	Description
4	Excellent overall organization and use of transitions between sentences and topics; effective use of rhetorical questions. Successful macro and micro rhetorical organization. Clearly organized discourse positively contributes to communication.
3	Good overall organization and use of transitions between ideas/sentences. Discourse is appropriately organized and structured. Ideas are logically connected to one another with appropriate cohesive devices. Organization does not significantly impede communication.
2	Minimal overall organization and/or incorrect use of transitions between ideas/sentences. Discourse not well organized and difficulty articulating main topics and/or subtopics. Errors in use of cohesive devices and organization of ideas somewhat impede communication.
1	No overall organization and/or ineffective use of transitions between ideas/sentences. Discourse is generally not organized or structured. Errors in use of cohesive devices and lack of organization of ideas severely impede communication.

REFERENCES

Aguilar, M. (2004). The peer seminar, a spoken research process genre. *Journal of English for Academic Purposes, 3*, 55-72.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Althen, G. (1991). Teaching culture to international teaching assistants. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach: Selected readings in TA training* (pp. 350-355). Dubuque, Iowa: Kendall/Hunt Publishing Company.

Anderson-Hsieh, J. (1990). Teaching suprasegmentals to international teaching assistants. *English for Specific Purposes, 9*(3), 195-214.

Anthony, L. (2011). *AntConc (Version 3.2.2)* [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Ard, J. (1987). The foreign TA problem from an acquisition-theoretic point of view. *English for Specific Purposes, 6*(2), 133-144.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bailey, K. M. (1982). *Teaching in a second language: The communicative competence of non-native speaking teaching assistants*. Unpublished doctoral dissertation. University of California, Los Angeles.

Bailey, K. M. (1984). The "foreign TA problem." In K. Bailey, F. Pialorsi, & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 3-15). Washington, D.C.: National Association for Foreign Student Affairs. p. 3-15.

Bailey, K. M. (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. C. Hauptman, R. Le Blanc, & M. B. Wesche (Eds.). *Second language performance testing* (pp. 153-180). Ottawa: University of Ottawa Press.

Bailey, K. M., Pialorsi, F., & Zukowski/Faust, J. (Eds.). (1984). *Foreign teaching assistants in U.S. universities*. Washington, D.C.: NAFSA.

Baldwin, P., Bernstein, J., & Wainer, H. (2009). Hip psychometrics. *Statistics in Medicine, 28*, 2277–2292.

Bauer, G. (1991). Instructional communication concerns of international teaching assistants: A qualitative analysis. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach: Selected readings in TA training* (pp. 420-426). Dubuque, Iowa: Kendall/Hunt Publishing Company.

Bauer, G. (1992). *Instructional communication concerns of international teaching assistants*. Unpublished doctoral dissertation, Pennsylvania State University, University Park, PA.

Bauer, G., & Tanner, M. (Eds.). (1994). *Current approaches to international TA preparation in higher education: A collection of program descriptions*. Seattle, Washington: Center for Instructional Development and Research, University of Washington.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, 21(2), 175-189.

Biber, D. (1986). On the investigation of spoken/written differences. *Studia Linguistica, 40*(1), 1-21.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written register*. Amsterdam: John Benjamins.

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9-48.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Bier, R., & Friedman, C. (1982, April). Certifying the English proficiency of foreign teaching assistants: The problem and the process. Paper presented at the Second Annual Midwest Regional TESOL Convention, Indianapolis, IN.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.

Black. P., & William. D. (1998) Assessment and classroom learning. *Assessment in Education,* 5, 7–74.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*(4), 364-375.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133-159). Mahwah, NJ: Lawrence Erlbaum Associates.

Bowles, H. (2006). Bridging the gap between conversation analysis and ESP – an applied study of the opening sequences of NS and NNS service telephone calls. *English for Specific Purposes, 25*, 332-357.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201-236). New York: Academic.

Bresnahan, M., & Kim, M. (1993). Predictors of receptivity and resistance toward international teaching assistants. *Journal of Asian Pacific Communication, 4*(1), 3-14.

Briggs, S. (1994). Using performance assessment methods to screen ITAs. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 63-80). Bloomington, IL: TESOL, Inc..

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrd, P., & Constantinides, J. (1992). The language of teaching mathematics: Implications for training ITAs. *TESOL Quarterly, 26*(1), 163-167.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612.

Cai, L. (2012). *flexMIRT™ version 1.86: A numerical engine for multilevel item factor analysis and test scoring*. [Computer software]. Seattle, WA: Vector Psychometric Group.

Cai, L., & Hansen, M. (2012). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*(2), 245–276.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221-248.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Chaudron, C., & Richards, J. C. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics, 7*(2), 113-127.

Chiang, S-Y. (2009). Dealing with communication problems in the instructional interactions between international teaching assistants and American college students. *Language and Education, 23*(5), 461-478.

Cho, S-J., Cohen, A. S., & Kim, S. H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*(2), 278-306.

Constantinides, J. C. (1987). The foreign TA problem – an update. *NASFSA Newsletter, 38*(5), 3-6.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 833-904.

Csomay, E. (2007). A corpus-based look at linguistic variation in classroom interaction: Teacher talk versus student talk in American University classes. *Journal of English for Academic Purposes, 6*, 336–355.

Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with cluster via model-based clustering. *Journal of the American Statistical Association, 93*, 294-302.

Davies, C., & Tyler, A. (1994). Demystifying cross-cultural (mis)communication: Improving performance through balanced feedback in a situated context. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 201-220). Bloomington, IL: TESOL, Inc..

Davies, C., Tyler, A., & Koran, J. (1989). Face-to-face with English speakers: An advanced training class for international teaching assistants. *English for Specific Purposes, 8*, 139-153.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333-356.

DeCarrico, J. & Nattinger, J. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes, 7*(2), 91–102.

De Jong, M. G., Steenkamp, J. B. E. M., & Fox, J. P. (2007). Relaxing crossnational measurement invariance using a hierarchical IRT model. *Journal of Consumer Research, 34*, 260-278.

De Jong, M. G., Steenkamp, J. B. E. M., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45*, 104-115.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society B, 39*, 1-38.

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. In J. Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 347–369). Philadelphia, PA: John Benjamins.

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, Psychometrics) (pp. 979–1027). Amsterdam: Elsevier.

Dick, R., & Robinson, B. (1994). Oral proficiency requirements for ITAs in U.S. colleges and universities: An issue in speech communication. *Journal for Association for Communication Administration, 2*, 77-86.

Douglas, D., & Selinker, L. (1989). Markedness in disource domain: Native and non-native teaching assistants. *Papers in Applied Linguistics, 13* (1), 69-81.

Douglas, D., & Selinker, L. (1994). Native and nonnative teaching assistants: A case study of discourse domains and genres. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 221-230). Bloomington, IL: TESOL, Inc..

Doughty, C. J., & Long, M. H. (Eds.). (2003). *The handbook of second language acquisition*. Malden, MA: Blackwell.

Douglas, D., & Myers, C. (1989). TAs on TV: Demonstrating communication strategies for international teaching assistants. *English for Specific Purposes, 8*, 169-179.

Duerksen, A. (1994). *A descriptive study of the discourse skills of non-native speaker teaching assistants*. Unpublished doctoral dissertation. Ball State University.

Dulay, H., & Burt, M. (1973). Should we teach children syntax? *Language Learning, 23*. 235–252.

Dulay, H., & Burt, M. (1974) Natural sequences in child second language acquisition. *Language Learning, 24*, 37–53.

Dulay, H., & Burt, M. (1975) Creative construction in second language learning and teaching. In M. Burt and H. Dulay (Eds.), *On TESOL '75: New directions in second language learning, teaching and bilingual education* (pp. 21–32). Washington, D.C.: Teachers of English to Speakers of Other Languages.

Dunn, T., & Constantinides, J. (1991). Standardized test scores and placement of international teaching assistants. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the*

professoriate of tomorrow to teach: Selected readings in TA training (pp. 414-419). Dubuque, Iowa: Kendall/Hunt Publishing Company.

Duskova, L. (1969). On sources of errors in foreign language learning. *International Review of Applied Linguistics in Language Teaching, 7*, 11–36.

Edwards, J. A., & Lampert, M. D. (Eds.). (1993). *Talking data: Transcription and coding in discourse research*. Hillsdale: Erlbaum.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). New York: Wiley.

Farr, F. (2003). Engaged listenership in spoken academic discourse: the case of student–tutor meetings. *Journal of English for Academic Purposes, 2*, 67-85.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326.

Feetham, E. (1988, April). *The faculty's role in the training of international teaching assistants: Part of the problem or part of the solution?* Paper presented at the Symposium on the Training of International Teaching Assistants, University of Pennsylvania, PA.

Flowerdew, J., & Tauroza, S. (1995). The effect of discourse markers on second language lecture comprehension. *Studies in Second Language Acquisition, 17*, 435-458.

Fortanet, I. (2004). The use of 'we' in university lectures: Reference and function. *English for Specific Purposes*, 23, 45-66.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354–375.

Fox, J-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Fox, W., & Geneva, G. (1994). Functions and effects of international teaching assistants. *The Review of Higher Education, 18*(1), 1-24.

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). *mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation.* Technical Report No. 597, Department of Statistics, University of Washington.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling, or wedding bells? *TESL-EJ, 8*(4), 1-37.

Gallego, J. (1990). The intelligibility of three nonnative English-speaking teaching assistants: An analysis of student-reported communication breakdowns. *Issues in Applied Linguistics, 1*(2), 219-237.

Gass, S. & Selinker, L. (Eds.). (1983). *Language transfer in language learning*. Rowley, Massachusetts: Newbury House Publishers.

Gass, S., & Selinker, L. (2001). *Second language acquisition: An introductory course* (2nd Ed.). Mahwah, NJ.: Lawrence Erlbaum Associates.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (pp. 131-143). London: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd Ed.). New York: Chapman & Hall.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511.

Gelman, A., & Shirely, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. L. Jones, & X, Meng (Eds.), *Handbook of Markov chain monte carlo* (pp. 163-174). London: Chapman & Hall.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid,& A. F. M. Smith (Eds.) ,*Bayesian Statistics 4* (pp. 169-193). Oxford: Oxford University Press.

Gibbons, J. (1984). Interpreting the English proficiency profile in Hong Kong. *RELC Journal, 15*, 64-74.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss. D. J., Segawa, E., Bhaumik, D. K., Kupfer, E. F., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4–19.

Gibbons R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57*(3), 423–436.

Gill, J. (2007). *Bayesian methods: A social and behavioral sciences approach* (2nd Ed.). New York: Chapman & Hall.

Granger, S. (2003a). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538-546.

Granger, S. (2003b). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465-480.

Green, C. F., Christopher, E. R., & Lam, J. (1997). Developing discussion skills in the ESL classroom. *ELT Journal, 51*(2), 135-143.

Halleck, G. & Moder, C. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly, 29*(4), 733-758.

Halliday, M. A. K. (1975). *Learning how to mean*. London: Edward Arnorld.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Han, Z. (1998). *Fossilization: An investigation into advanced L2 learning of a typologically distant language*. Unpublished doctoral dissertation, Birkbeck College, University of London.

Han, Z. (2004). *Fossilization in adult second language acquisition*. Tonawada, New York: Multilingual Matters.

Han, Z., & Odlin, T. (Eds.). (2006). *Studies of fossilization in second language acquisition*. Tonawada, New York: Multilingual Matters.

Hendel, D., Dunham, R., Smith, J., Solberg, J., Tzenis, C., Carrier, C., & Smith, K. (1993). Implications of student evaluations of teaching for ITA development. In L. K. Stillwater (Ed.), *The TA experience: Preparing for multiple roles* (pp. 390-400). OK: New Forums Press, Inc..

Hill, J. L., & Kriesi, H. (2001). Classification by opinion-changing behavior: A mixture model approach. *Political Analysis, 9*(4). 301-324.

Hinofotis, F., & Bailey, K. (1981). American undergraduates' reactions to the communication skills of foreign teaching assistants. In J. C. Fisher & J. S. Schacter (Eds.), *On TESOL '80, building bridges: Research and practice in teaching English as a second language* (pp. 120-136). Washington D.C.: TESOL, Inc..

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26*(2), 243-269.

Hudson, T. (1993). Nothing does not equal zero: Problems with applying developmental sequence findings to assessment and pedagogy. *Studies in Second Language Acquisition, 15*, 461-493.

Hyland, K. (2002). Genre: Language, context, and literacy. *Annual Review of Applied Linguistics, 22*, 113-135.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 24–49.

Jacobs, L., & Friedman, C. (1988). Student achievement under foreign teaching associates compared with native teaching associates. *Journal of Higher Education, 59*(5), 521-563.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213-241.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31-73.

Jenkins, S. (1997). *Cultural and pragmatic miscues: A case study of international teaching assistant and academic faculty miscommunication*. ERIC: Document No. ED411684, 35 p.

Jia, C., & Bergerson, A. (2008). Understanding the international teaching assistant training program: A case study at a northwestern research university. *International Education, 37* (2), 77-98.

Johncock, P. (1991). International teaching assistants: Tests and testing policies at U.S. universities. *College and University, 66*(3), 129-137.

Johnson, V. E. (1996), On Bayesian analysis of multirater ordinal data. *Journal of the American Statistical Association, 91*, 42-51.

Ju, M. K. (2000). Overpassivization errors by second language learners. *Studies in Second Language Acquisition, 22*, 85-111.

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

Kagan, O. (2005). In support of a proficiency-based definition of heritage language learners: The case of Russian. *International Journal of Bilingual Education and Bilingualism, 8*, 213-221.

Kaplan, R. (1989). The life and times of ITA programs. *English for Specific Purposes, 8*(2), 109-124.

Kim, A. (2011). *Cognitive diagnosis of second language reading ability based on a language ability framework.* Unpublished doctoral dissertation. Teachers College, Columbia University.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing, 19(1), 3-31.

Lee, Y., & Sawaki, Y. (Eds.) (2009). Cognitive diagnosis and Q-matrices in language assessment. *Language Assessment Quarterly, 6*(3), 169-263.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*(4), 675-724.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the national corpus*. London: Longman.

Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge, UK: Cambridge University Press.

Lesaux, N. K., & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal, 47*(3), 596-632.

Liao, S. (2009). Variation in the use of discourse markers by Chinese teaching assistants in the US. *Journal of Pragmatics, 41*, 1313-1328.

Linacre, J. M. (1989). *Many-faceted Rasch Measurement*. Chicago, IL: MESA Press

Locastro, V. (2001). Large classes and student learning. *TESOL Quarterly, 35*(3), 493-496.

Luoma, S. (2004). *Assessing speaking.* Cambridge, UK: Cambridge University Press.

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

Madden, C. & Myers, C. (Eds.) (1994). *Discourse and performance of international teaching assistants*. Bloomington, IL: TESOL, Inc..

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics, 32*(3), 287-314.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009–1020.

McChesney, J. (1994). The functional language of the U.S. TA during office hours. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 134-152). Bloomington, IL: TESOL, Inc..

McEnery, A., Xiao, R. Z., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Oxford: Routledge.

McLachlan, G.J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley-Interscience.

Mestenhauser, J. (1981). Foreign students as teachers: Lesson from the program in learning with foreign students. In G. Althen (Ed.), *Learning across cultures* (pp. 143-149). Washington, D.C.: NAFSA.

Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-33), Mahwa, NJ: Lawrence Erlbaum Associates.

Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*, 1050–1066.

Muthén, L., & Muthén, B. (2011). *Mplus user's guide, Version 6*. Los Angeles: Muthén & Muthén.

Myers, C. (1994). Question-based discourse in science labs: Issues for ITAs. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp.83-102). Bloomington, IL: TESOL, Inc..

Nattinger, J. & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nelson, G. (1990, March). *International teaching assistants: A review of research*. Paper presented at the TESOL Convention, San Francisco, CA.

Nelson, G. (1991). Effective teaching behavior for international teaching assistants. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach: Selected readings in TA training* (pp. 427-434). Dubuque, Iowa: Kendall/Hunt Publishing Company.

Nesi, H., & Basturkmen, H. (2009). Lexical bundles and discourse signalling in academic lectures. In J. Flowerdew & M. Mahlberg (Eds.), *Lexical cohesion and corpus linguistics* (pp. 23-43). Amsterdam: John Benjamins.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching?* (pp. 125-152). Amsterdam: John Benjamins.

Nyquist, J., Abbott, R., Wulff, D., & Sprague, J. (Eds.) (1991). *Preparing the professoriate of tomorrow to teach: Selected readings in TA training*. Dubuque, Iowa: Kendall/Hunt Publishing Company.

Oyama, S. C. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research, 5*(3), 261–83.

Patkowski, M. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning, 30*, 449–72.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Patz, R., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341-384.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A, 185*, 71–110.

Pienemann, M. (1998). *Language processing and second language development*. Amsterdam: John Benjamins.

Plakans, B., & Abraham, R. (1990). The testing and evaluation of international teaching assistants. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities*, (pp. 68-81). Washington, D.C.: NAFSA.

Plough, I. C., Briggs, S. L., & Bonn, S. V. (2011). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*, 7-11.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. New York: Springer.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal, 26*, 81-114.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs Sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 763-773). Oxford: Oxford University Press.

Rea-Dickins, P. (2007). Classroom-based assessment: Possibilities and pitfalls. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 505-520). New York: Springer.

Reinhardt, J. (2010). Directives in office hour consultations: A corpus-informed investigation of learner and expert usage. *English for Specific Purposes, 29*, 94-107.

Reppen, R. (2004). Academic language: an exploration of university classroom and textbook language. In U. Connor & T. Upton (Eds.), *Discourse in the professions* (pp. 65-86). Philadelphia, PA: John Benjamins.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

Robertson, D. (1983). *English language use, needs, and proficiency among foreign students at the UIUC.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Robinson, A. (1993). Responding to student questions: An analysis of teaching assistant discourse. In L. K. Stillwater (Ed.), *The TA experience: Preparing for multiple roles* (pp. 86-94). OK: New Forums Press, Inc..

Roeder, K., & Wasserman, L. (1997). Practical density estimation using mixtures of normals. *Journal of the American Statistical Association, 92*, 894-902.

Rosenthal J. S. (1993). Rates of convergence for data augmentation on finite sample spaces. *Annals of Applied Probability, 3*, 819-839.

Rosenthal J. S. (1995a). Minorization conditions and convergence rates for Markov chain monte carlo. *Journal of the American Statistical Association, 90*, 558-566.

Rosenthal J. S. (1995b). Rates of Convergence for Gibbs sampling for variance component models. *Annals of Statistics, 23*, 740-761.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282.

Rost, J. (1997). Logistic mixture models. In W.J. van der Linden & R.K. Hambleton, (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In I. W. Moelenaar (Ed.), *Rasch models: Foundations, recent developments and applications* (pp. 257-268). New York, NY: Springer Verlag.

Rounds, P. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly, 21*(4), 643-671.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. (1993). The other half of international teaching assistant training: Classroom communication workshops for international students. *Innovative Higher Education, 17*(3), 183-193.

Rubin, D., & Smith, K. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations, 14*, 337-353.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language, 50*, 696–735.

Sadler, R. (1989). Formative assessment and the design of instructional assessments. *Instructional Science, 18*, 119-144.

Scmidgall, J. E. (2012a, April). *Using an ITA assessment to provide detailed feedback on performance: Implications for learners, teachers, and validity*. Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Scmidgall, J. E. (2012b). *Evaluating the consistency of scores for a test of oral English within the framework of an argument for test use*. Unpublished qualifying paper, University of Los Angeles, California.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Berlin: Springer-Verlag.

Schwartz, B. (1997). The second language instinct. Plenary speech at GALA '97, Edinburgh.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Scott, M. (2013). *WordSmith Tools Help.* Liverpool: Lexical Analysis Software.

Seo, W. (1989). *A profile of communication skills of foreign teaching assistants in a major Midwestern university.* Unpublished doctoral dissertation, University of Cincinnati.

Sequeira, D., & Costantino, M. (1989). Issues in ITA training programs. In J. Nyquist, R. Abbott, & D. Wulff (Eds.), *Teaching assistant training in the 1990s (pp. 79-86).* San Francisco: Jossey-Bass, Inc.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*(4), 513-521.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512.

Smith, R., Byrd, P., Nelson, G., Barret, R., & Constantinides, J. (1992). *Crossing pedagogical oceans: International teaching assistants in U.S. undergraduate education* (ASHE-ERIC Higher Education Report No. 8). Washington D.C.: The George Wasington University, School of Education and Human Development.

Solka, J. L., Wegman, E. J., Priebe, C. E., Poston, W. L., & Rogers, W. (1998). Mixture structure analysis using the Akaike criterion and the bootstrap. *Statistics and Computing, 8*, 177-188.

Spiegelhalter, D. J., Best, N, G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B, 64*(4), 583–639.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., & Inskip, H. (1996). Hepatitis B: A case study in MCMC methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (pp. 21-43). London: Chapman & Hall.

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell, Oxford.

Tanner, M., Selfe, S., & Wiegand, D. (1993). The balanced equation to training chemistry ITAs. Innovative Higher Education, 17(3), 165-181.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Thomas, C., & Monoson, P. (1991). Issues related to state-mandated English language proficiency requirements. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach: Selected readings in TA training* (pp. 382-403). Dubuque, Iowa: Kendall/Hunt Publishing Company.

Thomas, C., & Monoson, P. (1993). Oral English language proficiency of ITAs: Policy, implementation, and contributing factors. *Innovative Higher Education, 17*(3), 195-209.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*(4), 713-729.

Vardi, I. (2009). The relationship between feedback and change in tertiary student writing in the disciplines. *International Journal of Teaching and Learning in Higher Education, 20*(3), 350-361.

vom Saal, D., Miles, R. J., & McGraw, R. L. (1988). A university-wide assessment and training program for international teaching assistants. *Journal of Agronomic Education, 17*(2), 68-72.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Warriner, D. S. (2007). ″It's just the nature of the beast": Re-imagining the literacies of schooling in adult ESL education. *Linguistics and Education, 18*, 305–324.

Wedel, M., & DeSarbo, W. S. (2002). Market segment derivation and profiling via a finite mixture model framework. *Marketing Letters, 13*(1), 17– 25.

Weissberg, B. (2000). Developmental relationships in the acquisition of English syntax: writing vs. speech. *Learning and Instruction, 10*, 37-53.

Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*(4), 693-697.

Williams, J. (1994, March). *Discourse analysis: Why do we need it?* Paper presented at the Annual TESOL Convention, Baltimore, MD.

Williams, J., Barnes, G., Finger, G., & Ruffin, P. (1987, April). *Training FTAs: Report of a needs analysis*. Paper presented at the TESOL Convention, Miami Beach, FL.

Yeni-Komshian, G. H., Flege, J. E., & Liu, S. (2000). Pronunciation proficiency in the first and second languages of Korean–English bilinguals. *Bilingualism: Language and Cognition, 3*(2), 131-149.

Young, R. (1989, May). *Curriculum renewal in training programs for international teaching assistants.* Paper presented at NAFSA conference, Minneapolis, MN.

Yung, Y-F. (1997). Finite mixtures in confirmatory factor-analysis model. *Psychometrika, 62*(3), 297-330.

Zobl, H., & Liceras, J. (1994). Review article: Functional categories and acquisition orders. *Language Learning, 44* (1), 159–180.