UNIVERSITY OF CALIFORNIA

Los Angeles

# Justifying the Use of a Second Language Oral Test
# as an Exit Test in Hong Kong:
# An Application of Assessment Use Argument Framework

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Applied Linguistics

by

Yujie Jia

2013

# ABSTRACT OF THE DISSERTATION

## Justifying the Use of a Second Language Oral Test
## as an Exit Test in Hong Kong:
## An Application of Assessment Use Argument Framework

by

Yujie Jia

Doctor of Philosophy in Applied Linguistics

University of California, Los Angeles, 2013

Professor Lyle F. Bachman, Chair

This study employed Bachman and Palmer's (2010) Assessment Use Argument framework to investigate to what extent the use of a second language oral test as an exit test in a Hong Kong university can be justified. It also aimed to help test developers of this oral test identify the most critical areas in the current test design that might need improvement. Candidates' oral responses to five integrated speaking tasks in this oral test were rated on five dimensions: Task fulfillment and relevance (TFR), Clarity of presentation (CoP), Grammar and Vocabulary (GV), Pronunciation (Pron), and Confidence and Fluency (CoFlu).

To provide backing for the meaningfulness of interpretations, confirmatory factor analysis (CFA) and item response theory (IRT) analyses were used to analyze 999 candidates' scores and raters' verbal reports were also analyzed to provide complementary information to the results of the quantitative analyses. Several CFA models were first tested and compared in terms of their statistical fit and substantive interpretability. And a graded response model was applied to the test data. The CFA results showed that the superior fit of the Higher-order trait-Uncorrelated Method model validated the test design, confirmed the current multicomponential view of language ability in the literature, and provided the most parsimonious explanation of the relationships among the five dimensions and overall speaking proficiency. The analytic scores were found to have much larger factor loadings on the trait factors than on the method factors, providing evidence that the component test scores could be meaningfully interpreted as indicators of the five dimensions. The presence of a higher-order speaking ability factor governing the five trait factors also supported the practice of reporting one composite score. Task Fulfillment and Relevance (TFR) measured on Task 4 had the highest method loading (.60) on Task 4 and the lowest trait factor loading (.36) on TFR, which suggested TFR4 might be too task specific and weak in measuring students' speaking ability to fulfill a speaking task in a relevant way. The trace lines of the graded response model also confirmed this. The raters' verbal reports showed that most raters did not have much difficulty differentiating across the performance levels. Hence, the problem of TFR4 can only be due to the nature of the task itself and its low discrimination. Both CFA and IRT results indicated that task types had great effects on test takers' speaking abilities especially TFR and that this language ability component might be too task specific.

In order to investigate the impartiality of interpretations, multi-group CFA and differential item functioning (DIF) were conducted to examine the extent to which the oral test had test bias and item bias across (1) gender and (2) disciplines. The multi-sample CFA results indicated that the factor structure was significantly different between males and females. However, the comparison of the factor loadings between females and males showed that only the factor loading of one item for the male group was significantly different from the female group at the 0.05 level. DIF results also suggested that the majority of the items displayed no DIF. The source of DIF may be attributed to the group mean difference on the latent trait and their real differences on certain aspects of language ability measured in this test. This provided backing for the impartiality of score interpretations, indicating that the rating-based interpretations from GSLPA SLT are impartial to a large extent across subgroups of test takers (males vs. females; business vs. non-business).

In order to examine the consistency of test scores, Generalizability theory (G theory) analyses were performed to investigate whether the test was dependable and whether the five dimensions were separable. G theory results showed that the phi coefficient for the whole test fell between .76 and .85 and Grammar and Vocabulary and Pronunciation proved to be the most dependable dimensions. G theory and CFA results both confirmed that the five speaking dimension were highly correlated with each other. The possible reasons of these findings were further discussed with reference to the raters' verbal reports.

Based on the above results, it can be concluded that the meaningfulness, impartiality, and consistency could be justified to a large extent. Some critical areas to be improved in the test design and administration were identified. Theoretical and practical implications were addressed and methodological limitations were also discussed. Overall, this study highlights the usefulness of Bachman and Palmer's Assessment Use Argument (2010) to justify the use of an existing language assessment.

The dissertation of Yujie Jia is approved.

Peter M. Bentler

Noreen M. Webb

Lyle F. Bachman, Committee Chair

University of California, Los Angeles

2013

# DEDICATION

To my parents
for their support and inspirations
throughout my pursuit of PhD degree.

## Table of Contents

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I owe thanks to a lot of people in the process of my dissertation writing as well as throughout my whole PhD study at UCLA. First of all, I would like to express my deepest gratitude to Professor Lyle F. Bachman, chair of my dissertation committee. Without his encouragement and support I could not have finished writing up this dissertation and completed my PhD degree. He was always there whenever I needed his help. He often told me that everything would work out when I had difficulties. His way of tackling tough problems in an easy manner helped me go through a lot of hardships. His extraordinary wisdom, professional expertise, and enormous charisma also made my three years at UCLA quite productive and enjoyable.

My sincere thanks also go to the other members of my dissertation committee: Prof. Peter Bentler, Prof. Noreen Webb, and Prof. John Schumann. Their rigorous scholarship and enthusiasm for research shaped my way of doing research. I took structural equation modeling class with Peter. His comments on my final paper reassured me of its good quality and strengthened my confidence with multivariate statistical analyses. He was very supportive when I worked on my qualifying paper, dissertation proposal, and dissertation. His timely endorsement has made my PhD study more efficient. Noreen’s classes were very appealing and she could introduce complex terms and knowledge with simple words. From her classes I began to know about generalizability theory. Her guidance and suggestions on my final paper provided many insights into my dissertation writing. Our dissertation meetings at the final stage of my dissertation helped solve one of the toughest problems and hence my dissertation writing could move much faster. John’s classes made me more familiar with doctoral research in applied linguistics. I was appreciative of his willingness to serve on my dissertation committee. His kindness and great personality made my dissertation completion smoother.

There are still some other professors at UCLA to whom I wish to express gratitude. Through Prof. Steven Reise's classes, I had a better understanding of the quantitative aspects of measurement. He also introduced item response theory to me. He could explain the key terms and concepts very clearly. He was also very helpful when I had problems with data analyses. Li Cai's advanced item response theory class strengthened my understanding about multi-dimensional IRT models. I was very grateful for his advice on my final papers.

I also feel lucky to have brilliant fellow students at UCLA: Hongwen Cai, Ikkyu Choi, Hsin-min Liu, Huan Wang, Jonathan Schmidgall, and Youngsoon So. Their companions motivated me to take many challenging classes in assessment, measurement and statistics. Our discussions at lab meetings and seminars inspired me a lot throughout my PhD study. Their suggestions on my presentations also helped refine my work. Besides, thanks also go to Lingyun Du from Department of Education. Her brightness and hard-working spirit made our collaboration a very pleasant experience.

I am indebted to CRESST for hiring me as a GSR and to Educational Testing Service for providing me TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment. I want to give my sincere thanks to these organizations for their financial support during my PhD study.

I am very grateful to Dr. Alan Urmston and Ms. Felicia Fang for their tremendous help with my data collection. Thank them for allowing me to use their test data for my dissertation. They were always ready to answer any test-related questions. Without their support I could have spent more years on my PhD study and dissertation writing.

Last but not least, I would like to give my special thanks to my parents for their continuing care, love and support. They tried all their means to help achieve my goals. Without their financial and spiritual support, I would not have the courage to pursue my PhD degree. The accomplishment of my dissertation is the best gift for them to acknowledge their pay in the past years.

# VITA

2003, BA in English Language and Literature, Shandong University

2007, MA in Applied Linguistics, Graduate University of Chinese Academy of Sciences

2010, Summer Intern, CTB/McGraw-Hill Summer Research Internship Program

2011, Special Reader, Department of Applied Linguistics, UCLA

2012, Teaching Assistant, Department of Asian Languages and Cultures, UCLA

2010-2012, Graduate Student Researcher, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles

## Publications

Jia, Y. (2009). Ethical standards for language testing professionals: A comparative analysis of five major codes. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *13* (2), 2-8.

Jia, Y. (2007). A cognitive study on the polysemy of the preposition *through*. *Journal of Central Chinese Normal University,* 42-46.

Jia, Y. & Zhang, W. (2006). Evaluating the constrcut validity of an EFL test for PhD candidates: A quantitative analysis of two versions. *Shiken:JALT Testing & Evaluation SIG Newsletter, 11*(1), 2-16.

Jia, Y. (2006). Improving undergraduates' writing ability by discourse teaching. *Journal of Wuhan University of Science and Technology. Vol. 8,* 225-227.

## Presentations

Jia, Y. (2012).“ Using Multivariate Generalizability Theory to Investigate the Dependability of a Computer-based Oral Test”. Paper to be presented at the National Council on Measurement in Education (NCME) Annual Meeting. Vancouver, British Columbia, Canada, April, 2012.

Jia, Y. (2011).“Investigating the Relationship between Self-assessment and Oral Test Performance”. Paper presented at the joint Conference of the Midwest Association of Language Testers and Technology for Second Language Learning. Iowa State University, Iowa, September, 2011.

Jia, Y. (2011). “entitled Justifying Score-based Interpretations from a Second Language Oral Test: Multi-group Confirmatory Factor Analysis”. Paper presented at the 33rd Language Testing Research Colloquium, Ann Arbor, Michigan, June, 2011.

Jia, Y., Urmston, A. & Fang, F. (2011). “Investigating the Dependability of Analytic Scoring for a Computer-based Oral Test”. Paper presented at the 33rd Language Testing Research Colloquium, Ann Arbor, Michigan, June, 2011.

Jia, Y. (2010). “ Using CFA Approach to Investigate the Construct Validity of the Analytic Rating Scales in a Semi-direct Oral Test”. Paper presented at the 13th Annual Conference of Southern California Association for Language Assessment Researchers, UCLA, May, 2010.

Jia, Y. (2009).“ Investigating test-taking strategies and test takers’ performance on a semi-direct academic oral test”. Paper presented at the 2nd International Conference on English, Discourse and Intercultural Communication, Macao, June, 2009.

Jia, Y. (2009).“Investigating the construct validity of an EFL reading test with two different question types.” Paper presented at the 2009 Language Training & Testing Center International Conference on English Language Teaching and Testing, Taipei, Taiwan, March 6-7, 2009.

Jia, Y. (2008). “Do we need a seperate code of ethics for language testing in China?”Paper presented at the 4th International Conference on Teaching English at Tertiary Level, Zhejiang, China, October, 2008.

Jia, Y. (2006). “Construct validity study of an EFL reading test for Chinese Doctoral candidates.” Paper presented at the International Conference on Language Testing, Guangzhou, China, December, 2006.

# Chapter 1 The Problem and Its Setting

## *1.1 Statement of the problem*

### 1.1.1 English language assessment in Hong Kong

The English language plays a central role in the everyday life of Hong Kong as an international center of finance, business, trade and tourism. With the return of sovereignty to mainland China in 1997 and the emergence of mainland China's influence in Hong Kong, there has been a concern about a perceived decline in the English language proficiency of Hong Kong university students. One common perception persists that students graduating from Hong Kong's tertiary institutions do not possess adequate English language skills to communicate effectively in the workplace settings. Those perceptions of declining standards of English among recent university graduates are so widespread that the Chief Executive of Hong Kong explicitly raised this as an issue in his first policy address in October 1997. One response to complaints of declining levels of English proficiency has been to propose an exit test that students should take shortly before graduation. The idea of introducing a language test as an exit test was first discussed at Hong Kong Polytechnic University (HKPU) in order to motivate students to improve their language performance. It also aimed to provide Hong Kong employers with reliable and accurate information about university students' English proficiency.

From 1994 to 1997 the Graduating Students' Language Proficiency Assessment (GSLPA) was developed at HKPU as an exit test for graduating students. Initially it was intended to be used for students from all the universities in Hong Kong. Between the years 1997 and 1999, the GSLPA went through extensive formal trialling and field testing at three universities: HKPU, Lingnan University and University of Hong Kong. It became apparent fairly soon that the use of a single English language exit test across all universities was impossible at that time. According to Lumley and Qian (2003), a major reason for this was the desire of institutions to maintain their autonomy, and the fear that a ‘league table’ of performance on such a test would be developed. A second reason was the difficulty of developing a test that was suitable for students in all disciplines. A further reason was the desire to avoid the negative washback of a shrinking curriculum often associated with standardized tests, which was a dominant feature of Hong Kong education. Nevertheless, work continued on the GSLPA within PolyU, resulting in the development and trailing of successive versions of a test instrument. In 2000, the test was fully operational at HKPU and administered to its final- year students.

The GSLPA has two components: speaking and writing. The content focuses on the professional workplace communication needs of recent graduates. In this way it looks forward to employment, rather than backwards at the academic context of university study (Lumley & Qian, 2003). This is consistent with its major aim of providing information to prospective employers about the English language proficiency of graduating students. As a standardized English proficiency test, the GSLPA has a number of characteristics that distinguish it from some popular commercial English proficiency tests commonly associated with university students, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS). In identifying an appropriate English proficiency test for graduating students at PolyU, a number of existing language tests (e.g., TOEFL, IELTS) were considered but rejected because their focus was on the university entry and mainly used for admission decisions whereas GSLPA focused more on exit and graduation and employment decisions in Hong Kong context were made based on the test use (Lumley & Qian, 2003). Predictions made by these tests were related primarily to courses of university study rather than the professional employment. This is also one of the most distinct features of the GSLPA.

### 1.1.2 Assessment use argument

In the real world language tests are developed to collect information for making decisions and serve for one or several purposes. These test uses or decisions may have serious consequences for the stakeholder groups of the tests. Bachman (1990) asserted that 'The single most important consideration in both the development of language tests and the interpretation of their results is the purpose or purposes which the particular tests are intended to serve' (p. 55). He also pointed out that it is test developers' responsibility to 'provide as complete evidence as possible that the tests that are used are valid indicators of the abilities of interest and that these abilities are appropriate to the intended use, and then to insist that this evidence be used in the determination of test use'(p.285). Bachman and Palmer (2010) further argued that language assessments are primarily used to promote beneficial consequences for the stakeholders, or the individuals, programs, or societies that will be affected by the assessments. They stressed two axioms for test developers and decision makers: 1) to be accountable to the stakeholder and 2) to demonstrate that the use of a particular assessment is justified through argumentation and the collection of supporting evidence. In light of concerns mentioned above, Bachman and Palmer (2010) proposed the Assessment Use Argument (AUA) as a conceptual and systematic framework to guide the development and use of a particular language assessment, including the interpretations and uses on the basis of the assessment. An AUA can be adopted to investigate the extent to which the intended use of a particular assessment is justified.

## *1.2 Research questions*

This study intends to investigate the extent to which the use of GSLPA Spoken Language Test (SLT) as an exit test at PolyU to the stakeholders can be justified. It employs the conceptual framework of Bachman and Palmer's (2010) Assessment Use Argument (AUA) to articulate claims and warrants about this oral assessment. More specifically, this study aims to address the following research questions:

1. Are the assessment records consistent across different assessment tasks?

   1.1. To what extent are the GSLPA Spoken Language Test and the individual speaking tasks dependable?

   1.2. To what extent are the analytic scores separable in terms of task fulfillment and relevance, clarity of presentation, grammar and vocabulary, pronunciation, and confidence and fluency?

2. Are the score-based interpretations about students' oral proficiency for workplace communication meaningful?

   2.1. Is the multi-componential factor structure assumed in this test design supported?

   2.2. Are there any problematic items that are weak in measuring test takers' speaking ability?

   2.3. Do tasks have effects on test takers' speaking performance?

   2.4. To what extent does what the test takers report correspond to their oral test performance?

3. Are the score-based interpretations impartial across different subgroups of test takers (males vs. females; business vs. non-business majors)?

This study identifies gender as one of the possible factors that cause differences in test performance. The second factor is the test takers' academic major, which is taken to be a reflection of subject background knowledge, since it is likely to be associated in some way with the topical content of the speaking tasks.

## *1.3 Definitions of key terms*

### 1.3.1 Assessment, assessment use, and assessment justification

An assessment is a procedure for collecting and recording information, and assessment use is an instance of using the assessment for making decisions. An Assessment Use Argument (AUA) is "a conceptual framework for guiding the development and use of a particular language assessment, including the interpretations and uses we make on the basis of the assessments" (Bachman & Palmer, 2010, p.99). An AUA comprises a set of claims that link test takers' performance to the consequences of using the assessment for making decisions.

Assessment justification is defined as the process that test developers will follow to investigate the extent to which the intended uses of an assessment are justified (Bachman & Palmer, 2010). This process involves two activities: 1) the articulation of specific statements in an Assessment Use Argument (AUA) and 2) the collection of relevant evidence or backing in support of the statements. The process of justification can guide the development and use of a given language assessment, provide the basis for quality control, and provide the basis for the accountability of test developers and decision maker held to the stakeholder groups. It should be noted that justification studies can never prove that the intended uses of the assessment are “true”, “valid”, or “correct”. Since assessment situations vary in different ways such as test takers and construct to be assessed, the justification process is local and relevant to every specific assessment situation. Given the fact that the conditions of the assessment situation can change over time, the process of justification is ongoing and the AUA to an assessment should be regularly reviewed and revised.

An AUA for a given assessment consists of two elements: claims and data. Claims are statements about the inferences to be made on the basis of data and the qualities of those inferences. A claim includes an outcome of the assessment process and one or more qualities of that outcome. Meaningfulness, impartiality and consistency addressed in this study are three qualities in an AUA framework. Qualities of the outcomes in AUA framework have no rank ordering in terms of their importance. These three qualities are chosen in this study because of the availability of the relevant evidence and backing for them.

### 1.3.2 Meaningfulness in AUA

Meaningfulness refers to “the extent to which a given assessment record 1) provides stakeholders with information about the ability or construct to be assessed, and 2) the extent to which this information is conveyed in terms that they can understand and relate to” (Bachman & Palmer, 2010, p.114). The meaningfulness of the interpretations is related to how we define the construct to be assessed and how we communicate this to stakeholders. In most language assessment settings, constructs are defined based on a language learning syllabus, a needs analysis of the abilities required to perform target language use tasks, or a language ability theory. There are many ways to provide backing or evidence for meaningfulness of the score interpretations, such as confirmatory factor analysis (CFA), discourse analysis of assessment performance or verbal protocol analysis.

### 1.3.3 Impartiality in AUA

Impartiality is defined as “the degree to which the format and content of the assessment tasks and all aspects of the administration of the assessment are free from bias that may favor or disfavor some test takers” (Bachman & Palmer, 2010, p.115). If the test takers at the same level on the construct perform differently on the assessment, there must be a test bias. The test format and content can both affect test takers’ performance on language assessments. In other words, if the differences in test takers’ language performance are not due to their differences in language ability, a test bias may occur. Survey of test takers or statistical analyses like multi-group CFA or differential item functioning (DIF) can be used to investigate the impartiality issue.

### 1.3.4 Consistency in AUA

Consistency refers to “the extent to which test takers’ performance on different assessments of the same construct yield essentially the same assessment records” (Bachman & Palmer, 2010, p.124). In AUA framework, “consistency” is a quality that is claimed for assessment records (scores, verbal descriptions). Evidence to support consistency comes from a variety of sources, including not only the quantitative analysis of test scores with measurement theory, but also the qualitative analyses of assessment performance, and the procedures that are followed in the administration, scoring/describing and reporting of the assessment results.

## *1.4 Significance of the Research*

### 1.4.1 Theoretical significance of the research

If the intended assessment use is critical to the development and evaluation of an assessment, as argued by Bachman and Palmer (2010), it is important to justify the assessment use to the stakeholders. The present study offers insight into justifications of the intended assessment use by investigating the use of a university second language oral proficiency test as an exit test in Hong Kong. It can serve as an example to illustrate how to justify the assessment uses in light of Assessment Use Argument (AUA) framework. The AUA framework can also guide efforts in relating the intended assessment uses to assessment design and administration, thus providing useful information to assessment developers for improvement of the test. In addition, the framework is promising in delineating the responsibilities of the assessment developers from those of assessment users, which is often a complicated issue in assessment evaluation and accountability.

### 1.4.2 Practical significance of the research

The present study has implications for language assessment practice in that the findings may provide valuable feedback to the assessment developers for improvement in the assessment. The study results may also provide implications for the school authority under concern regarding the test use.

In addition, this study has the potential to make contributions to the design and development of computer-based oral testing in general. It can provide a better understanding about English as a Second Language (ESL) students' test-taking processes on computer-based oral tests. The findings will also have practical implications for rater training and monitoring for computer-based oral tests.

# Chapter 2 Literature Review

## *2.1 Argument-based approach to validity*

In the first half of the twentieth century, assessment and measurement researchers have generally paid much attention to the reliability and validity of tests, with reliability providing an indication of the consistency of test scores, and validity addressing the meaning and utility of the scores. Following the lead of Cronbach (1980), Kane (1992), and Mislevy (2003), the idea of a validity argument has become widely respected within the educational measurement field. Kane (1992) developed the notion of an interpretive argument as providing framework for the gathering and disseminating of evidence supporting intended score interpretations. Drawing on the literature in practical argumentation, Kane described an interpretive argument as consisting of inferences and assumptions, which needed to be supported by relevant evidence. Building on Kane (1992), Kane, Crooks, and Cohen (1999) explicated the details of an interpretive argument for linking observations to interpretations. Kane (2006) has proposed four types of inferences in the network of inferences comprising the interpretive argument. These include *scoring, generalization, extrapolation,* and *decision.* Each inference 'involves an extension of the interpretation or a decision' (p. 23) which allows for checking and confirming of a previous interpretation or decision. As the first inference states that the observed score is a reflection of the observed behavior, the first level of the chain of inferences is to scrutinize the fidelity of the scoring procedures *(scoring)* in the way it is intended to be used. The second inference which Kane terms *generalization* links the observed score and what he refers to as the universe score. The third inference, extrapolation, is closely related to the concept of construct validity and can be evaluated using both analytical and empirical evidence. The fourth type of inference from the target score to the decision based on the test scores consigns the test to the realm of test use and consequences.

Kane (2004, 2006) has begun to address the role of test use, decisions and consequences by extending the interpretive argument described by Kane et al. (1999). The other key point made by Kane (2006) is the need for a systematic and organized way of formulating or framing validation research. In language testing, Bachman (2005) and Fulcher and Davidson (2007) have drawn on to the works of Kane (1992, 2002) and Toulmin (2003) in an effort to make validation more manageable, accessible, and transparent. Bachman (2005) stated that argument-based formulations provide the logic and a set of procedures for investigating and supporting claims about score-based inferences. He also addressed the issues of test use and its consequences. Bachman (2005) also discussed the feasibility of using the argument-based approach in language testing. The demands of a strong program of validation based on explicitly stated hypotheses and assumptions can be quite daunting and taxing on resources. Test developers would rather opt for a weaker construct validation program requiring the collection of easy evidence that provides support for their intended interpretations rather than adopting a strong program, for fear that their interpretations or arguments might be called into question. Fulcher and Davidson (2007) also provide numerous examples of how arguments can be used at the item or test levels.

Chapelle, Enright, and Jamieson (2004) have utilized an argument-based approach to build a validity framework for the internet-based Test of English as a Foreign Language (TOEFL iBT). With a restriction to descriptive interpretations and semantic inferences during the test development phase, they point out that the decision-based and policy inferences about test use and washback could only be carried out once the test is operational. This particular study represents a forward move in language testing from a highly abstract unified model of validity to a more transparent and usable argument-based approach to validation (Bachman, 2005; McNamara & Roever, 2006). In the book edited by Chapelle, Enright, and Jamieson (2008), they provided a detailed and reflective overview of the process of developing the TOEFL iBT. They also showed how the argument-based approach is useful for the test development and validation processes with some evidence from the operational use of the test.

Kadir (2008) drew on Kane's (2006) argument-based approach to validate an occupational language assessment and evaluate its usefulness and impact. In utilizing this approach to validation, Kadir examined the claims for the use of the test using a network of inferences forming the basis for the validity argument. This included the examination of the evidence from scoring procedures, generalization of observed scores to universe of scores, extrapolation of observed scores to non-test behavior, and investigating the impact of the test on the public service. Overall the argument framework for test validation as proposed by Kane worked well for the intended purpose of the study. Some weaknesses relating to the use of this framework were also mentioned in this study. These include the demands for comprehensiveness and the multitude of evidence needed in order to evaluate the strength of the validity argument for examining test use and impact and whether these demands can be met effectively by a single researcher in a single study.

### *2.2 Assessment Use Argument*

Bachman and Palmer (2010) developed a conceptual framework, an Assessment Use Argument (AUA), to guide the development and use of language assessments. In the framework, they subsume traditional notions of reliability and validity under qualities of claims in an AUA. To the extent that validity is defined in terms of how well an assessment program achieves its goals, it is necessary to pay some attention to consequences, positive and negative. Some researchers (Messick, 1989, 1994; Linn, 1997; Shepard, 1997; Kane, 2006) have advocated for a conception of validity that involves both the meaning of assessment scores and the consequences of their use. Bachman and Palmer (2010) have argued that an AUA provides a conceptual framework for justifying the assessment use. They state that,

> The AUA consists of a set of claims that specify the conceptual links between a test taker's *performance* on an assessment, an *assessment record*, which is the score or qualitative description we obtain from the assessment, an *interpretation* about the ability we want to assess, the *decisions* that are to be made, and the *consequences* of using the assessment and of the decisions that are made. (p. 30)

Bachman and Palmer make score uses and the consequences of score uses the centerpiece of their discussion: 'An AUA provides the conceptual framework for linking a claim about a particular set of consequences to the performance of individuals on a language assessment' (p. 156). Given the high stakes of many emerging uses of assessment systems (e.g. in school and teacher accountability, in employment and immigration decisions), the analysis of consequences in justifying assessment programs is becoming increasingly important.

They adopted an argument-based approach to validation for the development of the AUA framework. Following Toulmin's analysis of practical reasoning, the AUA includes "the following elements: data, claims, warrants, backing, rebuttals, and rebuttal backing' (p. 99). Bachman and Palmer also extend the argument-based framework in several ways. They base their approach to language assessment development and use fundamentally on ‘the need for a clearly articulated and coherent Assessment Use Argument (AUA)’ and on ‘the provision of evidence to support the statements in the AUA’ (p. 31). They adopt the framework and terminology of an argument-based approach, but they do not emphasize the validity of a proposed interpretation per se. Rather, they are concerned with the general question of the justification for assessment uses, with the justification of proposed interpretations constituting one of several major claims in the AUA.

Bachman and Palmer (2010) maintain that ‘Assessment justification consists of articulating an Assessment Use Argument (AUA) and collecting evidence to support this’ (p. 30). The process of *assessment justification* can be regarded as a process of articulating the claims and warrants in an AUA and providing backing or evidence to support the claims and warrants. Bachman and Palmer (2010) regard the AUA as an approach that can be tailored to guide the development and use of a specific assessment for a specific purpose for a specific group of test takers at a specific time in a specific situation. They state that ‘[a]ssessment development and use, and the process of justification are necessarily local’ (p. 438). The process of assessment justification is local because both the articulated AUA and the collected backing for support of the warrants or rebuttals are context-specific.

One example of a study utilizing Bachman and Palmer’s (2010) AUA framework is that of Wang (2010) where evidence was collected to justify an added use of college-level English as a Foreign Language (EFL) proficiency test. The study compared the originally intended and added assessment uses and linked the observed differences to desired modified or additional conditions for justifying the added use. Based on the comparison and linking results, five key modified warrants were identified for supporting the added use. Among them, the warrant of the equitability of decisions was identified as the potentially most questionable one. Wang concluded that the most critical area in the current test design and administration for supporting the added use is likely to be the observed substantial measurement errors of the paper test due to construct-irrelevant factors. Accordingly, she recommended that test developers need to focus most on identifying construct-irrelevant factors measured in the paper test and addressing them accordingly.

### *2.3 The multi-componential nature of L2 Speaking construct*

For language testers it is crucial to meaningfully measure the second language (L2) proficiency of test takers and make inferences from the test scores to a test taker's ability to use language for an identified purpose. Since the late 1960s, the language testing field has paid increasing attention to the nature of the L2 construct. Chalhoub-Deville and Deville (2005) traced the development in the field over the years of the construct definition of language proficiency. After Lado (1961), some of the most influential works have been (in chronological order): Oller (1979), Canale and Swain (1980), Omaggio (1986), Bachman (1990) and Bachman and Palmer (1996), and McNamara (1996). Chalhoub-Deville and Deville argued that the construct had largely been defined according to a psycholinguistic and cognitive paradigm. Language testers viewed the L2 construct as a stable and homogenous set of ability components.

One of the main problems underlying speaking tests is that 'speaking' is a difficult construct to define (Fulcher, 2003). One very popular although much criticized notion of spoken proficiency in second language contexts is that described in the ACTFL Guidelines (1985, 1999), where proficiency is described in terms of communicative growth. Different levels of proficiency are described in a hierarchical sequence of performance ranges. The guidelines describe proficiency as constituting of four factors: function, content, context, and accuracy. A number of researchers have considered the relative weight of individual features of performance in determining overall judgments of proficiency based on the ACTFL Scale and its predecessors. For example, Adams (1980) investigated the relationship between the five factors which were identified in assessing the Foreign Service Institute (FSI) Oral Interview Test of Speaking (i.e. accent, comprehension, vocabulary, fluency, and grammar) and the global speaking score (e.g. on a scale of 1–5) by analyzing analytic and overall score data drawn from test performances in various languages. The main factors distinguishing levels were found to be vocabulary and grammar, with accent and fluency failing to discriminate at several levels. Higgs and Clifford (1982) suggested that different factors contributed differently to overall language proficiency at the different levels defined in the FSI scale, and proposed the Relative Contribution Model (RCM) to describe rater perceptions of the relative role of each of five component factors making up global proficiency (i.e. vocabulary, grammar, pronunciation, fluency, and sociolinguistics). In their hypothesized model, vocabulary and grammar were considered to be the most important across all levels, but as the level increased, other factors such as pronunciation, fluency, and sociolinguistic factors would also become important.

Other researchers have also investigated the componential structure of proficiency at varying levels using other test instruments. De Jong and van Ginkel (1992) used speaking test data from 25 secondary school level students of French to investigate the relative contribution of different aspects of oral proficiency to the global proficiency score. The results revealed that the pronunciation category contributed most to global proficiency at the lower level, but as the level went up fluency became more important.

The contribution of accuracy and comprehensibility did not vary across the levels. McNamara (1990), validating the Speaking sub-test of the Occupational English Test (OET), a specific purpose test for health professionals, investigated the relationship between the global score (Overall Communicative Effectiveness) and five analytic scales (Resources of Grammar and Expression, Intelligibility, Appropriateness, Comprehension, and Fluency). An analysis using Rasch Item Response Modelling identified Resources of Grammar and Expression as the strongest determinant of the score for Overall Communicative Effectiveness; it was also the most ‘difficult’, that is the most harshly rated criterion (comprehension was scored most leniently). According to Fulcher (2003), speech can be broken down into pronunciation and intonation, accuracy and fluency, or it can be categorized in terms of strategies, or it can be regarded as a form of interaction and analyzed using the methods of pragmatics or discourse analysis. In the course of a normal conversation, all of these aspects are considered to be important. If testers try to separate out the strands, they may well find that the ecology of speaking is different in different successful speakers. This means that the accurate speaker may communicate effectively, but slowly, whereas the fluent speaker may sacrifice accuracy for the sake of rapid communication (Skehan, 1998).

Sawaki (2007) combined Confirmatory Factor Analysis (CFA) and multivariate generalizability theory (G theory) to analyze a Spanish speaking assessment designed for student placement and diagnosis. The results generally confirmed the key features of the assessment design:(1) the multicomponential and yet highly correlated nature of the five analytic rating scales: Pronunciation, Vocabulary, Cohesion, Organization and Grammar, (2) the high dependability of the ratings and the resulting placement decisions appropriate for the high-stakes decision-making context based on these analytic rating scales and the composite score, and (3) the largest contribution of Grammar to the composite score variance, which was consistent with the intention of program faculty members to reflect in the test design the relative importance of knowledge of grammar for students' academic success in the study-abroad program.

In conclusion, both theoretical arguments and empirical evidence have ascertained the multicomponentiality of the L2 speaking construct in the language testing field. However, previous findings varied in terms of the specific components identifies and their relative weighting in overall ratings of speaking. These divergent findings can be attributed, to some extent, to the different construct definitions of the oral tests and different characteristics of the populations of test takers involved. Nevertheless, the overall result of these studies is that L2 speaking is a multicomponential ability.

## *2.4 Effects of gender and academic majors on test performance*

### 2.4.1 Investigation of test fairness using DIF and Multi-group CFA

In measuring English as a second language (ESL) learners' ability, researchers and theorists have demonstrated that numerous factors other than language ability can affect test performance. The effects of test takers' characteristics (e.g., gender, language background) on test taker performance have been one of the primary concerns among language testers and researchers, relating to the issue of test fairness and equivalence. It is essential to investigate whether a test includes a potential bias against some particular groups of test-takers. Previous studies on the effect of test taker characteristics on language tests have been mainly concerned with two issues. One is to investigate whether the construct or the structure of the test is invariant across different groups. In other words, whether a test measures the same constructs for various groups has been one of the primary concerns of language testing researchers. The other issue is whether the difference in test performance is due to the task-takers' personal attributes. This issue has been investigated at the item level in differential item functioning (DIF). These two issues are highly interrelated because if the constructs measured by the test vary across the different language groups, this may be attributable to one or more DIF items on the test. Therefore, the studies that investigate the structural relationship of the test across different groups of test takers are relevant to the investigation of DIF.

The effects of test takers' language background on their test performance, has been the most frequently investigated factor in the English as a second or foreign language (ESL/EFL) testing field. Some researchers (e.g., Kunnan, 1994; Brown, 1999; Stricker & Rock, 2008;) employed confirmatory factor analysis or structural equation modelling to examine whether the test structure was invariant across different language groups, while others (Chen & Henning, 1985; Sasaki, 1991; Ryan & Bachman, 1992; Kim, 2001) attempted to identify DIF items across different native language backgrounds in the language tests.

### 2.4.2 Gender effects on test performance

Sunderland (1995) argued that there was evidence that a test or exam could favor female or male test takers in three possible ways: “Topic”, “Task” and “Tester”. Some topics could possibly be more accessible or familiar to males or females, although the only evidence to date supporting this claim remains anecdotal. On the contrary, O'Loughlin (2002) found no evidence of a gender effect either on the scores achieved by male and female candidates or on features of the discourse. With multi-faceted Rasch analysis, Lumley and O'Sullivan (2005) showed little effects for some of the hypothesized interactions of variables such as the task topic, the gender of the person presenting the topic and the gender of the GSLPA SLT candidates.

Most studies have employed DIF to investigate gender effects on second or foreign language test performance. Gafni (1991) examined gender DIF on two forms of the English subtest of the Israeli Psychometric Entrance Test (PET). Overall, Gafni found males were likely to perform better on items containing technical content than their female counterparts. Similarly, Ryan and Bachman (1992) reported the presence of gender DIF on the First Certificate of English (FCE) and Test of English as a Foreign Language (TOEFL) in the content categories of structure, vocabulary, and reading. Takala and Kaftandjieva (2000) studied gender DIF in the vocabulary subtest of the Finnish Foreign Language Certificate Examination, using the IRT One Parameter Logistic Model (OPLM). Takala and Kaftandjieva suggested that regardless of DIF findings at the item level, this vocabulary test did not show gender bias at the test level, because almost the same number of DIF items favored each group. More recently, Pae (2004) analyzed gender DIF on the English subtest of the KCSAT, using the MH as well as IRT-LR procedures, and found that whereas reading comprehension items classified as logical inference were highly likely to favor Korean males, the reading items pertaining to the mood, impression, or tone of a given passage were likely to favor Korean females.

### 2.4.3 The effects of background knowledge on test performance

For the effects of background knowledge on test performance, Clapham (1996) implemented the three-module reading test of the International English Language Testing System (IELTS) and found that students in general tended to perform significantly better on the reading module in their own subject area. Chung and Berry (2000) confirmed the findings from Clapham's study with a homogenous language group of secondary school students in Hong Kong. Using the IELTS reading test of science/technology module and a popular science text, they found that to some extent examinees' background knowledge of text content predicted reading comprehension. Jensen and Hansen (1995) compared 100 subjects' listening performance on 11 academic lectures with their self-reported prior knowledge of the lecture topics. Multiple regression analysis revealed that prior knowledge effect was significant for only 5 of the eleven lectures. The prior knowledge effect was stronger for technical lectures than nontechincal lectures, although the significance of the effect was still very trivial. The authors concluded that listening comprehension does not seem to be affected by prior knowledge. In contrast, Chiang and Dunkel (1992) discovered that prior knowledge played a significant role in understanding lectures. Long's (1990) study of learners of Spanish as a Foreign Language corroborated these results. Subjects were given a summary task after listening to the taped lectures, one familiar and one unfamiliar in terms of cultural themes. Long found that the recall tasks were per- formed significantly better for the topic familiar to the subjects than those for the unfamiliar topics.

As shown above, most studies have examined the effects of gender and background knowledge on listening, reading or writing performance. However, in the L2 speaking literature surprisingly little research has been undertaken to investigate the impacts of test takers' background knowledge and gender on the speaking performance. Particularly, DIF or multi-group CFA studies that examined the effect of gender and background knowledge on speaking performance are very rare. In this respect, DIF or multi-group CFA investigations would bridge a gap in the L2 speaking literature because such a study will provide information about whether speaking tasks function differently for examinees with different academic backgrounds or between males and females.

### *2.5 Holistic vs. analytic scoring*

In language testing literature, there has been considerable debate about the merits and limitations of holistic and analytic rating rubrics for speaking tests (Bachman & Savignon, 1986; Douglas & Smith, 1997; Fulcher, 1997; Ingram & Wylie, 1993; Underhill, 1987; Weir, 1990). In order to determine which rating scales are adopted, three factors may need to be considered: 1) the availability of rich information about examinees' language ability (Bachman, Lynch & Mason, 1995); 2) increased accuracy of ratings by drawing judges' attention to specific criteria (Brown & Bailey, 1984); and 3) consistency with the current multicomponential definition of language ability (Bachman, Lynch & Mason, 1995).

Xi (2006) summarized the advantages of holistic scoring as efficiency in scoring, ease in score reporting and a lesser cognitive load on raters. She also pointed out that holistic also had some problems. First, the relative weights of the sub-features defined in the scoring rubric are implicit. Based on raters' background and experience, the contributions of each component to the overall language ability may be weighted differentially. Another problem with holistic scoring is related to the interpretation of the scores. As Weir (1990) pointed out, the typical performance descriptions at each holistic score level might not work for candidates with varied performances on the components.

Analytic scores can provide the diagnostic information for examinees with varied profiles (Bachman & Savignon, 1986). The analytic scores can be reported in various forms. Multiple scores from the anlytic scales can be reported separately as a language profile. They can also be reported in the form of a composite score obtained by averaging or summing across the scores on analytic scales by weighting all components equally (e.g., Brown & Bailey, 1984; Kondo-Brown,

2002) or differentially (e.g., Weigle, 1998). Sawaki (2007) suggest that empirical evidence may be needed to justify the usefulness of the analytic ratings scales for an intended purpose. First, empirical studies among the analytic rating scales must demonstrate that the scales are not only related to one another but also have some distinctions from each other. In this way, each analytic score can provide some information about a component of a candidate's language ability. Second, when an overall score is reported, the empirical relationship of the analytic scales to the overall score, i.e., the weighting of individual analytic rating scales in an overall score, should be congruent with the relative importance of different aspects of language ability for a given purpose of assessment in a particular context.

For many language test developers, analytic scoring is preferable over holistic scoring. However, it also has its own problems which include potential rating inconsistencies due to high cognitive load on raters, difficulty in defining the dimensions in analytic rubrics precisely and getting raters well calibrated (Douglas & Smith, 1997; Underhill, 1987). Underhill (1987) reported the difficulty raters experienced when having to evaluate the candidate's performance on several criteria simultaneously. Douglas and Smith (1997) argued that it was very difficult to define the components precisely and raters may have different interpretations of the analytic rating scales. Raters must be well-trained so that they can differentiate the criteria reliably. Furthermore, the test population must demonstrate sufficiently varied profiles to warrant a more costly and complex analytic scoring system (Xi, 2006).

## *2.6 Verbal protocol in language testing*

### 2.6.1 Verbal protocol

After Ericsson and Simon (1984), second language researchers (e.g., Faerch & Kasper, 1987; Gass & Mackey, 2000; Green, 1998) began to pay attention to verbal protocol analysis (VPA). Many empirical studies also tended to adopt VPA as one research method in the L2 field. In particular, the use of verbal reports has gained an increasing popularity as a viable research methodology to elicit verbal reports of cognitive processes in language testing because the use of a process-oriented approach is considered crucial for test validation (Embretson, 1983; Messick, 1995). It has been used mainly to investigate test-taking strategies and processes. In recent years, language testing researchers also attempted to use it to gain a better understanding of raters' behaviors.

The strong assumptions of VPA are that subjects have "privileged access to their experiences" (Ericsson & Simon, 1993: xii), and that the information in their verbal reports is trustworthy. Gass and Mackey (2000) defined verbal protocols as the data one gets "by asking individuals to vocalize what is going through their minds as they are solving a problem or performing a task" (p. 13). VPA is a different research technique from others that involve verbal reports since they are to be used to make direct inferences about the cognitive processes of interest (Green, 1998). For language testing research, Cohen (1998, 2000) classifies verbal reports into the following three subcategories:

- Self-report: learners' general description of what they usually do when they respond to a test item or take a test (e.g., questionnaires and interviews on general test-taking behaviors).

- Self-observation: the examination of specific language behavior either introspectively (within 20 seconds; e.g., stimulated recall in Gass & Mackey, 2000 or immediate retrospection in Wu, 1998) or retrospectively (e.g., questionnaires, journal entries, and interviews on a specific test-taking instance).

- Self-revelation: concurrent think-aloud, i.e., stream-of-consciousness disclosure of thought processes while the information is being attended to.

The reliability and validity of verbal reports has been questioned especially the self-observational reports. For instance, once a cognitive skill becomes highly automatized, its underlying cognitive process may not be available for introspection. In order to improve the quality of the verbal reports, the following suggestions are often made: 1. minimize the time interval between the verbal report of cognitions and their actual occurrence; 2. use clear instructions that can help the subjects to better retrieve the information from their short-term memory; 3. train the subjects to conform to the protocol instructions.

### 2.6.2 Relevant studies on verbal reports for rater behavior

Orr (2002) used a verbal protocol analysis to investigate rater behavior. He found that raters paid attention to aspects that were not present in the rating scales. O’Donnell, Thompson, and Park (2006) conducted a verbal protocol study to understand rater behavior for second language oral assessment. O’Donnell et al. found that raters have their own internal criteria for oral rating and pay attention to those features even though they are not described in the rating bands. Yet, they were mostly successful to negotiate their internal criteria with the institutionalized criteria described in the rating bands. Brown, Iwashita, and McNamara (2005) conducted a rater orientation study using verbal reports to identify appropriate criteria for the assessment of test performance. Brown et al. found that all raters focused on the same general categories and tended to discuss the components of these categories in essentially similar ways. Fulcher (2003) argued that this type of rater behavior study using verbal protocols reveals important information about how valid the rating processes are in assigning grades. This procedure suggests valuable information for rater training and rating scale development/revision.

# Chapter 3 Methodology

## *3.1 Research approach and methodology*

The research approach used in this study employed mixed methods, including both quantitative and qualitative methods. Empirical data from the test under study were preferred because they provide the most convincing evidence in addressing the research questions in this study. In this approach, data obtained from different research methods were used conjunctively to provide a relatively comprehensive picture of the addressed issues, or to sequentially inform different stages of the research.

The qualitative methods employed in the study included verbal protocols. Regarding quantitative methods, the specific analyses to be used were determined by the nature of the warrants/rebuttals that were articulated in the AUA and the kinds of supporting evidence they required. Also considered was availability of data and the extent to which relevant model assumptions and data requirements were met. Statistical analyses are described in Section 3.6 below.

## *3.2 Population*

The present study is intended to generalize to all major stakeholders of the test program under concern. There were four major groups of stakeholders, namely, test takers, members of the testing program, ESL instructors and departments. Characteristics of each group are described below.

### 3.2.1 Test takers

There is only one administration per year, with over 3,000 test takers. All the test takers are undergraduates at the university concerned, most of whom are students in their final years of tertiary education. All undergraduates are required to take the test before they graduate. They may come from any of the following eight faculties or schools at the university that offer undergraduate programs: 1) Faculty of Applied Science and Textiles, 2) Faculty of Business, 3) Faculty of Construction and Land Use, 4) Faculty of Engineering, 5) Faculty of Health and Social Sciences, 6) Faculty of Humanities, 7) School of Design, 8) School of Hotel and Tourism Management.

### 3.2.2 Members of the testing program

There are seven members in the testing program, including one test coordinator, two test developers, three test administrators, and one testing consultant. The test coordinator is an English native speaker who holds a doctoral degree in English education. He has been in charge of the test development and administration for over 6 years. The testing consultant is a prestigious language testing expert with extensive experiences in language test development and research. The test developers are Mandarin or Cantonese native speakers who have held master degrees and are pursuing doctoral degrees in applied linguistics, or English education.

### 3.2.3 ESL instructors

There are approximately sixty ESL instructors, all of whom have Master’s or doctoral degrees in Linguistics or Applied Linguistics. Most of them have been teaching English for over ten years.

### 3.2.4 Departments

As listed earlier in this section, there are about eight faculties or schools comprising over 25 departments at the university. The major stakeholders are personnel from the academic affairs offices who are involved in making graduation decisions.

## *3.3 Samples*

Different samples were drawn depending on the nature and goal(s) of the data collection procedure. Details for each sample are described below.

### 3.3.1 Questionnaire

A questionnaire was administered to the test takers to elicit information about how they rated their own performance after they took the test. Given the feasibility of data collection and the large number of test takers at the university, the sample was obtained by the methods of convenience sampling. Specifically, those students who took the oral test in the last session of every testing day were chosen for the questionnaire study, with a total of about 359 participants.

### 3.3.2 Verbal protocols

Five trained and experienced raters from a large pool of accredited raters were recruited to take part in the rater cognition study. All are full-time teachers of English language / English linguistics at a tertiary institute in Hong Kong and hold Master’s or doctoral degrees in English language teaching / linguistics or equivalent. There were three males and two females among the five raters. Only one of them was Cantonese speaker and the other four were all English native speakers. The participants were given initial training and practice in verbal-report production.

3.3.3 Test records

In order to comply with the verbal protocols and questionnaire study, the researcher requested the complete test records of the GSLPA oral test administered in January 2012 for 999 test takers.

## *3.4 Materials*

Three types of materials were used for data analysis. They are tests administered by the testing program, questionnaires, and verbal protocols. Details about each type are presented below.

3.4.1 Test

The Spoken Language Test of the GSLPA takes place in a language laboratory and lasts for approximately 45 minutes. This test is currently rated using an analytic scoring rubric to assess examinees' performance on five dimensions: task fulfillment and relevance (TFR), clarity of presentation (CoP), grammar and vocabulary (GV), pronunciation (Pron), and confidence and fluency (CoFlu). It includes five speaking tasks, each with a different general function and purpose. These tasks require the candidates to use both listening and speaking skills. Note-taking is allowed throughout the test. Table 1 displays the task description, time allotment and dimensions to be measured for each task.

*Table 1. Descriptions of five speaking tasks in GSLPA SLT*

| Tasks | Description | Time allotment | Functions (dimensions to be |
|---|---|---|---|

|  |  |  | measured) |
|---|---|---|---|
| Task 1 | A summary of an interview | 2 minutes | TFR, CoP; GV; Pron |
| Task 2 | Answer questions as part of an interview. | 40 seconds for each of the four questions | CoFlu; GV |
| Task 3 | Provide an oral presentation of information from a written (graphic) source. | 3 minutes | Pron; CoP |
| Task 4 | Leave a telephone (or voice mail) message. | Between 30 seconds and 1 minute | TFR; CoFlu |
| Task 5 | Describe an aspect of life in Hong Kong. | 3 minutes | TFR; CoP; GV; Pron |

### 3.4.2 Questionnaire

A questionnaire was administered to 359 test takers to elicit information about the self-ratings of their performance after they completed the test. It is used to investigate whether what test takers report corresponds to their actual test performance. The questionnaire as shown in Appendix 1 consists of two sections. The first section provides profiles about the participants including their departments or schools, gender and their hometown. The second section of the questionnaire addresses how the test takers rated their performance on the dimensions of each tasks specified by the analytic rating scales. The rating is based on a 6-point Likert scale following the ratings scales of this oral test, ranging from “Very weak” to “Very strong”.

### 3.4.3 Verbal Protocols

Raters’ verbal protocols aim to provide information about how they give scores on each dimension and whether they could distinguish them well. Besides, the verbal protocol data is also expected to reveal whether raters could differentiate assessment levels and the dimensions on each task.

### 3.4.4 Test records

The researcher requested complete test records of the oral test administered in January 2012, including the analytic ratings of the students' oral responses and the students' composite test scores computed with FACETS analysis.

## *3.5 Procedures*

### 3.5.1 Test

*Test administration*. The oral test under study was administered in January 2012. It took place in a language laboratory and lasted for approximately 45 minutes. In this round of the GSLPA, over 3,000 undergraduates took the spoken language test. On each test day, 4 language labs were opened and each lab can accommodate around 30 students. Proctors of the oral test included ESL instructors and technical staff from English Language Center at Hong Kong Polytechnic University.

*Scoring*. Scoring of the oral test was performed in the month following the test administration. All the raters are qualified, experienced English language specialists. Each of them must be a full-time teacher of English language / English linguistics at a tertiary institute in Hong Kong and hold Masters of doctoral degrees in English language teaching / linguistics or equivalent. All raters undergo face-to-face training conducted by GSLPA staff and are required to meet stringent reliability criteria before they can receive an accreditation certificate. Accredited raters are required to re-certificate every two years. Dimensions of Tasks 1, 2 and 4 were rated by one rater while Task 3 and Task 5 were rated by two accredited raters. The test data with the analytic ratings were routinely analyzed by the test developers with FACETS analysis to check consistency of ratings.

### 3.5.2 Questionnaire

An electronic questionnaire was used for the purpose of easy administration and computerized data analysis. The electronic version of the questionnaire was created and compiled by the author using Survey Monkey. The technical officer of the Language Testing Unit uploaded the questionnaire onto the computer system, so when students clicked the “End” icon of the GSLPA spoken test, the questionnaire popped up automatically on their screens.

As time was very tight between each session which lasts about 45 minutes, it is not possible for students to stay behind to complete the questionnaire in most of the sessions. It is only feasible to carry out the questionnaires among the students who attend the last session of each day.

### 3.5.3 Verbal Protocols

For the raters’ cognition study, each participant was given one sample of test takers’ spoken responses. He or she was asked to provide verbal reports for each task in this sample of oral response. First, the raters were asked to listen to the performance when rating each dimension on each task (i.e., essentially straight through, but with repetition where needed) and then to describe how they made grading decisions, using any terms with which they felt comfortable. Second, they were asked to elaborate on their evaluations by pointing out whether the dimensions were distinct to each other and whether they had difficulties differentiating performance levels on each dimension of each task. After the rater awarded scores to the oral responses, the researcher may ask about the raters' rating processes retrospectively. For example, the researcher asked the raters to describe why they gave three points to the test taker or whether they had difficulties in making grading decisions. The raters' verbal reports were audio-recorded.

### 3.5.4 Procedures for protecting the rights of research participants

Since the study involves human subjects, an application for conducting the research was submitted to the Institutional Review Board (IRB), Office for Protection of Human Subjects, University of California, Los Angeles. Consent forms were obtained from test takers, ESL instructors and raters involved in the questionnaires and interviews.

## *3.6 Data analyses*

The study utilized both descriptive and inferential statistics calculated from empirical test data to facilitate investigation of the construct measured by the oral test. Descriptive statistics, confirmatory factor analysis (CFA), Generalizability theory (G theory) analyses, and item response theory (IRT) analyses were conducted with the test record. These statistical methods were selected to be the focus of the investigation because they were expected to identify the potentially most critical areas in the current test design and administration in supporting the intended assessment use. Raters' verbal protocols were qualitatively analyzed to investigate raters' rating processes.

### 3.6.1 Preliminary statistical analyses

First descriptive statistics were calculated and assumptions regarding univariate normality were inspected. Cronbach's alpha was also calculated to gain a rough idea of the extent to which the items of interest were reliable indicators of the intended constructs. Independent samples t-tests were conducted to investigate whether males and females have significant mean differences on the 14 analytic scores and whether business and non-business majors perform differently on the 14 items. Considering two tasks in this oral test were double rated and the other three tasks were rated only by one rater, the double ratings of each dimension on Task 3 and Task 5 were averaged and then rounded off for the easiness of the statistical analyses. For instance, the averaged 3.5 is rounded up to 4. Correlations among the averaged ratings, rounded ratings and the original ones were conducted to examine whether the averaged ratings or rounded ratings were more closely correlated with the original ones.

### 3.6.2 Confirmatory factor analyses (CFA)

CFA offers a sequential model testing framework for explicitly supporting or rejecting competing explanations about the relationships among analytic rating scales. Besides, CFA was also used to investigate whether what students' self-ratings of their own performance correspond to their actual test performance. Maximum-likelihood (ML) was used as the model parameter estimation method. Multiple criteria below were employed in order to assess the overall goodness of fit of the CFA models:

1. Minimum fit function chi-square: A statistically non-significant model chi-square statistic indicates an adequate model fit,

2. The Root Mean Square Error of Approximation (RMSEA): RMSEA values of 0.0, 0.05, 0.08 correspond to rough cut-off points for exact fit, close fit, and reasonable fit, respectively,

3. The comparative fit index (CFI), the normed fit index (NFI), the non-normed fit index (NNFI): Fit indices of .90 or above are used as indicators of adequate model fit.

All the CFA analyses were conducted using LISREL 8.80 ((Jöreskog & Sörbom, 2007). A series of CFA models were tested and compared in order to offer competing explanations of the structural relationships among the five rating scales. Structural Equation Modeling was utilized to examine the relationship between the CFA models of test performance and those of students' self-ratings.

### 3.6.3 Multi-sample analyses

Multi-sample analyses were performed with respect to gender and major fields of study based on the model of choice. Of particular interest were the mean differences in the five latent traits between different groups. To accomplish this, the group specific correlation matrices, means, and standard deviations were analyzed. It is worth mentioning that for meaningful multi-group mean trait comparisons, the scalar invariance assumption that the loadings and intercepts between groups are equal must be tested first by "forcing" the factor loadings to be equal between the two subgroups and assessing the deterioration of overall model fit. If the scalar invariance assumption is met, differences between group latent trait means will be obtained by applying the same factor loadings to both groups, fixing the latent trait means of the first group and freely estimating the latent trait means of the second group.

### 3.6.4 Item Response Theory (IRT) Analyses

*Data Recoding.* With only one rating included in the IRT analyses, there were altogether 14 items for each test candidate. Because extreme scores of 1 and 6 are rare, which could lead to very unstable parameter estimates in estimation, scores of 1 and 6 were collapsed together with the adjacent category scores whenever appropriate.

*IRT analyses.* The Samejima's graded response model was fit to the data. The slope and location parameters were inspected to investigate the problematic items. Raters' differentiations between assessment levels were examined through item trace lines. Peaks of test information curves capture the trait values at which the test differentiates students best.

*Differential Item Functioning (DIF)*. Lord (1980) observed that the trace line or item function curve is ideally suited to defining differential item functioning (DIF). The value of the trace line at each level is the conditional probability of a correct response given that level of ability of proficiency. If we are considering the possibility that an item may function differently (exhibit DIF) for some focal group relative to some reference group, then in the context of IRT we are considering whether the trace lines differ for the two groups. If the trace lines are the same, there is no DIF. If the trace lines differ, there is DIF. Because the trace line for an item is determined by the item parameters, Lord (1980) noted that the question of DIF detection could be approached by comparing estimates of the item parameters between groups. The present study is primarily concerned with procedures for DIF detection in an oral test in which the test-takers' responses are scored polytomously. DIF is used to examine whether item function curves or trace lines are the same across subgroups (males vs. females and business vs. non-business majors). Besides, the item information curves and test information curves were compared across sub-groups to investigate whether the oral test could discriminate the two groups equally well and provide the same amount of information between the two groups along the latent trait scale.

All IRT analyses were performed in IRTPRO 2.1 (Cai, du Toit, & Thissen, 2012) using Bock-Aitkin estimation method and Xpd algorithms.

### 3.6.5 Univariate and multivariate Generalizability theory (G theory) analyses

In 2012 test administration, Task 3 and Task 5 were rated by two raters while the other tasks were rated by only one rater. Besides, the same raters were assigned to Task 3 and Task 5 while the other tasks were rated by different raters from Task3 and Task5. This results in an unbalanced design p x (r: t: d) for the whole oral test in which raters are partially nested within tasks and tasks partially nested within dimensions. In a similar fashion, TFR, CoP, GV, and Pron also feature an unbalanced design-- p x (r: t) with persons crossed with raters and tasks and raters partially nested within tasks. These designs are confounded designs which do not allow us to calculate the variance components involving raters (r) and tasks (t). Hence, the confounded designs were treated as unbalanced nested designs in order to examine the dependability of TFR, CoP, GV, and Pron as well as the whole test. This generalizability study can only be an approximation of the real dependability estimates for TFR, CoP, GV, Pron and the whole test. CoFlu measured in Task 2 and Task 4 features a fully crossed design p x t with persons crossed with tasks, which enables us to calculate the exact dependability estimate of CoFlu.

D studies were conducted based on a balanced design for the whole test as well as the dimensions. The five dimensions were modeled as the fixed facet for the whole test. It was assumed that the five dimensions representing the five GLSPA speaking abilities of primary interest were not exchangeable with each other.

For this oral test, each examinee was double rated only on the dimensions measured in Task 3 and Task 5. Thus, multivariate G theory analyses were only conducted on Task 3 and Task 5 and the variance and covariance components were used to estimate the dependability of the composite scores for both tasks, which were averages or weighted averages of the analytic scores. Each task features a p x r design with person fully crossed with raters. Raters are considered as random facets. Task 3 was rated on two dimensions and Task 5 on four dimensions.

The univariate Generalizability theory (G theory) analyses were conducted with urGENOVA (Brennan, 2001a). Multivariate G theory analyses were performed with mGENOVA (Brennan, 2001b).

### 3.6.6 Raters' verbal protocols

Verbal reports from the individual raters were transcribed and double-checked by the researcher. Non-linguistic features such as pauses and laughing were excluded in the transcripts. The verbal reports were qualitatively described in Chapter 6 to address the research questions and also help better understand the quantitative findings.

# Chapter 4 Results I Confirmatory Factor Analyses

In this chapter the results of the confirmatory factor analysis (CFA) are presented. CFA was used to examine the factor structure of the GSLPA Spoken Language Test (SLT) and the relationship between test takers' perceptions of speaking ability and their actual test performance. Multi-sample CFA was conducted to address whether the score interpretations were consistent across subgroups of test takers.

## *4.1 Preliminary data analyses of the analytic scores*

In order to obtain information to help interpret the results from CFA and IRT analyses, several kinds of preliminary analyses were conducted. Descriptive statistics of the GSLPA SLT analytic scores were first calculated in order to examine the distributional characteristics of the scores and to check normality assumptions in the data. T-test was also conducted to investigate whether there were significant differences on the 14 analytic scores across subgroups of test takers (males vs. females; business vs. non-business majors). The correlations among the original ratings, averaged ratings and rounded ratings on Task 3 and Task 5 were examined in order to find out which set of ratings were more appropriate for the CFA and IRT analyses. Finally, Cronbach's alpha was calculated in order to gain a rough idea of the extent to which the items of interest were reliable indicators of the intended constructs.

### 4.1.1 Descriptive statistics

In order to examine the normality of the distributions of test scores and to check the comparability of the sub-groups (i.e., females vs. males, business vs. non-business majors), the descriptive statistics were calculated for the whole group on the one hand, and for each of the four sub-groups separately, on the other. The results for the whole group and four sub-groups are summarized in Appendix 2. These results indicate that the scores for the whole group and the four sub-groups had reasonably normal distributions since the univariate skewness and kurtosis values for all items fell within ±1.2. All the item means across all the groups were above 3.50 out of six points, suggesting that these items were relatively easy for this test population. The task fulfillment and relevance rating on Task 4 had the largest mean and standard deviation, indicating that this item was the easiest one and the most widely spread among all the items. The male group had lower means on all items than the female group, with between group score differences ranging from .13 to .27. The business group had higher means on all items compared to the non-business group. The score differences are relatively smaller, ranging from .01 to .20.

### 4.1.2 T-test for males vs. females and business vs. non-business majors

An independent samples t-test was conducted to investigate whether males and females have significant mean difference on each analytic rating. Similarly, a t-test was also used to examine the mean differences between business and non-business on the 14 analytic ratings. The t-test results are summarized in Appendices 3 and 4. As can be seen in Appendix 3, females have significantly higher mean scores than males on all the 14 analytic scores except TFR4. The t-test results between business and non-business majors (see Appendix 4) show that the business group performed significantly better than non-business students on Items TFR1, CoP1, GV1, GV2, CoFlu2, CoP3, CoP5, and GV5.

### 4.1.2 Correlations among the original, averaged and rounded ratings

Recall, from Chapter 3, that three different scores were calculated for ratings on Tasks 3 and 5 because of double ratings with Task 3 and Task 5. The intercorrelations among the original, averaged and rounded ratings on Task 3 and Task 5 are summarized in Appendix 5. It can be seen that the correlations among original ratings from two raters on both tasks were much lower compared to the correlations among their averaged and rounded ratings. The averaged ratings for all the dimensions on both tasks are highly correlated with the rounded ratings with coefficients of about .92. Given the fact that the rounded ratings are the same ordinal data type as the original ratings on tasks 1, 2, and 4, the rounded ratings were used as a basis for the CFA and IRT analyses.

### 4.1.3 Reliability analysis

Cronbach's alpha for the SLT total score—all 14 analytic scores, including the rounded ratings for Tasks 3 and 5—was .910, indicating a high internal consistency of the analytic ratings for GSLPA SLT. The Cronbach's Alpha if Item Deleted values in Table 2 suggest that all the items were reliable indicators of the test constructs.

*Table 2. Reliability estimates of the analytic scores*

| | Corrected Item-Total Correlation | Cronbach's Alpha if Item Deleted |
|---|---|---|
| T1TFR | .483 | .909 |
| T1CoP | .703 | .900 |
| T1GV | .734 | .900 |
| T1Pron | .713 | .900 |
| T2GV | .698 | .900 |
| T2CoFlu | .730 | .900 |
| T3Pron | .706 | .900 |

| | | |
|---|---|---|
| T3COP | .627 | .903 |
| T4TFR | .363 | .924 |
| T4CoFlu | .542 | .906 |
| T5TFR | .623 | .903 |
| T5COP | .676 | .901 |
| T5GVR | .751 | .899 |
| T5Pron | .690 | .901 |

## *4.2 Confirmatory factor analysis of GSLPA SLT*

### 4.2.1 Model specification

The correlation matrix (N=999) among the 14 analytic scores in this oral test is presented in Appendix 6. It can be seen that correlations vary from small to moderate. Five Confirmatory Factor Analysis (CFA) models were tested and compared in order to investigate competing explanations of the structural relationships among the five rating scales: 1) the correlated trait-uncorrelated method (CTUM) model, 2) the higher-order trait-uncorrelated method (HTUM) model, 3) the bi-factor model, 4) the correlated trait-correlated uniqueness (CTCU) model, and 5) the higher-order trait-correlated uniqueness (HTCU) model.

#### *4.2.1.1 The correlated trait-uncorrelated method (CTUM) Model*

In the CTUM model, the trait factors were correlated with each other while the method factors were uncorrelated with each other. This model, presented in Figure 1 below, depicts the multicomponential and yet correlated nature of the language ability assessed in the GSLPA SLT. The 14 rectangles in the center of Figure 1 represent the 14 observed variables, i.e., the 14 analytic ratings awarded to each candidate as all possible combinations of the five rating scales and five tasks. The five ovals to the left are for the traits of interest: task fulfillment and relevance

(TFR), clarity of presentation (CoP), grammar and vocabulary (GV), pronunciation (Pron), and confidence and fluency (CoFlu). The five ovals to the right are for the latent factors associated with the five speaking tasks. In this model, each of the 14 observed variables is specified as related to one trait factor and one method factor. However, this model does not fully represent the situation of the GSLPA SLT where a composite score is reported because a trait factor that represents the overall speaking ability is "missing" from this diagram.

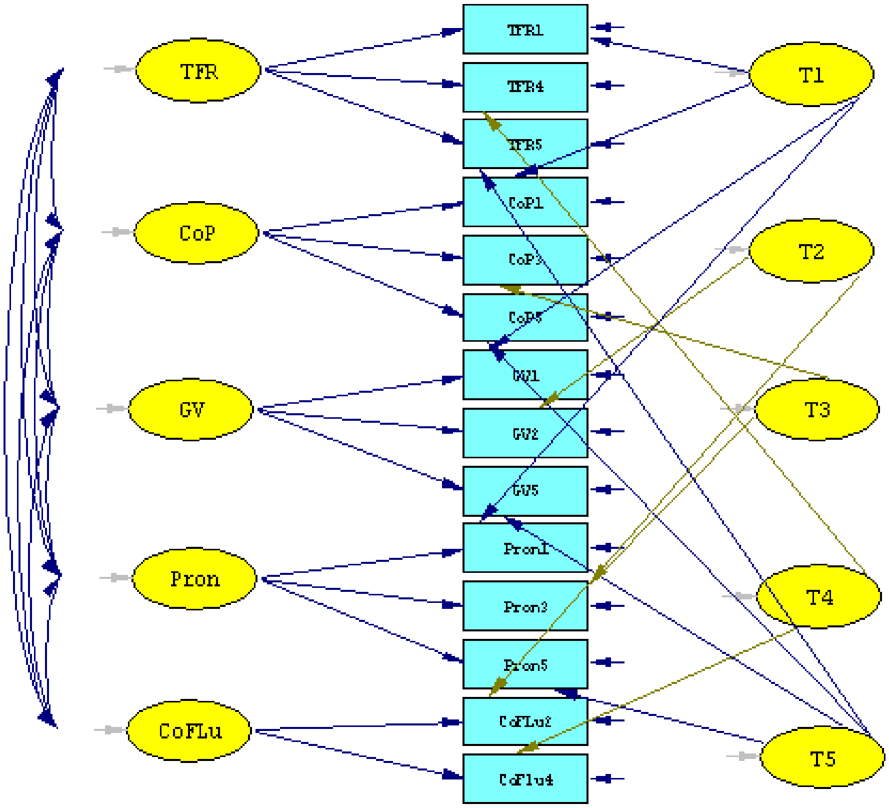

*Figure 1. CTUM model*

### *4.2.1.2 The higher-order trait-uncorrelated method (HTUM) Model*

In the HTUM model, a higher-order factor structure is imposed on the correlations among the five trait factors. This model, illustrated in Figure 2 below, specifies a higher-order factor that could account for correlations among the first-order trait factors representing the five rating scales. This trait factor structure reflects the assumption underlying the policy of reporting a single composite score.

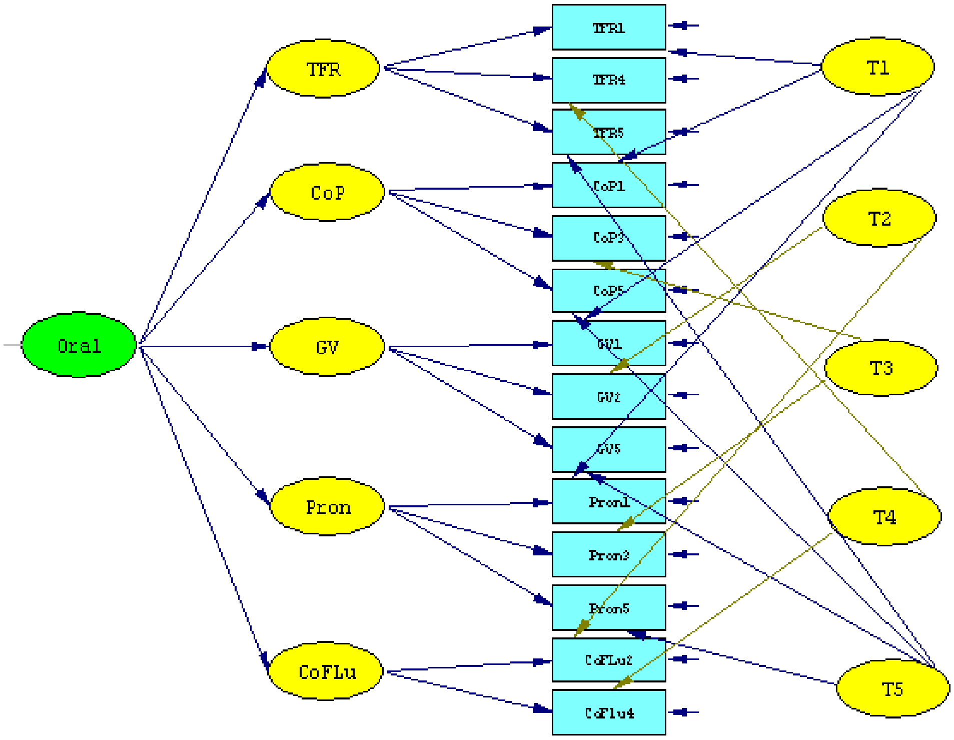

*Figure 2. HTUM model*

### *4.2.1.3 The Bi-factor Model*

The Bi-factor model specifies a single trait factor underlying the rating scales. In this model, the five traits representing the rating scales are essentially indistinguishable from one another. In order to demonstrate the multi-componential nature of the GSLPA SLT, the CTUM model must show significantly better fit than this model. Figure 3 displays the path diagram for this model.

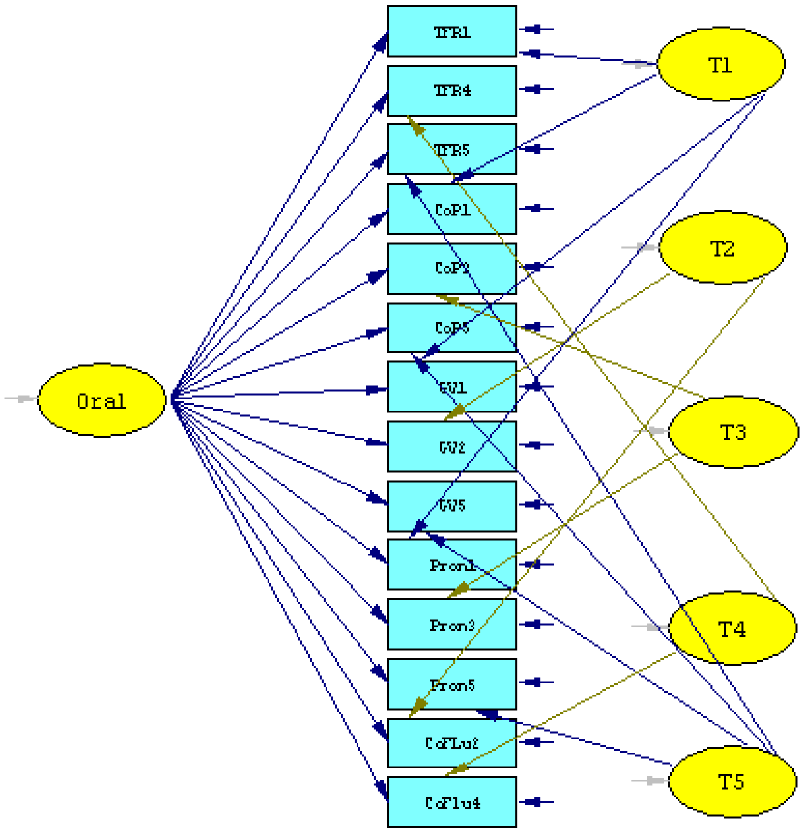

*Figure 3. Bi-factor model*

*4.2.1.4 The correlated trait-correlated uniqueness (CTCU) Model*

According to March (1989) and Kenny and Kashy (1992), iterative procedures in trait and method factor models often do not converge to a unique solution or they result in estimates that are outside the permissible range of values, for example, negative variances of the method factors or the error variables. These improper solutions might often be due to under-identified models. To overcome the problems of the CTCM model, Marsh (1989) recommended applying the correlated trait-correlated uniqueness (CTCU) model. In addition, the correlated uniqueness model provides a more convenient platform for conducting the multi-sample analysis, given the fact that the comparison of the trait factors, rather than method factors between two groups is the focus of this study.

While it presents a pragmatic solution to the methodological problems mentioned above, the correlated uniqueness model has a substantive disadvantage, in that it does not explicitly model the method factors, and thus fails to reflect the actual design of the SLT, which comprises five separate tasks. However, again, given the focus of this study on the traits that are measured by the GSLPA SLT, this interpretative weakness was seen as relatively unproblematic.

In the CTCU model, the five trait factors are correlated with each other and the correlations among the uniquenesses, or residuals, for the same task are also freely estimated. The path diagram for the CTCU model is shown in Figure 4.

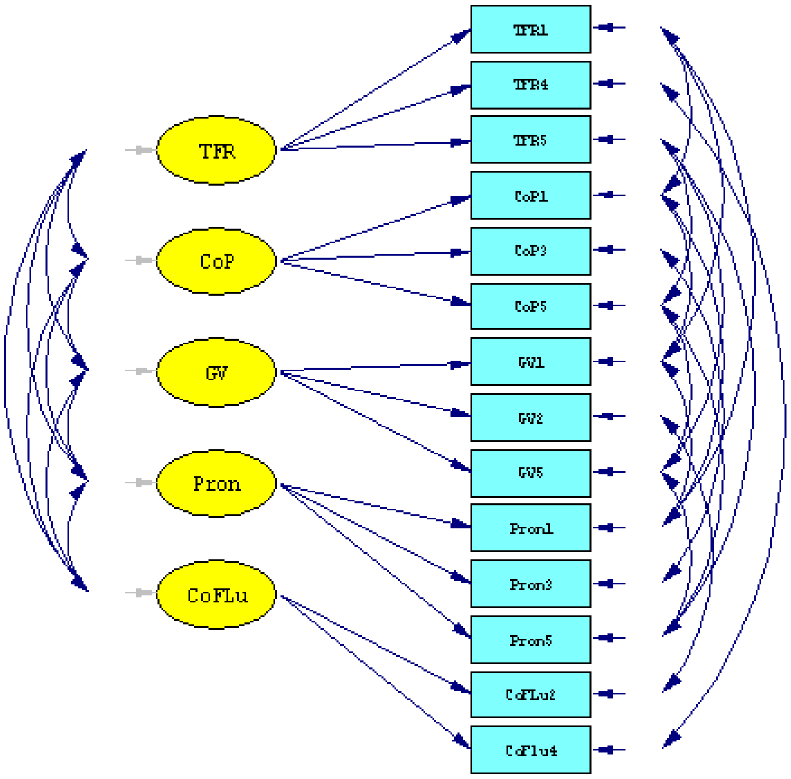

*Figure 4. CTCU model*

### *4.2.1.5 The higher-order trait correlated uniqueness (HTCU) model*

In the HTCU model, shown in Figure 5 below, a higher-order factor is included that could account for common variances among the first-order trait factors, while the uniquenesses for the same task are correlated.

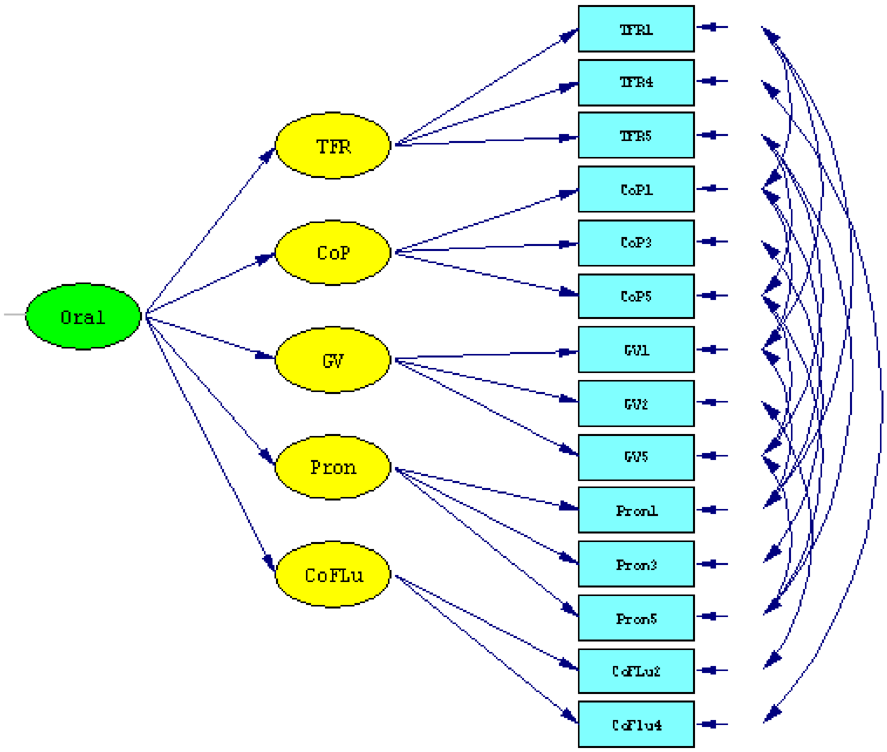

*Figure 5. HTCU model*

### 4.2.2 Model evaluation

The model fit indices for the five CFA models are summarized in Table 3.

*Table 3. Model fit indices for the five CFA models*

| Model | Description | CTUM | HTUM | Bi-factor | CTCU | HTCU |
|---|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| df | 56 | 61 | 66 | 52 | 57 |
| Minimum fit function chi-square | 117.52 | 181.83 | 796.80 | 95.23 | 141.89 |
| P value | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| RMSEA | 0.034 | 0.045 | 0.114 | 0.029 | 0.039 |
| RMSEA CI | 0.026-0.042 | 0.038-0.053 | 0.107-0.120 | 0.020-0.038 | 0.031-0.047 |
| CFI | 0.997 | 0.994 | 0.965 | 0.998 | 0.996 |
| NFI | 0.994 | 0.991 | 0.962 | 0.995 | 0.993 |
| NNFI | 0.995 | 0.991 | 0.952 | 0.996 | 0.994 |

As can be seen in Table 3, all CFA models except Bi-factor model provide acceptable fits to the test data. RMSEA values of CTCU, CTUM, HTCU, and HTUM models were all smaller than .05. The CTCU model showed the best fit to the data with a RMSEA value of .029. The obtained values of CFI, NFI, and NNFI for all the five models are higher than .95. The next section presents results of these five models from best fitting to least-well fitting: CTCU, CTUM, HTCU, HTUM, and Bi-factor model.

*The Correlated trait-correlated uniqueness (CTCU) model:* Overall, the fit of the CTCU model was excellent. With a large sample size (N=999), the minimum fit function chi-square statistic for the CTCU Model was still statistically significant (df=52; $\chi 2$ =95.23; p<0.001). RMSEA value of .029 indicated a very close fit. The standardized model parameter estimates for the CTCU model are presented in Appendix 7. Because they are adjusted for scale differences, the path coefficients are directly comparable among themselves as indicators of the strengths of relationships between the factors and the observed variables. All the trait factor loadings are significantly different from zero. The lowest one is the loading of TFR4 (.36) on TFR. The other trait loadings are relatively high ranging from .52 to .88. Moreover, the correlations among the five trait factors in the CTCU model are quite high. The lowest correlation coefficient is the one between Pron and TFR (.68). All the others are either around or above .80. The covariances between the residuals were all significantly different from zero, which provided evidence for the significant method factor loadings in the CTUM model. But all the covariances except that between TFR4 and CoFlu4 were less than .25 indicating the method effects were not strong. The unique variances and covariances among the residuals are provided in Appendix 8.

*The Correlated trait-uncorrelated method (CTUM) model:* The minimum fit function chi-square statistic for the CTUM Model was statistically significant (df=56; $\chi 2$ =117.52; p<.001), but this model nevertheless showed an excellent fit to the data. The obtained values of the CFI, NFI and NNFI met the pre-determined criteria of model fit. The RMSEA of .034 indicated a close fit of the model to the data. The results supported the distinct and yet correlated nature of the traits as defined by the five rating scales.

The standardized model parameter estimates for the CTUM model are summarized in Appendix 9. It can be easily seen that the factor loadings and the factor correlations were quite similar to the results in the CTCU model. Since tasks 2, 3 and 4 each only have two indicators, the method factor loadings of the two observed variables on them are forced to be equal. In this way, no identification problems could occur. The squared factor loading is equivalent to the covariance between these two observed variables after controlling for the trait factor. It should be noted that the trait factor loadings of the observed variables are larger than the method factor loadings except TFR4 and CoFlu4, which suggests that the test scores could be meaningfully interpreted as indicators of five dimensions although the methods also have significant effects on the test scores. TFR4 and CoFlu4 have the highest method factor loadings on Task 4 indicating that task 4 has the largest method effect on examinees' performance.

*The Higher-order trait-correlated uniqueness (HTCU) model:* As could be seen from Table 3, the HTCU model also provided a close fit to the data (RMSEA=0.039, 90% CI of RMSEA=.031-.047). The path coefficients between the five traits and the higher order trait were all above .82 and statistically significant($P<0.05$), supporting the hypothesis that the five first-order dimensions all measure a general speaking ability and the appropriateness of assigning a composite score. Appendix 10 shows the factor loadings of the observed ratings on the corresponding first-order factors as well as the regression coefficients from higher-order factor to the first-order factors. Most trait factor loadings in the HTCU model were exactly the same as the ones in the CTCU model. Other factor loadings had slight differences (e.g., 0.01). The path coefficients of the five first-order trait factors on the higher-order speaking factor were very high, ranging from .82 to .99. These above results indicated strong linear relationships between the first-order trait factors and the observed ratings, and between the higher-order factor and the first-order factors, respectively.

*The Higher-order trait-uncorrelated method (HTUM) Model:* Table 3 also suggests that the model fit of the HTUM model was satisfactory. The values of CFI, NFI and NNFI were all more than .90. The RMSEA of .045 indicated a close fit of the model to the data. Appendix 11 shows the factor loadings of trait and method factors and the regression coefficients from higher-order factor to the first-order factors. Most factor loadings of the observed ratings on the corresponding first-order trait factors in the HTUM model were exactly the same as the ones in the CTUM model. Other factor loadings had slight differences (e.g., 0.01). The factor loadings of the observed variables on the method factors also showed tiny differences between these two models. The largest difference was only .03.

*The Bi-factor Model:* Although the CFI, NFI and NNFI values were above .90, the RMSEA value of .114 suggests that this model fit was unacceptable. This could demonstrate that the language ability is multi-componential rather than unitary.

### 4.2.3 Model comparison

Since the HTUM model and bi-factor model are more restrictive and nested within the CTUM model, likelihood ratio tests (LRT) was conducted to see whether CTUM model could fit the data significantly better than the HTUM and Bi-factor models. By the same token, a likelihood ratio test was conducted to compare the model fit of the CTCU model with HTCU and CTUM models.

The likelihood ratio tests for model comparisons are shown in Table 4.

*Table 4. Likelihood ratio tests for five competing models*

| Models compared | df difference | Chi-square difference | Significance (p<.05) |
|---|---|---|---|
| CTUM VS.HTUM | 5 | 64.31 | Significant |
| CTUM VS. Bi-factor | 10 | 796.80 | Significant |
| CTCU VS. CTUM | 4 | 24.37 | Significant |
| HTCU VS. HTUM | 4 | 39.94 | Significant |

The chi-square difference test shows that the fit of the CTUM model was significantly better than that of the bi-factor model, supporting the hypothesis that the five trait factors are psychometrically distinct from one another. Regarding the comparison between the CTUM and the HTUM models, the likelihood ratio test (LRT) suggests that the CTUM model fit the data significantly better than the HTUM model. The CTCU model is also shown to improve significantly relative to the HTCU model. But RMSEA values of these two higher-order factor models still indicated a close fit to the test data. The LRT results show that the CTCU model fit the data significantly better than the CTUM model, while the HTCU model was significantly better than the HTUM model. From these results we could conclude that the CTCU model fit the data best among the five models. Besides, the CTCU model with only five trait factors makes it simpler to compare the trait factor loadings and latent trait means across subgroups of test takers. Hence, it is selected to be the baseline model for multi-sample analyses.

## *4.3 Multi-sample Confirmatory Factor Analyses*

### 4.3.1 Males vs. Females

In order to meaningfully compare the factor means between males and females, the factor structures for the two groups need to be assumed to be invariant. The factorial invariance assumption was tested, and with regard to the invariance in the factor loadings, error variances and factor correlations across samples, the fit indices were unacceptable. The minimum fit function chi-square was statistically significant (df=157; $\chi$2=217.81; p=0.0001). Hence, the factorial invariance assumption between the female and male student groups was rejected, so that meaningful latent trait mean comparisons were not possible. A graphic comparison of the factor loadings between the female and male groups is presented in Figure 6 below. Although most of the factor loadings were lower for the male group than the female group, only the factor loading of CoFlu4 for the male group was significantly different from the female group at the 0.05 level. The factor loading difference of CoFlu4 was significant yet relatively low (.16).

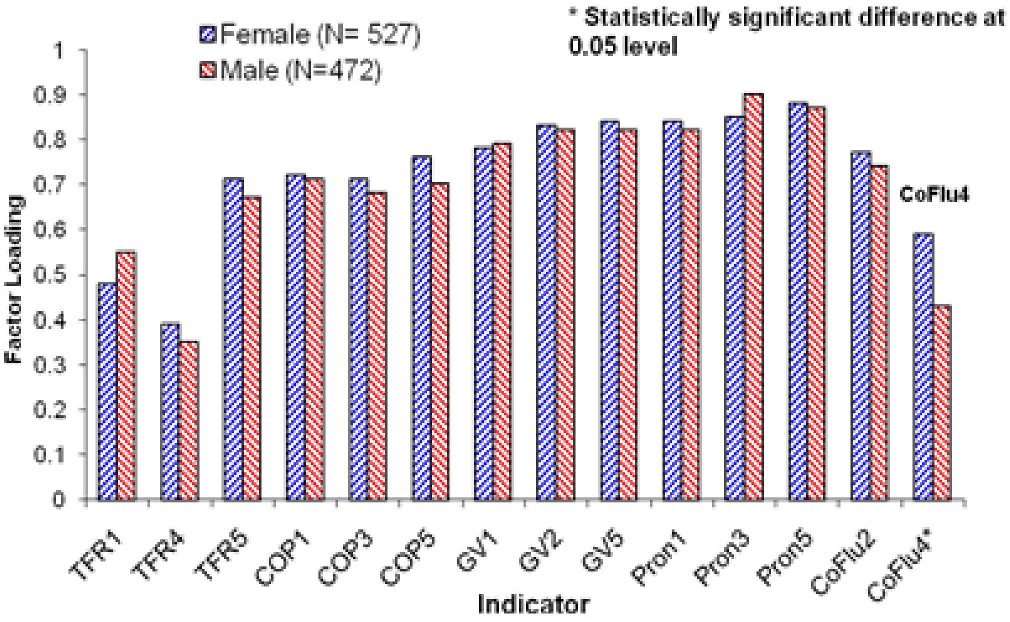

*Figure 6. Comparison of factor loadings between males and females*

4.3.2 Business vs. non-business majors

The factorial invariance assumption between the business and non-business student groups was also rejected (df=157; $\chi2$=247.05; p=0.000). Figure 7 provides a graphic display of the factor loadings of the business and non-business majors. Only two out of the 14 factor loadings were significantly different between the business and non-business groups. The factor loading of TFR1 for business group was significantly higher than the non-business group at the .05 level with a difference of .15. On the contrast, the factor loading of TFR5 for business group was significantly lower than the non-business group at the .05 level. The difference was relatively low (.10).

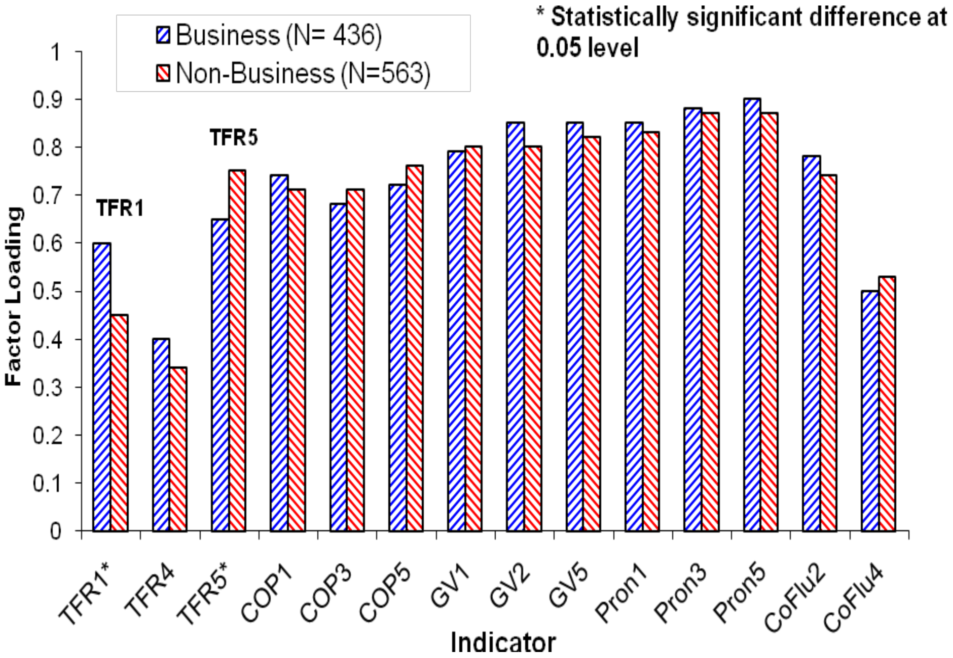

*Figure 7. Comparison of factor loadings between business and non-business majors*

### *4.4 The relationship between test takers' perceptions and their test performance*

#### 4.4.1 Preliminary analyses of the student questionnaire responses

The descriptive statistics for 14 self-rating (SR) items in the student questionnaire (Items 7 through 11 in the questionnaire) were first calculated. As shown in Table 5, the means of the 14 SR items in the questionnaire were similar to each other ranging from 3.49 to 3.65, and the standard deviations ranged from .891 to .993. All values for skewness and kurtosis fell within the acceptable range from -1 to 1 and most values were not significantly different from zero, indicating that all the variables were approximately normally distributed.

*Table 5. Descriptive statistics for 14 self-rating items from the questionnaire*

|  | N | Minimum | Maximum | Mean | Std. Deviation | Skewness | Std. Error | Kurtosis | Std. Error |
|---|---|---|---|---|---|---|---|---|---|
| TFR1S* | 349 | 1 | 6 | 3.49 | .993 | .091 | .131 | -.169 | .260 |
| TFR4S* | 349 | 1 | 6 | 3.65 | .942 | .059 | .131 | -.392 | .260 |
| TFR5S* | 349 | 1 | 6 | 3.58 | .933 | .164 | .131 | -.108 | .260 |
| CoP1S* | 349 | 1 | 6 | 3.48 | .948 | .141 | .131 | -.202 | .260 |
| CoP3S* | 349 | 2 | 6 | 3.59 | .923 | .284 | .131 | -.200 | .260 |
| CoP5S* | 349 | 1 | 6 | 3.57 | .922 | .144 | .131 | -.206 | .260 |
| GV1S* | 349 | 1 | 6 | 3.38 | .897 | .301 | .131 | .113 | .260 |
| GV2S* | 349 | 1 | 6 | 3.41 | .891 | .299 | .131 | .249 | .260 |
| GV5S* | 349 | 1 | 6 | 3.50 | .915 | .168 | .131 | .092 | .260 |
| Pron1S* | 349 | 1 | 6 | 3.65 | .961 | .209 | .131 | -.088 | .260 |
| Pron3S* | 349 | 1 | 6 | 3.62 | .932 | .250 | .131 | -.110 | .260 |
| Pron5S* | 349 | 1 | 6 | 3.59 | .926 | .312 | .131 | -.039 | .260 |
| CoFlu2S* | 349 | 1 | 6 | 3.52 | .981 | .084 | .131 | -.178 | .260 |
| CoFLu4S* | 349 | 1 | 6 | 3.63 | .955 | .193 | .131 | -.170 | .260 |

*S indicates the differences of the 14 variables in the survey from those in the oral test

The internal consistency estimate for the 14 SR items in the questionnaire was .971. Table 6 presents the corrected item-total correlations and Cronbach's alpha if deleted. The corrected item-total correlations for these 14 items were all above .74, suggesting that each item was highly correlated with the sum of the rest items and all the items seemed to measure one common construct. The statistics for Cronbach's alpha if deleted indicated that all the items contributed to the high internal consistency of the questionnaire and no problematic items were detected in this questionnaire.

*Table 6. Reliability estimates of the questionnaire*

|  | Corrected Item-Total Correlation | Cronbach's Alpha if Item Deleted |
|---|---|---|
| TFR1S | .743 | .971 |
| TFR4S | .808 | .970 |

| | | |
|---|---|---|
| TFR5S | .844 | .969 |
| CoP1S | .801 | .970 |
| CoP3S | .856 | .969 |
| CoP5S | .861 | .969 |
| GV1S | .808 | .970 |
| GV2S | .845 | .969 |
| GV5S | .856 | .969 |
| Pron1S | .801 | .970 |
| Pron3S | .851 | .969 |
| Pron5S | .857 | .969 |
| CoFlu2S | .836 | .969 |
| CoFLu4S | .841 | .969 |

4.4.2 The factor structure of the questionnaire

The correlation matrix of the 14 SR variables in questionnaire (N=349) is presented in Appendix 12. Since the CTCU model fit the SLT data best among all the CFA models, it was also fit to the self-rating data. Although the minimum fit function chi-square statistic for this model was statistically significant (df=52; $\chi 2$ =120.29; p=0.000), it still showed an acceptable fit to the questionnaire data, with an RMSEA=.058), and values for the CFI, NFI and NNFI all above .99. All the factor loadings are above .750 and the factor correlations are larger than .90. These results support the distinct and yet correlated nature of the self-ratings of speaking abilities. The standardized model parameter estimates for the CTCU model are presented in Table 7.

*Table 7. Standardized parameter estimates of the CTCU model for the self-ratings*

| | TFR | CoP | GV | Pron | CoFLu |
|---|---|---|---|---|---|
| TFR1S | 0.75* | | | | |
| TFR4S | 0.84* | | | | |
| TFR5S | 0.88* | | | | |

| | | | | |
|---|---|---|---|---|
| CoP1S | 0.78* | | | |
| CoP3S | 0.88* | | | |
| CoP5S | 0.88* | | | |
| GV1S | | 0.81* | | |
| GV2S | | 0.90* | | |
| GV5S | | 0.89* | | |
| Pron1S | | | 0.81* | |
| Pron3S | | | 0.89* | |
| Pron5S | | | 0.89* | |
| CoFlu2S | | | | 0.89* |
| CoFLu4S | | | | 0.87* |

*p<0.05

Correlations among the five traits:

| | TFRS | CoPS | GVS | PronS | CoFluS |
|---|---|---|---|---|---|
| TFRS | 1.00 | | | | |
| CoPS | 0.96* | 1.00 | | | |
| GVS | 0.92* | 0.94* | 1.00 | | |
| PronS | 0.90* | 0.93* | 0.95* | 1.00 | |
| CoFluS | 0.92* | 0.97* | 0.91* | 0.94* | 1.00 |

*p<.05

### 4.4.3 The relationship between test takers' perceptions of speaking abilities and their actual test performance

The CTCU model was also used as a baseline model to examine the correspondence between what test takers reported in their self-ratings and their actual test performance. The correlation matrix of the 14 variables in the questionnaire and the 14 items in the GSLPA SLT is shown in Appendix 13. The relationships between the trait factors in the questionnaire and the corresponding ones in the GSLPA SLT are represented in the structural equation modeling (SEM) as shown in Figure 8 below.

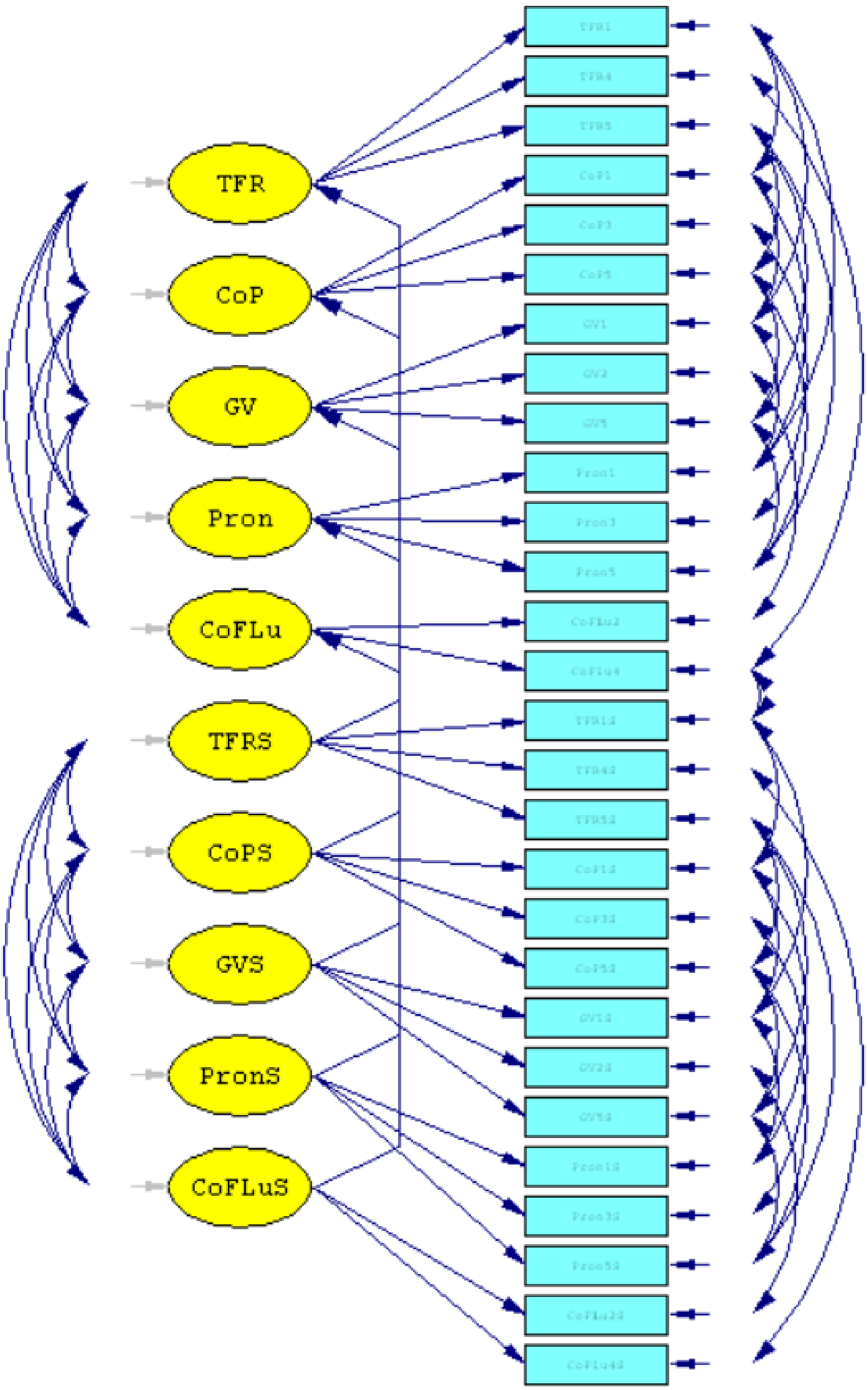

Figure 8. SEM for the relationship between the questionnaire and the GSLPA SLT

The SEM in Figure 8 is comprised of two measurement models and one structural model. In the first measurement model for GSLPA SLT, the five trait factors represent the five criteria reflected in the analytic rating scales: task fulfillment and relevance (TFR), clarity of presentation (CoP), grammar and vocabulary (GV), pronunciation (Pron), and confidence and fluency (CoFlu). These five trait factors are also correlated with each other and the correlations among the uniquenesses, or residuals, for the same task are freely estimated. The second measurement model specifies how the observed variables in the questionnaire are related to the five self-rating latent traits of interest: TFRS, CoPS, GVS, PronS, and CoFluS. The structural model shows the relationships between the five speaking traits and the corresponding traits in the self-ratings. In the structural model, the predictors are also latent factors: TFRS, CoPS, GVS, PronS, and CoFluS.

The regression coefficients from the trait factors in the questionnaire (TFRS, CoPS, GVS, PronS, and CoFluS) to those in the test (TFR, CoP, GV, Pron, and CoFlu) indicate the prediction strength. Overall, the model fit is excellent. The minimum fit function chi-square statistic is statistically significant (df=295; $\chi 2$ =385.38; p<.001). The RMSEA value of .027 indicates a very close fit to the data. The NFI, NNFi and CFI values are all above .98. The regression coefficients from trait factors in the questionnaire to those in the GSLPA SLT are reported in Table 8. It should be noted that the regression coefficients are all significantly different from zero at the .05 level, ranging from .221 to .319. This suggests a significantly low prediction effect from students' perceptions to their actual test performance. The highest one is from CoFluS to CoFlu, indicating that students can predict their Confidence and Fluency better compared to other dimensions.

*Table 8. Regression coefficients from the five trait factors in the questionnaire to the corresponding ones in the GSLPA SLT*

| Trait factors<br>in the questionnaire | Trait factors<br>in the oral test | Regression coefficients |
|---|---|---|
| TFRS | TFR | 0.221* |
| CoPS | CoP | 0.237* |
| GVS | GV | 0.234* |
| PronS | Pron | 0.234* |
| CoFluS | CoFlu | 0.319* |

* p<.05

### *4.5 Summary*

This chapter has addressed the meaningfulness and impartiality of rating-based interpretations with Confirmatory Factor Analytic approaches. The comparison of five competing CFA models indicated that the CTCU, HTCU, CTUM and HTUM models all fit the data well. Taking model parsimony and interpretability into consideration, the Higher-order trait-Uncorrelated method model was preferable since it confirmed the current multi-componential structure as reflected in the test design and provided the most parsimonious explanation of the relationships among the five dimensions and overall speaking proficiency. The multi-sample analyses showed that the strict factorial invariance assumption for males vs. females and business vs. non-business could not be supported. However, males and females differed significantly on only one factor loading. And business and non-business students showed significant measurement variance on only two factor loadings. Results also indicated a significant yet weak relationship between what test takers reported and their actual test performance.

## Chapter 5 Results II Item Response Theory Analyses

This chapter presents the results of the graded response model and Differential Item Functioning (DIF). The graded response model can be used to detect some problematic items by examining the parameter estimates and the trace lines. DIF was conducted to investigate whether there was any item bias related to group membership.

### *5.1 Graded response model*

Taking into consideration the ordinal nature of the analytic rating scales, Samejima's graded response model was fit to the data. Raters' differentiations between assessment levels were examined through item trace lines. Peaks of test information curves captured the trait values at which the test differentiated students best. The item parameter estimates from the graded response model are presented in Table 9.

*Table 9. The item parameter estimates from the graded response model*

| Label | *a* | *s.e.* | $b_1$ | *s.e.* | $b_2$ | *s.e.* | $b_3$ | *s.e.* | $b_4$ | *s.e.* |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | $^{4}$ 1.12 | 0.09 | -2.24 | 0.15 | -0.25 | 0.09 | 1.96 | 0.20 | | |
| CoP1 | $^{8}$ 2.32 | 0.16 | -2.22 | 0.11 | -0.31 | 0.07 | 1.50 | 0.15 | | |
| GV1 | $^{12}$ 3.07 | 0.24 | -2.31 | 0.11 | -0.43 | 0.06 | 1.45 | 0.14 | | |
| Pron | $^{16}$ 2.98 | 0.24 | -2.46 | 0.13 | -0.64 | 0.06 | 1.25 | 0.13 | | |
| CoFlu2 | $^{20}$ 2.19 | 0.16 | -2.38 | 0.13 | -0.50 | 0.06 | 1.25 | 0.14 | | |
| GV2 | $^{24}$ 3.03 | 0.26 | -2.54 | 0.14 | -0.55 | 0.06 | 1.34 | 0.14 | | |
| Pron3 | $^{28}$ 2.92 | 0.25 | -2.60 | 0.15 | -0.72 | 0.05 | 1.16 | 0.13 | | |
| CoP3 | $^{32}$ 1.82 | 0.15 | -3.02 | 0.21 | -0.51 | 0.06 | 1.39 | 0.15 | | |
| TFR4 | $^{37}$ 0.65 | 0.07 | -3.21 | 0.34 | -1.52 | 0.17 | 0.41 | 0.13 | 2.79 | 0.35 |
| CoFlu4 | $^{41}$ 1.15 | 0.10 | -3.84 | 0.33 | -1.06 | 0.09 | 1.33 | 0.16 | | |
| TFR5 | $^{45}$ 1.72 | 0.15 | -3.09 | 0.22 | -1.13 | 0.07 | 1.07 | 0.13 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| COP5 | 49 | 2.24 | 0.20 | -2.64 | 0.16 | -0.53 | 0.06 | 1.50 | 0.15 |
| GV5 | 53 | 3.33 | 0.37 | -2.42 | 0.14 | -0.50 | 0.05 | 1.33 | 0.14 |
| PRON5 | 57 | 2.69 | 0.26 | -2.51 | 0.15 | -0.74 | 0.05 | 1.18 | 0.13 |

*a* and *b* parameters in Table 9 dictate the shape and location of the trace lines for all the four categories in each item. In the graded response model, each item is described by one slope parameter (*a*) and the number of categories minus one location parameters ($b_j$) – one for each threshold between the response categories. The higher the slope parameters (*a*), the more narrow and peaked the trace lines, indicating that the response categories differentiate among individuals at different levels of the latent variable well. The location parameters ($b_j$) determine the location of the trace lines along the latent variable continuum. Specifically, the trace lines peak in the middle of two adjacent location parameters.

In the CFA analyses reported in Chapter 4, Task Fulfillment and Relevance on Task 4 (TFR4) proved to be a problematic item with the lowest trait loading and the highest task loading among all items. IRT analyses of the analytic scores from the 999 respondents further show that the slopes of TFR4 ranked the lowest among all the items. The slope for TFR4 was 0.65 and the slopes for all the other items were above 1.1, indicating that TFR4 did not discriminate test candidates as well as the other items.

Trace lines for all the categories represent the probability of an individual responding in a particular category conditional on the latent variable. They can help us better interpret the previous results and find out whether raters could differentiate between assessment levels within each item. Since all the other items had high slopes and discriminated candidates well, their trace lines show that raters were more likely to place candidates into higher levels as the trait level goes up. Figure 9 displays the trace line for item GV1 as an example of the good items.

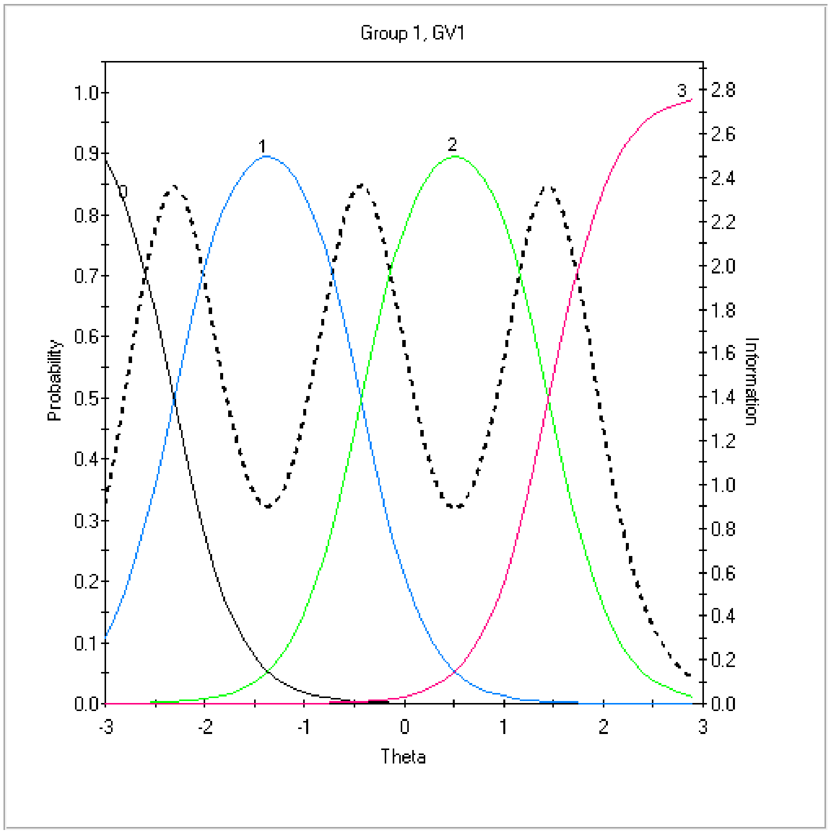

*Figure 9. Trace lines for GV1*

In contrast, the trace lines for TFR4 as shown in Figure 10 below indicate that raters did not always assign candidates with higher trait ability into higher proficiency levels. The trace lines for levels 2 and 3 (labeled G1,0 and G1,1) somewhat overlapped with each other. This means that at the same trait level, a candidate has the same chance of being awarded 2 or 3. It can be inferred that raters might have had difficulties differentiating between performance levels 2 and 3. In addition, all the trace lines for TFR4 were much flatter compared to those for GV1. Along the latent trait from -3 to +3, the probabilities of getting scores 2, 3, 4, 5, and 5 do not differ much.   Overall, the lack of discrimination among assessment levels for task fulfillment and relevance on Task 4 attests to its low slopes and trait loadings.

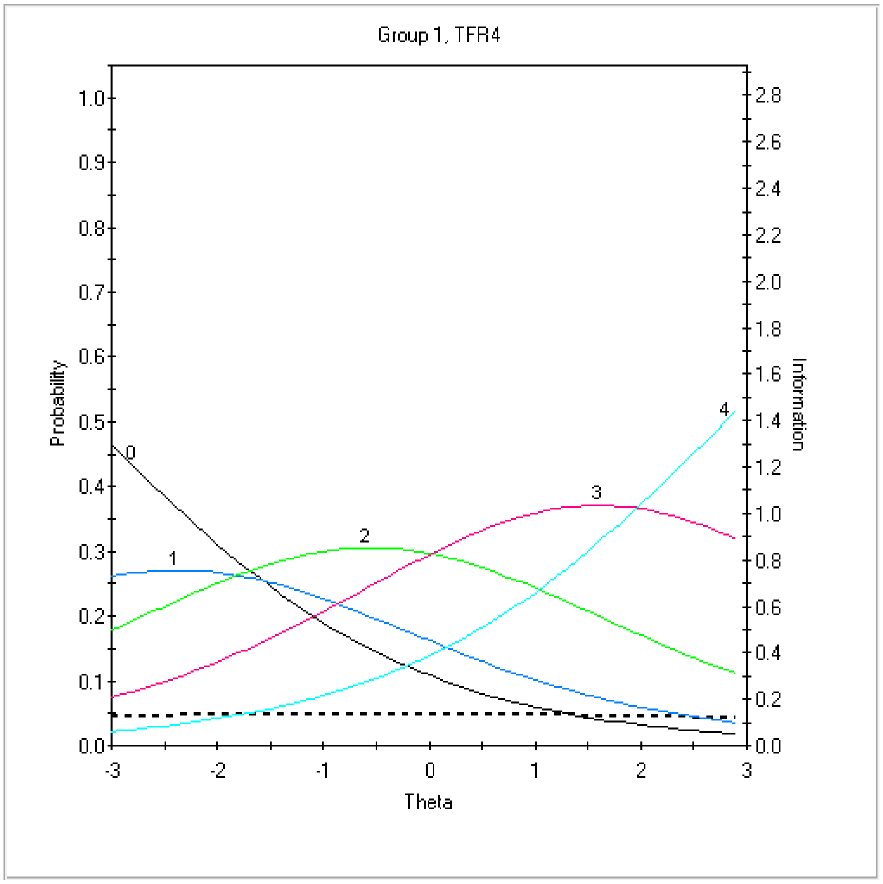

*Figure 10. Trace lines for TFR4*

Item information curves as shown in the bottom of Figures 9 and 10 further demonstrate that TFR4 provided very little information about students' latent trait levels. For all these reasons, TFR4 proved to be a weak item in measuring a candidate's true speaking ability. Both the IRT and previous CFA results provide evidence for the test coordinator's decision to discard TFR4 in calculating the composite score.

The test performance curve is presented in Figure 11 below. The solid line in the figure is the total test information curve and the dashed line represents standard errors of measurement. The test information curve peaks at trait levels of -2.3, -0.5 and 1.5, indicating that this test discriminate students best at these trait levels. The test provided the least amount of information at trait levels of higher than 2.5 and the majority of information about students at trait levels of lower than 2. These results support the decisions made based on test scores. This speaking test is mainly used to make graduation and employment decisions. Therefore, the decisions are more crucial for those students at lower trait levels.

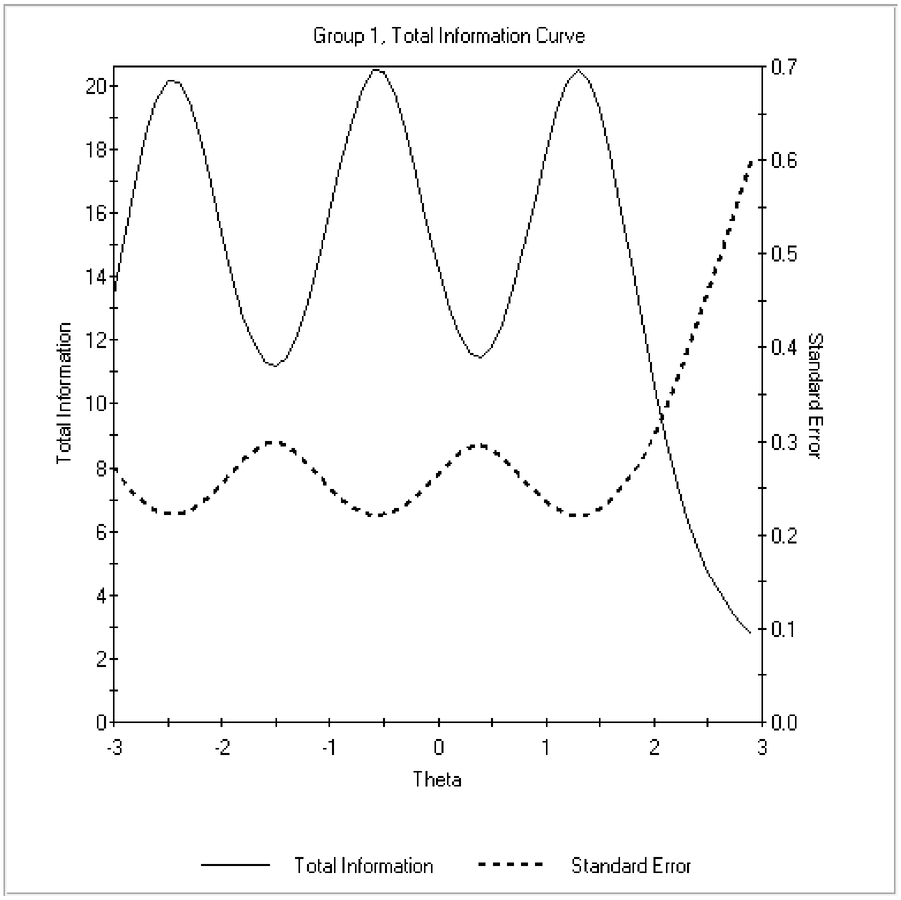

*Figure 11. Test information curve*

The variability of task difficulties was examined by looking at the location parameters of each item as shown in Table 9. All the standard errors are very small compared to the location parameters, indicating that they are all statistically significant. It is worth noting that the location parameters for the two dimensions measured by Task 4 (TFR and CoFlu) are lower than all the other items, suggesting that Task 4 is the easiest task. Interestingly, TFR on Task 1 seems to be the most difficult item. Furthermore, the location parameters for TFR5 are relatively low compared to other items. The varied difficulty levels for TFR on different tasks suggest that this dimension is heavily influenced by tasks and tends to interact with task types. For the other dimensions, the locations parameters don't display much difference for different tasks.

## *5.2 Differential Item Functioning (DIF)*

### 5.2.1 Detection of DIF between males and females

Table 10 presents the DIF statistics for all the items, comparing males and females. The $X^2$ statistics in the header row are the Wald $X^2$ test comparing parameter estimates for an item between the reference and focal groups. $X^2_a$ with a *p* value of less than .05 indicates that the slope parameters (*a*) are significantly different between the two groups and $X^2_{c|a}$ with a *p* value of less than .05 shows that the location parameters (*b*) are significantly different. The total $X^2$ can tell us whether the item is biased or not. The significant differences on parameter estimates between male and female groups are highlighted in bold. These values show that CoFlu4 was more discriminating for female group than for male group. With regard to difficulty parameters, the $b_2$ estimate suggests that it was harder for the male group to score a 4 than the female group at the same level of ability The Wald $X^2$ test shows that Items CoFlu4 and Pron5 contained DIF with p values less than .05. Table 12 also indicates that the slope parameters CoFlu4 differed significantly between male and female groups and there was a significant difference on the location parameters for Pron5 between these two groups.

*Table 10. DIF statistics for males vs. females*

| Male | Female | Total $X^2$ | *d.f.* | *p* | $X^2_{\mathrm{a}}$ | *d.f.* | *p* | $X^2_{\mathrm{c\|a}}$ | *d.f.* | *p* |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | TFR1 | 6.8 | 4 | 0.1489 | 3.2 | 1 | 0.0743 | 3.6 | 3 | 0.3122 |
| CoP1 | CoP1 | 1.6 | 4 | 0.8098 | 0.5 | 1 | 0.4870 | 1.1 | 3 | 0.7746 |
| GV1 | GV1 | 1.7 | 4 | 0.7871 | 0.9 | 1 | 0.3409 | 0.8 | 3 | 0.8467 |
| Pron | Pron | 3.6 | 4 | 0.4712 | 0.1 | 1 | 0.7353 | 3.4 | 3 | 0.3303 |
| CoFlu2 | CoFlu2 | 2.9 | 4 | 0.5708 | 0.7 | 1 | 0.4160 | 2.3 | 3 | 0.5202 |
| GV2 | GV2 | 0.4 | 4 | 0.9859 | 0.1 | 1 | 0.7506 | 0.3 | 3 | 0.9682 |
| Pron3 | Pron3 | 2.9 | 4 | 0.5782 | 0.6 | 1 | 0.4463 | 2.3 | 3 | 0.5129 |
| CoP3 | CoP3 | 1.8 | 4 | 0.7809 | 0.0 | 1 | 0.9989 | 1.8 | 3 | 0.6251 |
| TFR4 | TFR4 | 7.0 | 5 | 0.2224 | 0.7 | 1 | 0.4084 | 6.3 | 4 | 0.1766 |
| **CoFlu4** | **CoFlu4** | **13.1** | **4** | **0.0108** | **9.4** | **1** | **0.0022** | **3.7** | **3** | **0.2928** |
| TFR5 | TFR5 | 2.0 | 4 | 0.7427 | 0.0 | 1 | 0.9934 | 2.0 | 3 | 0.5804 |
| CoP5 | CoP5 | 1.8 | 4 | 0.7722 | 0.1 | 1 | 0.8124 | 1.7 | 3 | 0.6271 |
| GV5 | GV5 | 3.3 | 4 | 0.5164 | 1.2 | 1 | 0.2812 | 2.1 | 3 | 0.5539 |
| **Pron5** | **Pron5** | **11.3** | **4** | **0.0231** | **0.5** | **1** | **0.4866** | **10.8** | **3** | **0.0126** |

The item parameter estimates for these two items of both male and female groups are shown in Table 11. It can be easily seen that females had a significantly larger slope parameter (1.32) than males (.83). Regarding the location parameters, females have a larger $b_1$parameter and smaller $b_2$ and $b_3$ parameters than males, indicating that females have a higher probability of getting scores 4 and 5 at the higher latent trait levels yet a relatively smaller probability of getting scores 2 and 3 at the lower latent trait levels.

*Table 11. Item parameter estimates for CoFlu4 and Pron5*

| Group | Label | *a* | *s.e.* | $b_1$ | *s.e.* | $b_2$ | *s.e.* | $b_3$ | *s.e.* |
|---|---|---|---|---|---|---|---|---|---|
| Male | CoFlu4 | **0.83** | 0.11 | -4.56 | 0.63 | -0.98 | 0.16 | 1.97 | 0.29 |
| Female | CoFlu4 | **1.32** | 0.12 | -3.52 | 0.39 | -0.79 | 0.12 | 1.55 | 0.11 |
| Male | Pron5 | 2.63 | 0.31 | **-2.60** | 0.22 | **-0.47** | 0.07 | **1.63** | 0.15 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Female | Pron5 | 2.38 | 0.20 | **-2.24** | 0.19 | **-0.66** | 0.08 | **1.44** | 0.08 |

In order to examine the level at which the score point favors a particular group, trace lines for four categories (2, 3, 4, and 5) or score points of items CoFlu4 and Pron5, are plotted in Figures 12 and 13 below. The solid and dotted lines indicate the probability of obtaining each score point for male group and female group, respectively.

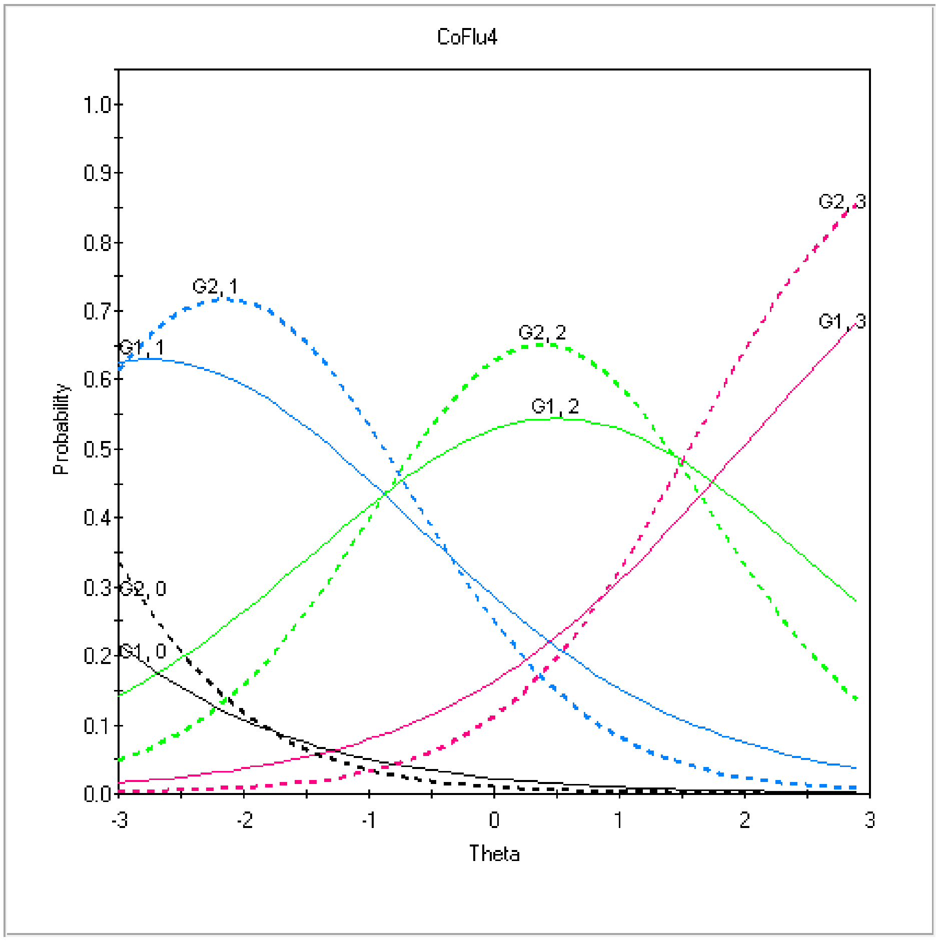

*Figure 12. Trace lines of CoFlu4 for both males and females*

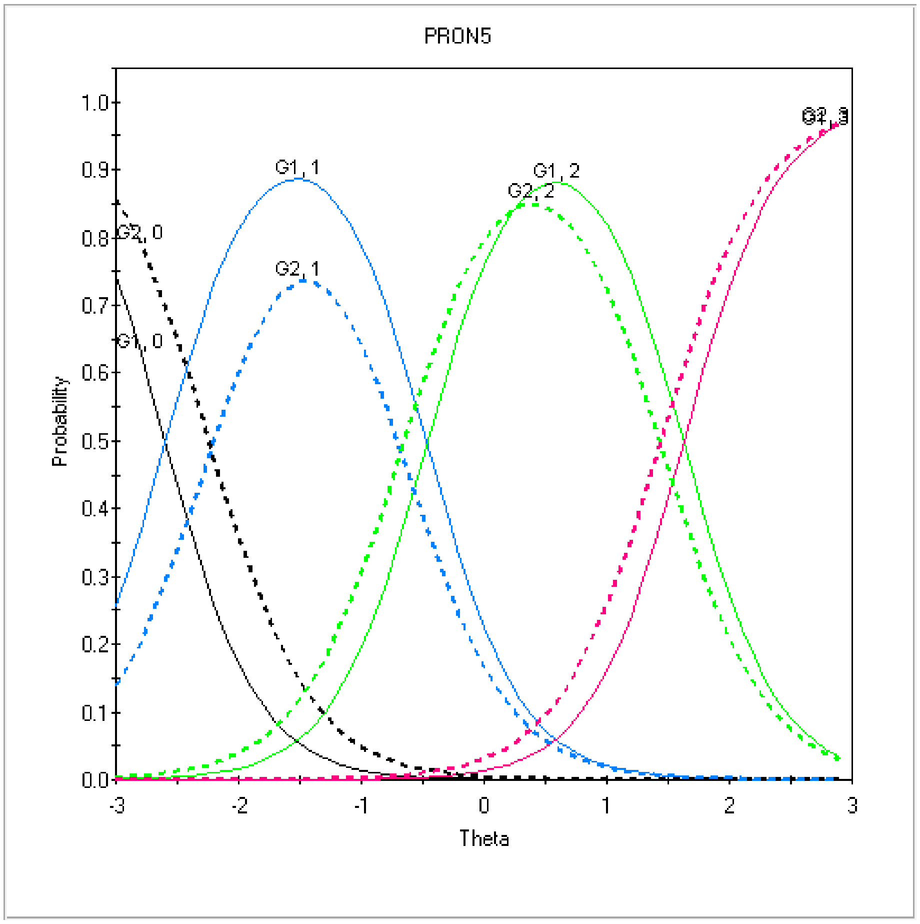

*Figure 13. Trace lines of Pron5 for both males and females*

As seen in Figures 12 and 13, the probability of obtaining the lowest score point 2 for CoFlu4 and Pron5 was high at the very low ability level and it decreased as the ability increases. Conversely, the probability of obtaining the highest score point 5 was high at the high ability level. At the ability level below –0.5, very few test-takers were expected to score a 5. The trace lines also showed differences between the two groups for these two items. For item CoFlu4 as shown in Figure 12, the trace lines for male group were flatter than the female group. At each interval between the location parameters, there was always a higher probability of scoring a 2, 3, 4 or 5 for the female group than the male group. This suggests that this item can discriminate females significantly better than males. For item Pron5, along the latent trait continuum with standard deviations larger than -.66, female group had a higher chance of being scored 4 than the male group. However, although CoFlu4 and Pron5 displayed DIF, the magnitude of DIF was quite low with a slope difference of .49 on CoFlu4 and location parameter differences of less than .40 on Pron5.

### 5.2.2 Detection of DIF between business and non-business majors

Table 12 below presents the DIF statistics for all the items, comparing business and non-business majors. The information included in this table is the same as that given in Table 10 above. DIF statistics as shown in Table 12 indicate that there was no DIF detected between business and non-business majors.

*Table 12. DIF statistics for business vs. non-business majors*

| business | non-business | Total $X^2$ | *d.f.* | *p* | $X^2_a$ | *d.f.* | *p* | $X^2_{c\|a}$ | *d.f.* | *p* |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | TFR1 | 5.9 | 4 | 0.2064 | 1.5 | 1 | 0.2159 | 4.4 | 3 | 0.2240 |
| CoP1 | CoP1 | 0.6 | 4 | 0.9682 | 0.2 | 1 | 0.6256 | 0.3 | 3 | 0.9573 |
| GV1 | GV1 | 1.8 | 4 | 0.7792 | 0.2 | 1 | 0.6572 | 1.6 | 3 | 0.6672 |
| Pron | Pron | 5.4 | 4 | 0.2471 | 0.0 | 1 | 0.9100 | 5.4 | 3 | 0.1432 |
| CoFlu2 | CoFlu2 | 1.6 | 4 | 0.8012 | 0.0 | 1 | 0.9020 | 1.6 | 3 | 0.6534 |
| GV2 | GV2 | 3.2 | 4 | 0.5213 | 1.0 | 1 | 0.3245 | 2.3 | 3 | 0.5221 |
| Pron3 | Pron3 | 2.2 | 4 | 0.6941 | 0.1 | 1 | 0.8020 | 2.2 | 3 | 0.5393 |
| CoP3 | CoP3 | 3.3 | 4 | 0.5080 | 1.0 | 1 | 0.3303 | 2.4 | 3 | 0.5017 |
| TFR4 | TFR4 | 2.5 | 5 | 0.7802 | 0.2 | 1 | 0.6613 | 2.3 | 4 | 0.6839 |
| CoFlu4 | CoFlu4 | 2.6 | 4 | 0.6329 | 0.6 | 1 | 0.4346 | 2.0 | 3 | 0.5820 |
| TFR5 | TFR5 | 5.2 | 4 | 0.2719 | 3.2 | 1 | 0.0744 | 2.0 | 3 | 0.5757 |
| CoP5 | CoP5 | 2.6 | 4 | 0.6267 | 0.2 | 1 | 0.6768 | 2.4 | 3 | 0.4888 |
| GV5 | GV5 | 2.5 | 4 | 0.6426 | 0.0 | 1 | 0.9079 | 2.5 | 3 | 0.4759 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pron5 | Pron5 | 7.2 | 4 | 0.1253 | 0.3 | 1 | 0.5708 | 6.9 | 3 | 0.0757 |

Figure 10 displays the trace lines for item T1GV with the solid line representing the business group and the dotted line for the non-business group. It can be easily seen that the trace lines for these two groups overlap with each other. At the same ability level, these two groups had quite similar chances of being scored 2, 3, 4, or 5. The trace lines for GV1 for these two groups were displayed in Figure 14 as an example of the 14 items without DIF. Obviously, the trace lines for business and non-business students almost overlap with each other.

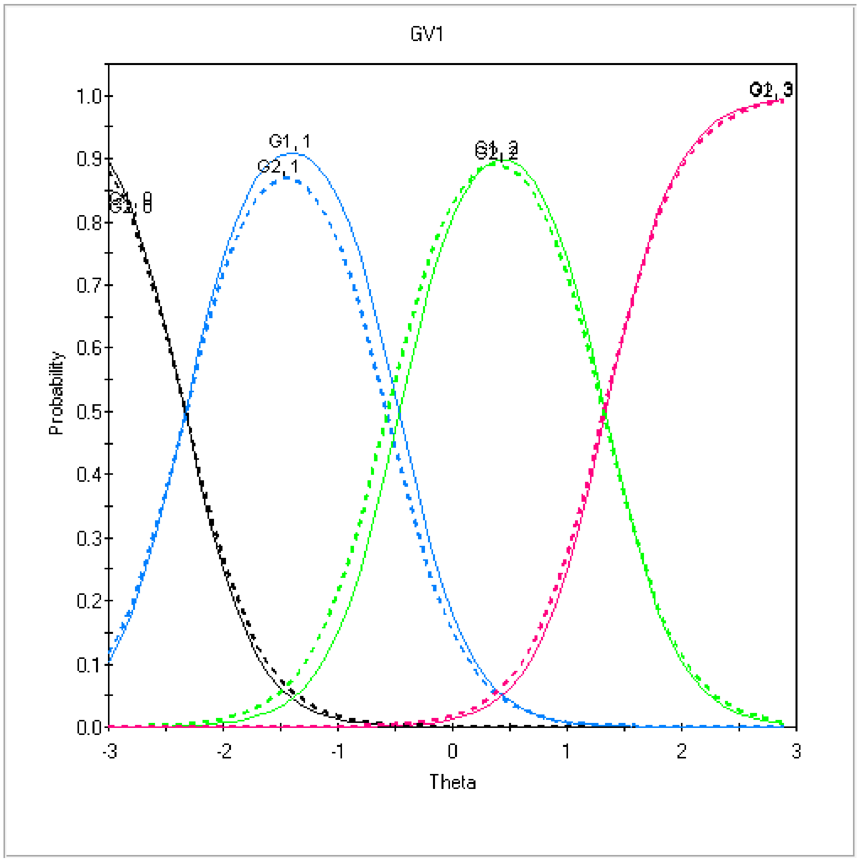

*Figure 14. Trace lines of GV1 for business and non-business students*

### 5.2.3 Group mean difference between female and male groups

Two items are found to present DIF between female and male groups. The other items without DIF presence are set as anchor items. Latent trait mean difference can be estimated based on the anchor items with the mean and standard deviation for the male group set as 1 and 0. As shown in Table 13, female group mean is significantly higher on the latent trait than male group.

*Table 13. Group means and standard deviations for males and females*

| Group | Label | $\mu$ | *s.e.* | $\sigma^2$ | *s.e.* | $\sigma$ | *s.e.* |
|---|---|---|---|---|---|---|---|
| male | G1 | 0.00 | ----- | 1.00 | ----- | 1.00 | ----- |
| female | G2 | 0.45 | 0.05 | 1.14 | 0.16 | 1.07 | 0.08 |

## *5.3 Summary*

This chapter has demonstrated the application of DIF to provide backing for the impartiality warrant of rating-based interpretations. The graded response model results confirmed that Task4 was weak in measuring students' language abilities especially Task Fulfillment and Relevance. DIF results indicated the majority of the items displayed no DIF between males and females and the magnitude of the DIF for CoFlu4 and Pron5 was very small.   Besides, all the items didn't show DIF between business and non-business students. The latent trait mean comparison between males and females suggested that on average females were more proficient in their true speaking ability than males.

## Chapter 6 Results III Generalizability Theory Analyses and Raters' Verbal Reports

The results of Generalizability theory (G theory) analyses and raters' verbal reports are reported in this chapter. First, univariate G theory analysis was conducted to examine the dependability of the five speaking dimensions as well as the whole test. Next, multivariate G theory analysis was applied to calculate the phi coefficients for Task 3 and Task 5. In addition, the disattenuated correlations between the dimensions were also calculated. Finally, raters' verbal reports were used to provide some complementary information to the quantitative results and specifically, to investigate the extent to which raters were able to differentiate the dimensions and performance bands.

### *6.1 Univariate G theory analyses on the dependability of the five dimensions and the whole test*

As mentioned earlier, the GSLPA SLT is intended to be used as a compulsory exit test for graduating students. Thus, score users such as employers would be primarily interested in using GSLPA scores for making absolute decisions as to whether students have high English ability levels for professional communication they will face in their future career. For this reason the test users need to know how dependable candidates' scores are for absolute decisions (the phi coefficient). Moreover, although rigorous processes are followed in task development, minor differences in difficulty across forms may still exist. In this case, phi coefficients are more appropriate than generalizability coefficients since test users are using students' scores that may be based on different forms as a basis for making absolute decisions.

In the 2012 test administration, Task 3 and Task 5 were rated by two raters while the other tasks were rated only by one rater. Furthermore, the same raters were assigned to Task 3 and Task 5 while the other tasks were rated by different raters from Task 3 and Task 5. This pattern of ratings resulted in an unbalanced design *p* x *(r: t: d)* for the whole oral test in which raters were partially nested within tasks and tasks partially nested within dimensions. Similar to the whole test, TFR, CoP, GV, and Pron also feature an unbalanced design-- *p* x (*r: t*) with persons crossed with raters and tasks and raters partially nested within tasks. These designs are confounded designs which do not allow us to calculate separate variance components for raters (r) and tasks (t). Hence, the confounded designs were treated as unbalanced nested designs in order to examine the dependability of TFR, CoP, GV, and Pron as well as the whole test. This generalizability study can only be an approximation of the real dependability estimates for TFR, CoP, GV, Pron and the whole test. CoFlu measured in Task 2 and Task 4 features a fully crossed design *p* x *t* with persons crossed with tasks.

D studies were conducted based on a balanced design for the whole test as well as the dimensions. The five dimensions were modeled as the fixed facet for the whole test. It was assumed that the five dimensions representing the five GLSPA speaking abilities of primary interest were not exchangeable with each other. Table 14 shows the G study variance components and the percentages of the total variance accounted for by each source of variation.

*Table 14. G study variance components for the dimensions and the whole test*

| Source of variation | Variance component | | | | | | Percent of total variation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | TFR | CoP | GV | Pron | CoFlu | Overall | TFR | CoP | GV | Pron | CoFlu |
| *p* | 0.234 | 0.139 | 0.226 | 0.282 | 0.298 | 0.255 | 38.82% | 14.41% | 34.23% | 46.25% | 44.29% | 39.25% |
| *t* | 0.017 | 0.073 | 0.004 | 0.001 | 0.001 | 0.014 | 2.86% | 7.61% | 0.57% | 0.21% | 0.11% | 2.11% |
| *r:t* | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | | 0.00% | 0.00% | 0.10% | 0.04% | 0.04% | |
| *pt* | 0.031 | 0.396 | 0.063 | 0.064 | 0.088 | 0.382 | 5.17% | 41.04% | 9.55% | 10.45% | 13.02% | 58.64% |
| *pr:t,e* | 0.320 | 0.357 | 0.366 | 0.262 | 0.286 | | 53.15% | 36.97% | 55.55% | 43.04% | 42.54% | |

As can be seen in Table 14, examinees showed the greatest variations in GV and Pron and are least variable in their TFR as indicated by the variance components associated with persons on these dimensions. This suggests that examines varied considerably in terms of their ability on GV and Pron while they did not differ much in their ability on TFR. The largest proportion of the total variance in the scores on GV and Pron could be explained by examinees' universe score differences, suggesting that examinees' GV and Pron scores are the most reliable. For TFR, the largest proportion of the total score variance was accounted for by the person-by-task interaction effect (*pt*). For all dimensions except TFR, as well as the whole test, the largest source of error was the *pr:t* interaction effect and other random errors. As mentioned above, the person-by-task interaction (*pt*) was the largest source of error for TFR, indicating that examinees were rank ordered very differently on TFR across the tasks. The task (*t*) effect was also relatively large for TFR compared to the other dimensions, which suggests that considerably tasks were at different levels of difficulty for TFR. For other dimensions, examinees may not differ much across the tasks. Given that TFR is the most task-specific dimension in the analytic rating scales, it could be expected that task would have a great effect on examinee's performance and that examinees' TFR scores would be rank ordered very differently across the tasks.

Due to the nested design for the speaking tasks, the independent variance components involving raters such as *r* and *pr* could not be estimated. The *r:t* effects for all the dimensions and the whole test were nearly zero, indicating that raters did not differ much in their leniency or harshness or in judging where an examinee stand compared to other students within each task.

Using the variance components in the G study, D studies were conducted in which the number of raters and tasks were varied to examine their impact on the phi coefficients of the analytic scores.

In the D studies for the whole test with the dimension effect fixed, the *d* and *pd* effects were not present when calculating the phi coefficients. Table 15 provides the phi coefficients for the five dimensions and the whole test when different combinations of numbers of raters and tasks are used for a balanced *p* x (*r: t*) design.

*Table 15.D studies for the five dimensions and the whole test*

| Alternative D studies for *p* x (*r: t*) design | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phi Coefficients | No of tasks | One rater | | | | | | Two raters | | | | | |
| | | Overall | TFR | CoP | GV | Pron | CoFlu | Overall | TFR | CoP | GV | Pron | CoFlu |
| | 1 | 0.39 | 0.14 | 0.34 | 0.46 | 0.44 | 0.39 | 0.53 | 0.18 | 0.47 | 0.59 | 0.56 | |
| | 2 | 0.56 | 0.25 | 0.51 | 0.63 | 0.61 | 0.56 | 0.69 | 0.30 | 0.64 | 0.74 | 0.72 | |
| | 3 | 0.66 | 0.34 | 0.61 | 0.72 | 0.70 | 0.66 | 0.77 | 0.39 | 0.73 | 0.81 | 0.79 | |
| | 4 | 0.72 | 0.40 | 0.68 | 0.77 | 0.76 | 0.72 | 0.82 | 0.46 | 0.78 | 0.85 | 0.84 | |
| | 5 | 0.76 | 0.46 | 0.72 | 0.81 | 0.80 | 0.76 | 0.85 | 0.52 | 0.82 | 0.88 | 0.87 | |

One obvious observation is that the phi coefficients increase when more raters and more tasks are used. When one rating is obtained for each response, using more tasks yields much higher phi coefficients for all the dimensions and the whole test. The impact of increasing the number of ratings per response from one to two is large for CoP, GV, and Pron, and is relatively small for TFR. As shown in Table 15, the phi coefficients were .39, .73, .81, and .79 for TFR, CoP, GV, and Pron scores respectively when 3 tasks and 2 raters were used. The phi coefficient was .56 for CoFlu when 1 rater and 2 tasks were used. The dependability estimates for TFR and CoFlu would be considered as relatively low for high-stakes decisions. The phi coefficient for the whole test is .76 when one rater and five tasks are used. It would increase up to .85 if two raters and five tasks are used. Taking into consideration the unbalanced design of this whole test, we might conclude that the phi coefficient for GSLPA SLT fell between .76 and .85. The results of the D studies offer us useful information about optimizing assessment designs. On the one hand, the dependability and validity of the assessment need to be assured. On the other hand, cost for test development and scoring and efficiency of an assessment need to be factored in when designing an assessment. It is worth noting that the phi coefficients for CoP, GV and Pron with a combination of two raters and three tasks are expected to be higher than when three tasks and one rater are used. Two raters and three tasks are recommended as an optimal design for these dimensions. The dependability estimate of .85 when two raters and five tasks are used is also acceptable for large-scale computer-based oral assessments. Controlling for the number of the raters, more tasks are required for TFR to produce more dependable scores than the other four dimensions. The reasons for the low dependability of TFR will be discussed in Chapter 7.1.

## *6.2 Multivariate G theory analyses on the dependability of composite scores for Task 3 and Task 5*

### 6.2.1 The dependability of composite scores for Task 3 and Task 5

For this oral test, each examinee was double rated on only the dimensions measured in Task 3 and Task 5, that is, CoP and Pron for Task 3 and TFR, CoP, GV, and Pron for Task 5. Thus, multivariate G theory analyses were only conducted on Task 3 and Task 5 and the variance and covariance components were used to estimate the dependability of the composite scores for both tasks, which were averages or weighted averages of the analytic scores. Each task features a p x r design with person fully crossed with raters. Raters were considered as random facets. Task 3 was rated on two dimensions and Task 5 on four dimensions.

The variance-covariance matrices for Task 3 and Task 5 are shown in Table 16 and Table 17, respectively. In these tables, the diagonal values are the variance components for examinees' universe scores and the error components on the dimensions. The lower diagonal elements are the covariance components which provide some additional information about how examinees' universe scores on the dimensions for each task covary and also show how different error components associated with those dimension scores covary. The upper diagonal values for persons are the correlations among the universe scores on the dimensions for each task, which will be discussed in detail in the section on the distinctness of the dimensions.

*Table 16. Estimated G study variance and covariance components for Task 3*

| Effect | CoP | Pron |
|---|---|---|
| p | 0.14582 | **0.79408** |
|  | 0.13770 | 0.20622 |
| r | 0.00153 |  |
|  | 0.00012 | -0.00025 |
| pr | 0.39587 |  |
|  | 0.12901 | 0.28354 |

Note. Lower diagonal elements are covariances.

Upper diagonal elements are correlations.

*Table 17. Estimated G study variance and covariance components for Task 5*

| Effect | TFR | CoP | GV | Pron |
|---|---|---|---|---|
| p | 0.17317 | **1.03344** | **0.87188** | **0.87668** |
|  | 0.17704 | 0.16947 | **0.99184** | **0.90726** |
|  | 0.16892 | 0.19010 | 0.21677 | **0.93352** |
|  | 0.16784 | 0.17183 | 0.19996 | 0.21166 |
| r | -0.00032 |  |  |  |
|  | -0.00012 | -0.00028 |  |  |
|  | -0.00013 | -0.00015 | -0.00026 |  |
|  | -0.00011 | -0.00015 | -0.00009 | -0.00025 |

| | | | | |
|---|---|---|---|---|
| pr | 0.35668 | | | |
| | 0.16779 | 0.33711 | | |
| | 0.12125 | 0.13028 | 0.26202 | |
| | 0.07168 | 0.09925 | 0.10520 | 0.28904 |

Note. Lower diagonal elements are covariances.

Upper diagonal elements are correlations.

If analytic scores are available for each task, it is possible to generate a composite score, which can be an average or a weighted average of the analytic scores on a given task. The composite score indicates the overall quality of an examinee’s performance on one task. However, the dependability of the composite score is affected by the weights of the dimensions. It is a common practice to give equal weights to all the dimensions. However, some dimensions may outweigh the others for substantive reasons. One can explore the impact of using different weighting schemes on the dependability of composite scores and determine which scheme can maximize the dependability.

In this study, three weighting schemes were compared for Task 3 in terms of the dependability of the resulting composite scores. The first weighting scheme applied equal weights to the two dimensions. The second one gave more weight (.60) for CoP compared to Pron (.40) and the third one assigned more weight to Pron (.60) and less weight to CoP (.40). By the same token, five weighting schemes were compared for Task 5. The first one also applied equal weights to all four dimensions. The other four gave more weight to different dimensions. In the subsequent D studies, different weighting schemes were applied to the dimensions, and phi coefficients of the composite scores were estimated for single rating versus double ratings. The phi coefficients of the composite scores when different weighting schemes are used are shown in Table 18 for Task 3 and Table 19 for Task 5.

*Table 18. Changes in composite score phi coefficients for Task 3 with different weights*

| Number of raters | 1 rater | | | 2 raters | | |
|---|---|---|---|---|---|---|
| Weighting scheme | CoP(.40)<br>Pron(.60) | CoP(.50)<br>Pron(.50) | CoP(.60)<br>Pron(.40) | CoP(.40)<br>Pron(.60) | CoP(.50)<br>Pron(.50) | CoP(.60)<br>Pron(.40) |
| phi coefficients | 0.41826 | 0.40050 | 0.37708 | 0.58982 | 0.57194 | 0.54766 |

*Table 19. Changes in composite score phi coefficients for Task 5 with different weights*

| Number of raters | 1 rater | | | | | 2 raters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Weighting scheme | TFR(.20)<br>CoP(.20)<br>GV(.20)<br>Pron(.40) | TFR(.20)<br>CoP(.20)<br>GV(.40)<br>Pron(.20) | TFR(.20)<br>CoP(.40)<br>GV(.20)<br>Pron(.20) | TFR(.40)<br>CoP(.20)<br>GV(.20)<br>Pron(.20) | TFR(.25)<br>CoP(.25)<br>GV(.25)<br>Pron(.25) | TFR(.20)<br>CoP(.20)<br>GV(.20)<br>Pron(.40) | TFR(.20)<br>CoP(.20)<br>GV(.40)<br>Pron(.20) | TFR(.20)<br>CoP(.40)<br>GV(.20)<br>Pron(.20) | TFR(.40)<br>CoP(.20)<br>GV(.20)<br>Pron(.20) | TFR(.25)<br>CoP(.25)<br>GV(.25)<br>Pron(.25) |
| phi coefficients | 0.53359 | 0.53160 | 0.50388 | 0.50250 | 0.52593 | 0.69587 | 0.69418 | 0.67010 | 0.66889 | 0.68932 |

It can be easily seen that the coefficients increased considerably from one rating to double ratings across all the weighting schemes, suggesting that double ratings may be necessary for all the tasks. The differences among the weighting schemes were also relatively large within both the single and double ratings. For Task 3, with equal weights for CoP and Pron, the phi coefficient was .57 when double ratings were used. Giving more weight to Pron seems more preferable in terms of maximizing the dependability of the composite score. For Task 5, the phi coefficient was .69 when double ratings are used and equal weights were applied to all the dimensions. Similarly, giving more weight to Pron can also maximize the dependability of the composite score.

### 6.2.2 Correlations among the dimensions in Task 3 and Task 5

The covariance components for persons (*p*) indicate how persons' universe scores on the dimensions covary with one another. High covariance components among the dimension scores relative to the variance components indicate that examinees who have high GV scores tend to have high scores on TFR, CoP, or Pron. Disattenuated correlations among the dimensions were estimated based on the covariance components for persons (*p*) and the variance components for persons (*p*) in the multivariate G theory framework. Since CoFlu was not measured in Task 3 and Task 5, only the correlations among TFR, CoP, GV, and Pron were examined here.

Table 20 compares the observed and disattenuated correlations among the dimensions for Task 3 and Task 5. Examining the disattenuated correlations can answer the question, "If the measurement had been perfectly reliable, what correlations would be seen?" They thus indicate the degree to which two sets of measurements have low observed correlations because of measurement error or because they are really uncorrelated. As is shown in the table, the observed correlation between CoP and Pron in Task 3 was moderate (.585). After correction for score unreliability, the disattenuated correlation became relatively large (.794). For Task 5, the observed correlations among the dimensions ranged from moderate to high. After correcting for score unreliability, closer relationships among the four dimensions are observed, as indicated by the disattenuated correlations among them. The disattenuated correlation between TFR and CoP is the highest (1.033). Interestingly, the correlations between TFR and the other two dimensions rank the lowest with coefficients of .872 and .877, respectively. These results will be discussed in Chapter 7.1 to address the separability of these dimensions.

*Table 20. Observed and disattenuated correlations among the dimensions in Task 3 and Task 5*

| | CoP vs Pron | | TFR vs. CoP | | TFR vs GV | | TFR vs Pron | | CoP vs GV | | GV vs Pron | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Obs | Dis | Obs | Dis | Obs | Dis | Obs | Dis | Obs | Dis | Obs | Dis |
| Task 3 | .585 | .794 | | | | | | | | | | |
| Task 5 | .638 | .907 | .757 | 1.033 | .657 | .872 | .576 | .877 | .744 | .992 | .718 | .933 |

Note: Obs=Observed correlations; Dis=Disattenuated correlations

### *6.3 Raters' verbal reports*

The primary purpose of collecting verbal protocols from the raters was to investigate raters' rating processes when they awarded scores to test takers. Of particular interest to the researcher was to determine how well raters could differentiate among the rating criteria and the performance bands. Hence, while the researcher listened to the recordings with the raters, she also asked questions such as 'Can you differentiate between these criteria?' The questions aimed to help us have a better understanding of the rating behaviors.

The excerpts given below were selected from the whole transcription and provide complementary information to the quantitative results and further reveal how well the raters can differentiate the rating criteria and the performance bands. The transcriptions were organized by the researcher's questions along with the raters' identification (ID) code. 'ME' in the ID code indicates that the rater is a male English speaker. 'FE' means that the rater is a female English speaker. Similarly, 'MN' specifies that the participant is a male non-Native English speaker. Since the verbal reports represent exactly the raters' words, some grammatical errors may be present in the data.

When asked about their perceptions of the overlap among the five dimensions, most of the raters indicated that it was easy to differentiate among these. One rater mentioned that she may tend to give similar scores to all the dimensions. However, she also reported that when she noticed this tendency she would correct it. Another rater indicated that there was considerable overlap between grammar and vocabulary and pronunciation. In answering the reasons for this, he pointed out that these two dimensions always went together and students usually possessed both good grammar and pronunciation. Another also mentioned that task fulfillment and relevance was quite similar to clarity of presentation because students could not fulfill the task well without a clear presentation. Overall, the raters could differentiate the dimensions easily based on the ratings scales or rating matrix. In addition, they paid great attention to the sub-features specific to each dimension as listed in the rating scale. In general, the raters concluded that there were close relationships among some dimensions. For example, students with high scores on CoP tended to do well in TFR.

*Are those criteria, e.g., task fulfillment and relevance, grammar and pronunciation, easy to differentiate?*

> "Yes. I would look at key words, for example, clarity of presentation, the difference between 3 and 4 is how hard it is to follow at times; more confused I would give 3, 4 would only have minor problems. For clarity of presentation, between 3 and 4, is more difficult to differentiate. Some problems I would give 4, more problems I would give 3." (FE1)

> "In terms of grammar, basic grammar was a 3, pretty good grammar/vocabulary was a 4. If it is 2, they only followed the script, as I mentioned before." (FE2)

> "I think for pronunciation, if they were extremely slow in speech, I gave a 3. If they can produce more fluent structures in speech, I tended to give a 4." (ME1)

> "Differentiating between criteria is easy for me, since I have been part of IELTS[1] for so long, so no problem for me." (FE2)

*Do you have an attempt to award similar scores to each criterion on the same task?*

[1] IELTS is the International English Language Testing Service (www.ielts.org), which includes an oral assessment that is rated by trained raters.

> "There are times where I notice it, and there are times that I don't. I notice I would more likely give similar scores all the way down, I thought that grammar/vocab/accuracy/pronunciation kind of go together. It is very rare that a student has great pronunciation but very poor grammar." (FE1)

> "When I started paying attention to grammar, I would compare to pronunciation, then I think they deserve the same scores. They tended to be the same. But I did try to pay attention to that." (FE2)

> "Task fulfillment and relevance, sometimes, may go with clarity and presentation as well. Because if it's not very clear, then they didn't fulfill the task. You may have to think which score you are giving, I had to make sure I wasn't double marking. Sometimes I have to differentiate between the two. But I noticed for task fulfillment, I paid attention to them fulfilling the task." (FE1)

When raters were asked about the difficulties in giving scores to each dimension, two of them pointed that bands 3 and 4 were sometimes hard to differentiate especially for Clarity of Presentation. Conversely, one rater mentioned that the low end (1) and high end (6) were hard to give and that the middle values (3 and 4) were easy to differentiate. But raters did report that there were differences on the dimensions between 3 and 4.

*Do you have any difficulties in giving scores 3 or 4, or 2 and 3? Do you have difficulties differentiating between those levels?*

> "No, it's difficult to explain. When I hear the candidate, a 3 for me is when it is very difficult for them to express what they want to say. A 4 for me is they are able to express what they

say, but is not very clear all the time. So just thinking a 4 student is generally easier to listen to and easier to follow. A 3 student I have to follow more carefully." (ME1)

"I think I rarely gave 2. In terms of 2, I don't think they said very much, and whatever they said was only following the script, not able to add any more of their own vocabulary, and the grammar was basic, and the pronunciation was extremely tentative. But very rarely will I give a 2." (FE2)

"2 and 3 are easy. 3, 4 are quite hard to differentiate, because they tend to be a borderline. If candidate's performance is generally higher than another's, I will shift the curve up for them." (MN)

*Can you differentiate between 3 and 4 on these criteria?*

"You have to work out where the middle ground is? I think the average tended to be 3 or 4. If they are on the better side of average, I gave a 4. 6 was the difficult score to give to anyone. I didn't give any 1s. I think it's pretty standard, the high and the low is hard to discriminate, but the middle values were pretty easy to differentiate." (FE2)

"I would look at key words, for example, clarity of presentation, the difference between 3 and 4 is how hard it is to follow at times; more confused I would give 3, 4 would only have minor problems. For clarity of presentation, between 3 and 4, is more difficult to differentiate. Some problems I would give 4, more problems I would give 3." (FE1)

Overall, raters' verbal reports show that they had little difficulties differentiating between the performance levels. In terms of the scoring criteria, they could also easily tell the differences among the five dimensions but they thought some criteria such as GV and Pron had very close associations with each other.

### *6.4 Summary*

This chapter presented the results of univariate and multivariate G theory analyses as well as the raters' verbal reports. Univariate G theory analyses showed that the phi coefficient for TFR and CoFlu were relatively low compared to the other dimensions. The large person-by-task interaction effect and relatively large task effect for TFR indicated that this dimension was heavily influenced by task types and examinees were rank ordered differently on TFR across Task 1, Task 4, and Task 5. Multivariate G theory analysis indicated that there were very high correlations among the dimensions. Raters' verbal reports revealed that raters could discriminate the criteria and performance level well although they thought some criteria had closer relationships.

# Chapter 7 Discussions and Implications

## *7.1 Discussion of research question 1*

The first research question addressed consistency of assessment records in the GSLPA SLT: *Are the assessment records consistent across different assessment tasks?* In this section, each sub-question of the research question 1 will be discussed based on the results reported in Chapters 5 and 6. The results of Generalizability Theory (G theory) analyses, Item Response Theory (IRT) Analyses, and the raters' verbal reports will be referenced either to address each sub-question or to explain possible explanations for the results.

### 7.1.1 Sub-question 1:

*To what extent are the GSLPA Spoken Language Test and the individual speaking tasks dependable?*

The first sub-question addressed the dependability of the whole SLT as well as of the individual speaking tasks. Since only Task 3 and Task 5 were double rated, it was impossible to examine the dependability of the other three tasks with single ratings. In order to answer this sub-question, univariate G theory analyses with an unbalanced design $p$ x $(r: t)$ for the dimensions as well as the whole speaking test and multivariate G theory analysis for Task 3 and Task 5 were conducted.

The results showed that GV and Pron proved to be the most reliable dimensions with over 40% of variances accounted for by variance components associated with persons. The person-by-task interaction (*pt*) was the largest source of error for TFR, indicating that examinees were rank ordered very differently on TFR across the tasks. Task (*t*) was relatively large for examinees’ TFR scores, which means that Task 1, Task 4 and Task 5 had different difficulty levels for examinees’ TFR. For all the other four dimensions, the largest source of error is the pr:t interaction and other random errors.

With regard to the score dependability, the phi coefficients of the analytic scores for one task with a single rating were quite low. The impact of increasing the number of ratings per response from one to two was large for CoP, GV, and Pron and was relatively small for TFR.

When the analytic scores were averaged across three tasks, there were double ratings for each task, the phi coefficients for CoP, GV, and Pron rose to relatively high levels (.73–.81) while the phi coefficient for TFR was still quite low (.39). With a single rating and 2 tasks, the phi coefficient for CoFlu was not high (.56). The phi coefficient for the whole test fell between .76 with 1 rater and five tasks, and to .85 when two raters and five tasks were applied. However, since tasks are partially nested within dimensions and raters also partially nested within tasks, this generalizability study can only be an approximation of the actual test design and these results should be interpreted with caution.

In the language assessment field, the reported findings were inconsistent with regard to the magnitude of the person-by-task interaction. This is mainly due to differences in the characteristics of the tasks and the scoring criteria used. If an assessment includes tasks that are not very differentiated in task types and uses scoring criteria that are relatively stable across tasks, it is less likely to see variation in performance across tasks. However, in those studies that reported large person-by-task interaction (e.g., Brennan et al., 1995; Lee, 2005; Lee & Kantor, 2005), the tasks were richly contextualized and/or the scoring rubrics contained features that are more task-specific, thus the quality of these features is more likely to vary across tasks.

In this study, the large person-by-task interaction and relatively large task effects for TFR are probably attributable to two factors: variations in task types and the unique tasks-specific characteristics of Task Fulfillment and Relevance. The scoring rubrics for TFR, which include task-specific assessment focuses, contribute to variation in examinees' scores across tasks. These findings are actually consistent with current theoretical models of communicative competence, which claim that communicative competence is to some extent stable while recognizing that some components may be local and dependent on the contexts in which the interactions occur (Chalhoub-Deville, 2003).

The variability due to tasks can be reduced by increasing the number of tasks and reducing the person-by-task interactions that would not diminish domain representation.

There are two potential ways to reduce the variability due to tasks: one is to increase the number of tasks, and the other is to reduce the person-by-task interactions in ways that would not weaken domain representation. There is a limit on the number of tasks that can be used in large-scale assessments due to logistic and efficiency concerns. If the person-by-task interactions attempt to be reduced by incorporating task that are not differentiated in task types, there is a tradeoff between reducing variation in performance across tasks and weakening domain representation.

It is possible to compute a composite score for each task, which is an average or weighted average of the analytic scores. The composite score indicates the overall quality of a particular task response. However, the magnitude, interpretation, and dependability of the composite scores depend on how the dimensions are weighted. The dependabilities of composite scores for Task 3 and Task 5 were compared across different weighting schemes when a different number of ratings were used. The results in Chapter 6 showed that the coefficients increased considerably from one rating to double ratings across all the weighting schemes. This suggested that double ratings may be necessary for all the tasks. It was found that the use of different weighting schemes had some impact on the dependability of task-level composite scores. The impact of weighting schemes on composite score dependability was influenced by a few factors: the variances and reliability of the dimensions and the correlations among the analytic scores. Since the universe score variances of the five dimensions differed somewhat (See Table 16), the phi coefficients for Task 3 and Task 5 were influenced by the choice of weighting schemes. For Task 3, with equal weights to CoP and Pron, the phi coefficient was .57 when double ratings were used. Giving more weight to Pron seemed preferable in terms of maximizing the dependability of the composite scores. For Task 5, the phi coefficient was .69 when double ratings were used and equal weights were applied to all the dimensions. Similarly, giving more weight to Pron could also maximize the dependability of the composite scores. However, Webb and Shavelson (1981) suggested that expert weights are construct and theory driven and may be preferable to sets of weights which maximize generalizability of composite scores.

### 7.1.2 Sub-question 2:

*To what extent are the analytic scores separable in terms of task fulfillment and relevance, clarity of presentation, grammar and vocabulary, pronunciation, and confidence and fluency?*

As mentioned in Chapter 6, multivariate generalizability analysis could only be conducted for Task 3 and Task 5 with double ratings. Therefore, the separability of dimensions on the other three tasks could not be investigated. Given the fact that CoFlu was not examined in either of these two tasks, only the distinctness of TFR, CoP, GV and CoFlu were investigated. The disattenuated correlations among the analytic scores by task ranged from moderately high to perfect (larger than 1). The disattenuated correlation between CoP and Pron on Task 3 ranked the lowest among all the correlations. Those between TFR and GV and Pron on Task 5 were also relatively low with coefficients of .872 and .877, respectively. Conversely, TFR was highly correlated with CoP with the correlation coefficient even larger than 1.

The CFA results in Chapter 4 indicated that the factor correlations among TFR, CoP, GV, Pron and CoFlu in the CTCU and CTUM models were quite high with coefficients above .80 suggesting that the language components were highly related to each other, although the speaking ability was multi-componential. The lowest correlation was between TFR and Pron (.67). Further, the correlations between Pron and the other dimensions were also not very high, around or below .80.

The raters' verbal reports were somewhat consistent with the quantitative results. The raters thought it was more difficult to separate task fulfillment and relevance and clarity of presentation because test takers might not fulfill the task well if they did not express themselves clearly. Hence, very high disattenuated correlations between TFR and CoP were observed (even larger than 1). Raters also mentioned that test takers with a better command of vocabulary and grammar were more likely to have a good pronunciation. This was reflected in the factor correlations where Pron had a closer association with GV than the other dimensions. The relative distinctness of Pron and TFR is probably due to three factors: Firstly, conceptually there is little overlap between TFR and Pron. Pron is one facet of language proficiency while TFR is with a different and more generic nature (i.e., it could apply to non-language related tasks). Secondly, operationally the descriptors for TFR and Pron might be distinct and it is easier for the raters to rate TFR in isolation from Pron. Thirdly, while there are standardized scoring rubrics for all the criteria which raters use for five speaking tasks, two of these tasks (Task 1 and Task 4) also have rating grids for TFR which are specific to the particular version of the tasks. Raters thus may use a combination of rating grids and scoring rubric for TFR on Task 1 and Task 4.

Overall, all the five dimensions across the tasks are highly correlated with each other. This might be due to a combination of the following reasons: (a) Conceptually the construct underlying the analytic scores are highly correlated; (b) The dimensions may be distinct conceptually, but raters are unable to consistently interpret the descriptors for each dimension either because they could not distinguish among them or because there is overlap among the descriptors in the dimensions; (c) There might be a ‘halo effect’: raters first decided that a candidate was at a certain level and then awarded that level for each dimension; Two raters mentioned in their verbal reports that their general impressions on the test takers’ performance might influence their ratings on different dimensions.   (d) Examinees might be more likely to do equally well in TFR, CoP, GV, Pron and CoFlu. Most examinees are in their final year in a Hong Kong university, so similarities in their English learning practices and exposure to English may have lead to high correlations among their analytic scores.

### *7.2 Discussion of research question 2*

The second research question addressed the meaningfulness of score interpretations: *Are the interpretations about students’ oral proficiency for workplace communication meaningful?* In this section, each sub-question of the research question 2 is discussed based on the results reported in Chapters 4, 5 and 6. The results of Confirmatory Factor Analyses, Item Response

Theory (IRT) Analyses, and the raters' verbal reports are referenced either to address each sub-question or to explain possible causes of the results.

### 7.2.1 Sub-question 1:

*Is the multi-componential factor structure assumed in this test design supported?*

The CTUM, HTUM, CTCU and HTCU models all had a close fit to the test data. They did not show much difference in terms of the standardized parameter estimates. The factor loadings of the analytic scores on the five dimensions were all statistically significant. All the loadings were substantial (above .50) except the one of TFR4 on TFR (only .36). The unacceptable fit of Bi-factor model demonstrated that language ability measured was multi-componential rather than unitary. The fit of the CTUM model was significantly better than that of the bi-factor model, supporting the hypothesis that the five trait factors are psychometrically distinct from one another. The significant and substantive trait factor loadings suggested that the multi-componential structure as proposed by the test developers were supported by the statistical analysis of the test scores. Most trait factor loadings were larger than the method factor loadings, indicating that the analytic ratings could be meaningfully interpreted as indicators of five dimensions although method factors also have significant effects on analytic scores. Hence, meaningfulness of the score interpretations could be assured. Taking model complexity and interpretability into consideration, the higher-order trait factor models might be more preferable since they not only confirmed the current multi-componential view of language ability in the literature but also provided the most parsimonious explanation of the relationships among the five dimensions and overall speaking proficiency. The presence of a higher-order speaking ability factor governing the five trait factors also supported the practice of reporting one composite score.

### 7.2.2 Sub-question 2:

*Are there any problematic items that are weak in measuring test takers' speaking ability?*

Amongst the generally high (>0.50) trait factor loadings, TFR4 stood out as an anomaly with a factor loading of less than 0.40. This is not surprising as the abnormally low TFR4 correlations with all other 13 items harbingered this phenomenon. TFR4 had the highest method loading (.60) on Task 4 and the lowest trait factor loading (.36) on TFR, which suggested TFR4 might be too task specific and weak in measuring students' speaking ability to fulfill a speaking task in a relevant way.

By examining the item trace lines, it was found that the assessment levels were ordered within all the items except TFR4. TFR4 had a very low discrimination. The trace lines for TFR4 appeared much flatter compared to those good items. The trace lines also indicated that raters might have difficulties differentiating levels 2 and 3. Along the latent trait continuum, test takers had similar chances of being awarded 2, 3, and 4. However, the raters' verbal reports showed that most raters did not have much difficulty differentiating across the performance levels, especially the middle ground such as 2, 3, and 4. The problem of TFR4 can only be due to the nature of the task itself and its low discrimination. The above results suggested that TFR4 was weaker in measuring students' speaking ability to fulfill a speaking task in a relevant way. In retrospect, Task 4 assesses fulfillment and relevance from a task in which students are given a maximum of one minute to leave a voice mail message. The one-minute time limit and its rigid format might have left little room for the superior speakers to fully display their speaking ability and differentiate themselves from the others. Furthermore, Task 4 is purely an information providing task. Students who were weak on the other dimensions could still score highly on TFR4 as long as they are able to provide the information required. On the other hand, students with superior speaking abilities might have tried to elaborate on their responses, run out of time, and hence get low scores on this dimension. In other words, test wiseness may play a big role in answering this item. It was learned that an administrative decision was recently made to discard TFR4 score in the composite score computation and the findings seemed to support this decision.

### 7.2.3 Sub-question 3:

*Do tasks have effects on test takers' speaking performance?*

In the CFA results of Chapter 4, all method factor loadings for the method factors were statistically significant, suggesting the presence of task effects and these effects could not be neglected in the test design and development. Particularly, the method factor loadings of TFR4 and CoFlu4 were both larger than their trait factor loadings, suggesting that Task 4 had significantly high impacts on test takers' oral performance. This might be due to the task nature itself as discussed in the previous section. The IRT results demonstrated that there were interactions between TFR and task types. The large person-by-task interaction effect and relatively large task effect for TFR in the G theory analysis also showed that examinees performed differently on TFR across Task 1, Task 4, and Task 5. The above results indicated that task types had great effects on test takers' speaking abilities especially TFR and that this language ability component might be too task specific. Chalhoub-Deville (2003) claims that some communicative components may be local and dependent on the contexts in which the interactions occur. Task 1 seemed to be the most difficult task for TFR while Task 4 was the easiest one. In Task 1, test candidates were asked to summarize an interview while Task 4 only required them to leave a telephone message. Obviously, Task 1 placed a heavier cognitive load on test candidates. In second language acquisition research, it has been demonstrated that task features have impacts upon the demand that tasks place on test takers. According to Tarone (1988), the construct of a 'stable competence' is untenable and performance data can only support the weaker construct of 'variable capability'. Similarly, Ellis (1985) argues for a heterogeneous capability that is manifested differentially depending upon task conditions in operation at the time of production.

Language testers have also begun to pay attention to the impact of task conditions on task difficulties and test performance (e.g., Bachman & Palmer, 1981; Bachman & Palmer, 1996; Bachman, 2002; Carr, 2006; Fulcher, 1995). Fulcher (1995) did not show positive attitudes towards the variable competence model of Second Language Acquisition and argued that if the language abilities were task specific, then each test would be a test of performance in the specific situation defined by the facets of the test situation. It would thus be impossible to generalize the meaning of test scores from any test task to other tasks, or any non-test situation, unless there is a precise match between every facet of the test and the criterion. He suggested that these insights from SLA research might be relevant to discriminating between what is contextually determined and what resides in competence. This study further demonstrated that task features had relatively large impacts upon certain language abilities such as Task Fulfillment and Relevance (TFR) while there was little variation in test takers’ performance on other ability components like Grammar and Vocabulary and Pronunciation. TFR proves to be more context-dependent while language knowledge components such as vocabulary and pronunciation seem to be more stable and context-free.

### 7.2.4 Sub-question 4:

*To what extent does what the test takers report correspond to their oral test performance?*

The first key issue in this sub-question concerned the factor structure of the self-assessment questionnaire. The internal consistency estimate of .971 indicated that this self-assessment questionnaire was a reliable measure. The CFA results showed that the correlated trait-correlated uniqueness model fit the questionnaire data very well with a RMSEA value of .058 and CFI, NFI, and NNFI value all above .99. All the factor loadings were substantial (above .75) and all the factors were highly correlated with each other. The above results supported the multi-componential nature of the 14 questionnaire items.

The second issue examined the relationship between test takers' perceptions of their speaking abilities and their actual test performance. The overall model fit was excellent (df=295; $\chi 2$ =385.38; p<.05; RMSEA=0.027). The regression coefficients from the trait factors in the questionnaire to the corresponding ones of the GSLPA SLT were significant at the .05 level but the values were relatively low, ranging from .221 to .319. This suggested that the prediction effects of students' self-ratings were significant yet low. The highest prediction effect of CoFlu with a value of .319 indicated that students can predict their Confidence and Fluency better compared to other dimensions.

These findings are consistent with earlier empirical studies where self-assessment has been found to be positively correlated with test scores, ability and achievement (Birckbichler et al., 1993; Brantmeier, 2005; Hargan, 1994; Heilenman, 1991; Krausert, 1991; Oscarson, 1978; Ross, 1998).The results for this study concur with the Blanche and Merino's (1989) conclusion that self-assessment typically provides robust concurrent validity with criterion variables.

### *7.3 Discussion of research question 3*

The third research question addressed the impartiality of score interpretations: *Are the score-based interpretations impartial across different subgroups of test takers?* In this section, the results of multi-sample Confirmatory Factor Analysis (CFA) and Differential Item Functioning (DIF) are discussed to answer this research question. The possible explanations of the findings are also explored, if applicable.

DIF and multi-sample CFA both examined the measurement equivalence across different subgroups of test takers. In the scale-level analysis, the set of items comprising a test are usually examined using multi-sample CFA, while DIF is often used to examine the item-level invariance (Zumbo, 2003). Hence, multi-sample CFA allows one to investigate whether the construct is measured equivalently and further whether the model parameters are invariant across the different groups. In essence, the purpose of the multi-sample CFA is to reproduce the observed covariance matrix of the items based on a specified number of factors and factor correlations. The IRT framework is a powerful way to detect item-level DIF.   If the item response functions or item parameters are different for two groups, it is clear that the item has DIF.

The multi-sample CFA results showed that the factor structure was significantly different between males and females (df=157; $\chi 2$=217.81; p=0.0001). Hence, the factorial invariance assumption between the female and male student groups was rejected, indicating that the latent trait means could not be meaningfully compared. The comparison of the factor loadings between females and males showed that only the factor loading of CoFlu4 for the male group was significantly different from the female group at the 0.05 level. DIF results showed that the majority of the items displayed no DIF. Males and females had significantly different slope parameters on CoFlu4 and different location parameters on Pron5. CoFlu4 showed DIF with significantly different slope parameters between males and females, which were consistent with the multi-sample CFA results. This suggested that CoFlu4 did not do an equally good job in measuring test takers' language ability and could not discriminate males and females equally well. Nevertheless, the magnitude of DIF for CoFlu4 was not high with a slope difference of .49. Similarly, the magnitude of DIF for Pron5 was quite low with a location parameter $b_2$ difference of .19. The examination of descriptive statistics for females and males indicated that these two groups had the largest item mean difference (.27) on Pron5 and a relatively large mean difference (.21) on CoFlu4. T-test also showed that these mean differences were significant at the .05 level. Hence, the occurrence of DIF on Pron5 could be attributable to the group mean difference on this item. In addition, the group means comparison between females and males also indicated that on average females scored significantly higher than males on the construct of GSLPA SLT. This could further confirm that within this cohort and possibly typical of university students in Hong Kong, females possess higher oral proficiency in English than males. The occurrence of DIF items between males and females may have come from these two groups' real differences in certain aspects of language ability that the GSLPA SLT targets. DIF is a necessary but not sufficient condition for identifying item bias. Item bias occurs "when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some identifiable characteristic of the test item or testing situation that is not relevant to the test purpose" (Zumbo, 1999, p.12). Examination of test characteristics such as scoring rubrics, rating process and task types would be required to determine whether the difference in performance represents any bias related to group membership.

The assumption of the factorial invariance between business and non-business students was also rejected. Most factor loadings of business students and non-business students were similar to each other while only TFR1 and TFR5 displayed some significant differences which were less than .15. Interestingly, no items showed DIF although there were some minor differences on the slope and location parameters. Although the test construct was not measured exactly equivalently between business and non-business students, the slope and location parameters for all the items had no significant differences between these two groups. Each test item did an equally good job in measuring students' language ability across these two subgroups of test takers.

Although the assumptions of the strict factorial invariance for males vs. females and business vs. non-business majors were both rejected, most factor loadings had no significant differences related to group membership. All the items showed no DIF between business and non-business students and the majority of items displayed no DIF between females and males. The source of DIF in items CoFlu4 and Pron5 may be attributed to the group mean difference on the latent trait and their real differences on certain aspects of language ability measured in the GSLPA SLT. This provides backing for the impartiality of score interpretations, indicating that the rating-based interpretations from GSLPA SLT are impartial to a large extent across subgroups of test takers (males vs. females; business vs. non-business).

### *7.4 Implications for the testing program under study*

The present study is intended to aid the test developers in identifying the most problematic area(s) in the current test in order to make the test use more justifiable. The low dependability of task types indicates that two raters are necessary for each task. Although CFA results indicated the five dimensions were psychometrically distinct from each other, the low dependability of the analytic scores and the high disattenuated correlations in this study both suggested that there would not be much gain by using analytic rating scales in operational settings. Hence, taking the cost effectiveness into consideration, holistic scoring seems preferable in this testing context. Information about performance on individual dimensions would be very useful for practice and self-learning where the stakes are much lower. Analytic scoring can be provided as an additional service for potential GSLPA examinees who want to practice their speaking skills and improve their performance on the speaking section.

The large task effects of Task 4 and the low discrimination of CoFlu and TFR measured in this task showed that the test developers may need to make some revisions to this task. The relatively short time limit and its rigid format might have left little room for the superior speakers to fully display their speaking ability and differentiate themselves from others. There are two possible ways to solve this: 1) increase the complexity of this task and place heavier cognitive loads on the test takers; 2) give examinees more time to answer this question and have more space to display their language ability.

The test information curve peaked at trait levels of -2.3, -0.5 and 1.5, indicating that this test discriminated students best at these trait levels. The test provided the least amount of information at trait levels of higher than 2.5 and the majority amount of information about students at trait levels of lower than 2. This finding may contribute to the standard setting of GSLPA LT and further minimize the classification errors in decision making.

### *7.5 Implications for language assessment theory and practice*

The present study serves as an example of applying and Assessment Use Argument (AUA) in a language testing program and illustrates its usefulness for guiding research into the development and justification of a language assessment. It is hoped that this study will deepen the understanding of the argument-based approaches to test validation and particularly the Assessment Use Argument framework. Specifically, the study demonstrates the practicability of an AUA in the following three aspects: 1) guiding the justification process of a language assessment, 2) collecting different types of evidence for the claims and warrants, and 3) identifying the most critical areas in the current test. An AUA proved to be a powerful framework for evaluating the language assessments in terms of how well they work in practice and not just in terms of technical characteristics. One strength of an AUA lies in its clear articulation about which types of evidence need to be collected for which claims or warrants. The clear linkage between the evidence types and the claims and warrants makes much easier for the researchers to develop, evaluate, and justify a language assessment.

In this study, the large person-by-task interaction and task effects for TFR was probably attributable to the characteristics of this dimension as well as the nature of the tasks. There are two potential ways to reduce the variability due to tasks: one is to increase the number of tasks, and the other is to reduce the person-by-task interactions in ways that would not weaken domain representation. However, it seems difficult to balance the domain representation and the consistency in scores on this dimension. This may pose some challenges to Bachman and Palmer's attempt to design assessment as a sample of performance from the Target Language Use domain.

In the process of justifying the language assessment, an AUA demands that the evaluation of the test be carried out at many levels and this required different data types and analyses of many kinds. The size and complexity of the justification study may be a big challenge for a single researcher to undertake.

### *7.6 Limitations and further research*

This study was limited in several ways. First of all, the five dimensions were not included in the DIF analyses. DIF was conducted with a graded response model as a baseline model. A full IRT model with a general speaking dimension, five componential speaking dimensions, and five tasks might produce different DIF results from the unidimensional model. For example, some items may display DIF on the dimensions even if not on the general speaking ability. It is hoped to fit this full model with Metropolis Hastings Robbins Monroe Algorithm in the future.

Secondly, normality is always a concern for linear models when parameters are interpreted based on their P values. Although the skewness and kurtosis values of the outcome variables show roughly normal distributions, it is safer to address potential non-normality via the Satorra-Bentler approach (Satorra & Bentler, 1988). Model fit might further improve when non-normality, if any, is properly taken into account. The parameter estimates would remain the same using the Satorra-Bentler approach. The standard error estimates would differ, but probably not to the extent that they would change the substantive interpretations of the major findings assuming normality. After all, the factor loadings and structural regression coefficients in this study surpass their standard error estimates by a very wide margin, requiring unrealistically high non-normality adjustment for the estimates to be non-significant.

Another limitation is with G theory analyses. The complex rating and task schemes featured an unbalanced design *p* x *(r: t: d)* for the whole test with raters partially nested within tasks and tasks partially nested within dimensions. Similarly, for the five dimensions raters were also partially nested within tasks. This design is a confounded design in which the conditions of crossing or nesting are not fully met although it is quite typical in large-scale assessments. This design does not allow for certain estimates of the confounding effects involving rater and tasks. The researcher treated it as an unbalanced nested design in order to be able to calculate the variance component associated with raters and tasks. Hence, G theory analyses results in this study are only an approximation of the actual dependability estimates for the whole test as well as the dimensions.

Finally, the present study intends to be informative rather than judgmental. Since its primary purpose is to help test developers identify areas for improvement in the current test design and administration to support the intended use, it is beyond the scope of the study to reach any conclusion as to whether the intended use is justified or not. Furthermore, since this study is conducted in the context of an ESL oral test administered in a Hong Kong university, the results may not generalize to other research contexts (e.g. different tests, different test takers, different universities).

# Appendices

## Appendix 1. Questionnaire

Thank you for taking the GSLPA (English) Spoken Language Test. As part of our on-going research and development, from time to time we carry out student surveys. We would be very grateful if you could spare a few minutes to complete this survey before you leave. Completion of this survey will not affect your GSLPA (English) result. All responses and identities will be kept strictly confidential.

1. Your GSLPA Candidate Number: ___________________
2. Gender:
    1. Male
    2. Female
3. Where are you from?
    1. Hong Kong
    2. Mainland China
    3. Others ___________(Please specify)
4. Did you attend a GSLPA Spoken Language Test workshop?
    1. Yes
    2. No
5. Before you took the GSLPA Spoken Language Test, to what extent were you familiar with the criteria on which you were to be tested?
    1. Unfamiliar

2. Not very familiar
3. Familiar
4. Very familiar

6. In the GSLPA Spoken Language Test, each task tests students' ability on different criteria. Before you took the test, to what extent were you familiar with which criteria are tested by each task?
    1. Unfamiliar
    2. Not very familiar
    3. Familiar
    4. Very familiar

The following items are about the criteria of English oral proficiency you do well in. Please click the appropriate options based on your own experiences.

7. In Task 1 of the GSLPA Spoken Language Test, you listened to an interview and then were asked to provide a summary of the information in the interview. This task was measured on four criteria: Task fulfillment and relevance (i.e. content), Clarity of presentation (i.e. organization), Grammar and vocabulary, and Pronunciation. How would you rate your performance in the Test on each of these four criteria?

|  | 1(Very weak) | 2 | 3 | 4 | 5 | 6(Very strong) |
|---|---|---|---|---|---|---|
| Task Fulfillment and Relevance |  |  |  |  |  |  |
| Clarity of Presentation |  |  |  |  |  |  |
| Grammar and |  |  |  |  |  |  |

| | | | | | | |
|---|---|---|---|---|---|---|
| Vocabulary | | | | | | |
| Pronunciation | | | | | | |

8. In Task 2 of the GSLPA Spoken Language Test, you were asked to answer questions as part of an interview. This task was measured on two criteria: Grammar and vocabulary, and Confidence and fluency. How would you rate your performance in the Test on each of these four criteria?

| | 1(Very weak) | 2 | 3 | 4 | 5 | 6(Very strong) |
|---|---|---|---|---|---|---|
| Grammar and Vocabulary | | | | | | |
| Confidence and Fluency | | | | | | |

9. In Task 3 of the GSLPA Spoken Language Test, you were asked to provide an oral presentation of information from a written (graphic) source. This task was measured on two criteria: Pronunciation and Clarity of presentation (organization). How would you rate your performance in the Test on each of these four criteria?

| | 1(Very weak) | 2 | 3 | 4 | 5 | 6(Very strong) |
|---|---|---|---|---|---|---|
| Clarity of Presentation | | | | | | |
| Pronunciation | | | | | | |

10. In Task 4 of the GSLPA Spoken Language Test, you were asked to leave a telephone message. This task was measured on two criteria: Task fulfillment and relevance (content), and Confidence and fluency. How would you rate your performance in the Test on each of these four criteria?

| | 1(Very weak) | 2 | 3 | 4 | 5 | 6(Very strong) |
|---|---|---|---|---|---|---|
| Task Fulfillment and Relevance | | | | | | |
| Confidence and Fluency | | | | | | |

11. In Task 5 of the GSLPA Spoken Language Test, you were asked to talk to a visitor informally, giving opinions and information in response to the question. This task was measured on four criteria: Task fulfillment and relevance (content), Clarity of presentation (organisation), Grammar and vocabulary, and Pronunciation. How would you rate your performance in the Test on each of these four criteria?

| | 1(Very weak) | 2 | 3 | 4 | 5 | 6(Very strong) |
|---|---|---|---|---|---|---|
| Task Fulfillment and Relevance | | | | | | |
| Clarity of Presentation | | | | | | |
| Grammar and Vocabulary | | | | | | |
| Pronunciation | | | | | | |

# Appendix 2. Descriptive statistics for the analytic scores

| | Mean for All N=999 | Mean for M N=472 | Mean for F N=527 | Mean for B N=436 | Mean for N N=563 | Std. Deviation for All N=999 | Std. Deviation for M N=472 | Std. Deviation for F N=527 | Std. Deviation for B N=436 | Std. Deviation for N N=563 | Std. Deviation for N N=999 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | 3.59 | 3.52 | 3.65 | 3.70 | 3.50 | .903 | .900 | .903 | .922 | .880 | .880 |
| CoP1 | 3.68 | 3.57 | 3.77 | 3.74 | 3.63 | .742 | .745 | .726 | .776 | .711 | .711 |
| GV1 | 3.73 | 3.60 | 3.84 | 3.80 | 3.67 | .690 | .672 | .687 | .708 | .672 | .672 |
| Pron1 | 3.84 | 3.70 | 3.98 | 3.85 | 3.84 | .709 | .677 | .712 | .750 | .676 | .676 |
| CoFlu2 | 3.79 | 3.69 | 3.88 | 3.88 | 3.73 | .786 | .778 | .782 | .784 | .782 | .782 |
| GV2 | 3.80 | 3.69 | 3.90 | 3.88 | 3.75 | .697 | .687 | .692 | .715 | .678 | .678 |
| Pron3R1 | 3.89 | 3.76 | 4.01 | 3.89 | 3.90 | .713 | .698 | .707 | .719 | .709 | .709 |
| Pron3R2 | 3.90 | 3.77 | 4.02 | 3.89 | 3.92 | .686 | .682 | .668 | .740 | .641 | .641 |
| Pron3Average | 3.90 | 3.77 | 4.02 | 3.89 | 3.91 | .590 | .579 | .575 | .616 | .569 | .569 |
| Pron3Round | 4.12 | 3.99 | 4.24 | 4.12 | 4.13 | .640 | .623 | .631 | .676 | .611 | .611 |
| COP3R1 | 3.79 | 3.68 | 3.89 | 3.83 | 3.76 | .736 | .735 | .723 | .738 | .734 | .734 |
| COP3R2 | 3.85 | 3.78 | 3.91 | 3.93 | 3.79 | .736 | .721 | .744 | .805 | .673 | .673 |
| CoP3Average | 3.82 | 3.73 | 3.90 | 3.88 | 3.78 | .586 | .569 | .591 | .622 | .554 | .554 |
| CoP3Round | 4.07 | 4.00 | 4.13 | 4.12 | 4.03 | .641 | .625 | .650 | .680 | .606 | .606 |
| TFR4 | 4.16 | 4.13 | 4.19 | 4.19 | 4.13 | 1.300 | 1.317 | 1.286 | 1.275 | 1.320 | 1.320 |
| CoFlu4 | 3.96 | 3.85 | 4.06 | 4.01 | 3.92 | .810 | .813 | .796 | .808 | .811 | .811 |
| TFR5R1 | 4.01 | 3.92 | 4.09 | 4.04 | 3.98 | .737 | .721 | .742 | .724 | .746 | .746 |
| TFR5R2 | 4.02 | 3.94 | 4.09 | 4.08 | 3.97 | .719 | .702 | .727 | .755 | .686 | .686 |
| TFR5Average | 4.014 | 3.926 | 4.093 | 4.062 | 3.977 | .5929 | .5799 | .5938 | .5966 | .5878 | .5878 |
| TFR5Round | 4.25 | 4.15 | 4.34 | 4.29 | 4.22 | .658 | .640 | .662 | .664 | .653 | .653 |
| COP5R1 | 3.77 | 3.64 | 3.88 | 3.83 | 3.72 | .703 | .705 | .683 | .712 | .694 | .694 |
| COP5R2 | 3.78 | 3.67 | 3.88 | 3.86 | 3.72 | .720 | .708 | .718 | .740 | .698 | .698 |
| CoP5Average | 3.77 | 3.66 | 3.88 | 3.84 | 3.72 | .581 | .565 | .576 | .598 | .563 | .563 |
| CoP5Round | 4.00 | 3.89 | 4.10 | 4.07 | 3.94 | .629 | .608 | .632 | .643 | .613 | .613 |
| GV5R1 | 3.78 | 3.68 | 3.87 | 3.84 | 3.74 | .693 | .695 | .678 | .684 | .697 | .697 |
| GV5R2 | 3.78 | 3.66 | 3.88 | 3.83 | 3.74 | .691 | .690 | .675 | .696 | .685 | .685 |
| GV5Average | 3.779 | 3.667 | 3.880 | 3.835 | 3.736 | .5897 | .5788 | .5819 | .5904 | .5861 | .5861 |
| GV5Round | 4.00 | 3.90 | 4.09 | 4.05 | 3.96 | .637 | .642 | .619 | .650 | .625 | .625 |
| Pron5R1 | 3.90 | 3.74 | 4.03 | 3.92 | 3.88 | .724 | .690 | .726 | .770 | .687 | .687 |
| Pron5R2 | 3.89 | 3.75 | 4.01 | 3.90 | 3.88 | .691 | .689 | .668 | .740 | .650 | .650 |
| Pron5Average | 3.891 | 3.744 | 4.024 | 3.906 | 3.880 | .5968 | .5699 | .5898 | .6327 | .5678 | .5678 |
| Pron5Round | 4.11 | 3.97 | 4.24 | 4.15 | 4.08 | .635 | .602 | .637 | .676 | .601 | .601 |

| | Skewness for All N=999 | | Skewness for M N=472 | | Skewness for F N=527 | | Skewness for B N=436 | | Skewness for N N=563 | |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | -.052 | .077 | -.092 | .112 | -.019 | .106 | .036 | .117 | -.163 | .103 |
| CoP1 | .099 | .077 | .269 | .112 | -.038 | .106 | .129 | .117 | .029 | .103 |
| GV1 | .122 | .077 | .375 | .112 | -.100 | .106 | .115 | .117 | .103 | .103 |
| Pron1 | .166 | .077 | .249 | .112 | .067 | .106 | .148 | .117 | .176 | .103 |
| CoFlu2 | .134 | .077 | .246 | .112 | .038 | .106 | .164 | .117 | .113 | .103 |
| GV2 | .268 | .077 | .478 | .112 | .096 | .106 | .258 | .117 | .255 | .103 |
| Pron3R1 | .222 | .077 | .361 | .112 | .111 | .106 | .271 | .117 | .183 | .103 |
| Pron3R2 | -.023 | .077 | .001 | .112 | -.024 | .106 | .118 | .117 | -.168 | .103 |
| Pron3Average | .107 | .077 | .276 | .112 | -.022 | .106 | .159 | .117 | .062 | .103 |
| Pron3Round | .253 | .077 | .271 | .112 | .254 | .106 | .253 | .117 | .253 | .103 |
| COP3R1 | .234 | .077 | .391 | .112 | .115 | .106 | .288 | .117 | .192 | .103 |
| COP3R2 | .167 | .077 | .324 | .112 | .027 | .106 | .107 | .117 | .134 | .103 |
| CoP3Average | .258 | .077 | .474 | .112 | .072 | .106 | .241 | .117 | .218 | .103 |
| CoP3Round | .165 | .077 | .207 | .112 | .116 | .106 | .149 | .117 | .131 | .103 |
| TFR4 | -.456 | .077 | -.469 | .112 | -.442 | .106 | -.481 | .117 | -.436 | .103 |

|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| CoFlu4 | .185 | .077 | .207 | .112 | .190 | .106 | .036 | .117 | .301 | .103 |
| TFR5R1 | .014 | .077 | -.043 | .112 | .043 | .106 | .007 | .117 | .026 | .103 |
| TFR5R2 | -.189 | .077 | -.318 | .112 | -.111 | .106 | -.295 | .117 | -.127 | .103 |
| TFR5Average | -.098 | .077 | -.110 | .112 | -.111 | .106 | -.265 | .117 | .030 | .103 |
| TFR5Round | .045 | .077 | .154 | .112 | -.067 | .106 | -.068 | .117 | .132 | .103 |
| COP5R1 | .184 | .077 | .337 | .112 | .083 | .106 | .113 | .117 | .236 | .103 |
| COP5R2 | .018 | .077 | -.010 | .112 | .032 | .106 | -.044 | .117 | .041 | .103 |
| CoP5Average | .110 | .077 | .286 | .112 | -.046 | .106 | -.015 | .117 | .193 | .103 |
| CoP5Round | .073 | .077 | .173 | .112 | -.034 | .106 | .041 | .117 | .077 | .103 |
| GV5R1 | .177 | .077 | .497 | .112 | -.099 | .106 | .219 | .117 | .156 | .103 |
| GV5R2 | -.040 | .077 | .020 | .112 | -.080 | .106 | .075 | .117 | -.141 | .103 |
| GV5Average | .192 | .077 | .468 | .112 | -.037 | .106 | .200 | .117 | .186 | .103 |
| GV5Round | .162 | .077 | .338 | .112 | .034 | .106 | .255 | .117 | .069 | .103 |
| Pron5R1 | .160 | .077 | .468 | .112 | -.112 | .106 | .055 | .117 | .257 | .103 |
| Pron5R2 | .025 | .077 | -.129 | .112 | .215 | .106 | .133 | .117 | -.109 | .103 |
| Pron5Average | .177 | .077 | .243 | .112 | .110 | .106 | .098 | .117 | .246 | .103 |
| Pron5Round | .323 | .077 | .247 | .112 | .358 | .106 | .163 | .117 | .459 | .103 |

|  | Kurtosis for All |  | Kurtosis for M |  | Kurtosis for F |  | Kurtosis for B |  | Kurtosis for N |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | N=999 |  | N=472 |  | N=527 |  | N=436 |  | N=563 |  |
| TFR1 | -.167 | .155 | -.130 | .224 | -.212 | .212 | -.155 | .233 | -.269 | .206 |
| CoP1 | -.168 | .155 | .075 | .224 | -.242 | .212 | -.119 | .233 | -.300 | .206 |
| GV1 | .050 | .155 | .553 | .224 | -.048 | .212 | .241 | .233 | -.145 | .206 |
| Pron1 | .282 | .155 | .673 | .224 | .149 | .212 | .238 | .233 | .273 | .206 |
| CoFlu2 | .039 | .155 | .234 | .224 | -.022 | .212 | -.044 | .233 | .099 | .206 |
| GV2 | .327 | .155 | .798 | .224 | .175 | .212 | .515 | .233 | .137 | .206 |
| Pron3R1 | .318 | .155 | .821 | .224 | .105 | .212 | .158 | .233 | .461 | .206 |
| Pron3R2 | .568 | .155 | .135 | .224 | 1.102 | .212 | .480 | .233 | .569 | .206 |
| Pron3Average | .409 | .155 | .841 | .224 | .360 | .212 | .351 | .233 | .451 | .206 |
| Pron3Round | .558 | .155 | 1.145 | .224 | .174 | .212 | .367 | .233 | .724 | .206 |
| COP3R1 | -.242 | .155 | -.056 | .224 | -.271 | .212 | -.366 | .233 | -.154 | .206 |
| COP3R2 | -.242 | .155 | .172 | .224 | -.480 | .212 | -.452 | .233 | -.153 | .206 |
| CoP3Average | .257 | .155 | .850 | .224 | .020 | .212 | .052 | .233 | .395 | .206 |
| CoP3Round | .469 | .155 | .920 | .224 | .172 | .212 | .106 | .233 | .820 | .206 |
| TFR4 | -.436 | .155 | -.511 | .224 | -.364 | .212 | -.346 | .233 | -.496 | .206 |
| CoFlu4 | .006 | .155 | -.015 | .224 | .052 | .212 | -.157 | .233 | .190 | .206 |
| TFR5R1 | .319 | .155 | .480 | .224 | .180 | .212 | .319 | .233 | .332 | .206 |
| TFR5R2 | .495 | .155 | 1.119 | .224 | -.034 | .212 | .638 | .233 | .368 | .206 |
| TFR5Average | .173 | .155 | .627 | .224 | -.169 | .212 | .546 | .233 | -.040 | .206 |
| TFR5Round | -.079 | .155 | .352 | .224 | -.301 | .212 | -.054 | .233 | -.052 | .206 |
| COP5R1 | .209 | .155 | .488 | .224 | .148 | .212 | .362 | .233 | .119 | .206 |
| COP5R2 | -.045 | .155 | .323 | .224 | -.371 | .212 | -.240 | .233 | .168 | .206 |
| CoP5Average | -.082 | .155 | .555 | .224 | -.376 | .212 | -.282 | .233 | .180 | .206 |
| CoP5Round | .074 | .155 | .865 | .224 | -.402 | .212 | -.302 | .233 | .436 | .206 |
| GV5R1 | .063 | .155 | .540 | .224 | -.040 | .212 | .155 | .233 | -.006 | .206 |
| GV5R2 | .219 | .155 | .176 | .224 | .362 | .212 | .340 | .233 | .080 | .206 |
| GV5Average | .245 | .155 | .691 | .224 | .249 | .212 | .457 | .233 | .090 | .206 |
| GV5Round | .424 | .155 | .632 | .224 | .470 | .212 | .531 | .233 | .288 | .206 |
| Pron5R1 | .414 | .155 | .830 | .224 | .572 | .212 | .173 | .233 | .644 | .206 |
| Pron5R2 | .550 | .155 | .087 | .224 | .849 | .212 | .405 | .233 | .610 | .206 |
| Pron5Average | .557 | .155 | .742 | .224 | .652 | .212 | .343 | .233 | .752 | .206 |
| Pron5Round | .766 | .155 | 1.003 | .224 | .614 | .212 | .455 | .233 | 1.114 | .206 |

# Appendix 3. T-test results between males and females on the 14 analytic ratings

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| TFR1 | Equal variances assumed | .054 | .816 | -2.207 | 997 | .028 | -.126 | .057 | -.238 | -.014 |
| | Equal variances not assumed | | | -2.207 | 985.550 | .028 | -.126 | .057 | -.238 | -.014 |
| CoP1 | Equal variances assumed | 4.992 | .026 | -4.303 | 997 | .000 | -.200 | .047 | -.292 | -.109 |
| | Equal variances not assumed | | | -4.297 | 978.709 | .000 | -.200 | .047 | -.292 | -.109 |
| GV1 | Equal variances assumed | 6.475 | .011 | -5.492 | 997 | .000 | -.237 | .043 | -.321 | -.152 |
| | Equal variances not assumed | | | -5.499 | 989.318 | .000 | -.237 | .043 | -.321 | -.152 |
| Pron1 | Equal variances assumed | 9.613 | .002 | -6.313 | 997 | .000 | -.278 | .044 | -.365 | -.192 |
| | Equal variances not assumed | | | -6.331 | 993.488 | .000 | -.278 | .044 | -.365 | -.192 |
| GV2 | Equal variances assumed | 4.558 | .033 | -3.915 | 997 | .000 | -.194 | .049 | -.291 | -.097 |
| | Equal variances not assumed | | | -3.917 | 986.251 | .000 | -.194 | .049 | -.291 | -.097 |
| CoFlu2 | Equal variances assumed | 8.076 | .005 | -4.767 | 997 | .000 | -.208 | .044 | -.294 | -.123 |
| | Equal variances not assumed | | | -4.769 | 986.681 | .000 | -.208 | .044 | -.294 | -.123 |
| Pron3 | Equal variances assumed | 29.460 | .000 | -6.423 | 997 | .000 | -.255 | .040 | -.333 | -.177 |
| | Equal variances not assumed | | | -6.428 | 987.508 | .000 | -.255 | .040 | -.333 | -.177 |
| CoP3 | Equal variances assumed | 13.926 | .000 | -3.133 | 997 | .002 | -.127 | .040 | -.206 | -.047 |
| | Equal variances not assumed | | | -3.140 | 991.888 | .002 | -.127 | .040 | -.206 | -.048 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR4 | Equal variances assumed | .236 | .627 | -.688 | 997 | .492 | -.057 | .082 | -.218 | .105 |
| | Equal variances not assumed | | | -.687 | 979.333 | .492 | -.057 | .083 | -.219 | .105 |
| CoFlu4 | Equal variances assumed | 3.490 | .062 | -4.064 | 997 | .000 | -.207 | .051 | -.307 | -.107 |
| | Equal variances not assumed | | | -4.059 | 979.900 | .000 | -.207 | .051 | -.307 | -.107 |
| TFR5 | Equal variances assumed | 19.951 | .000 | -4.686 | 997 | .000 | -.193 | .041 | -.274 | -.112 |
| | Equal variances not assumed | | | -4.695 | 991.085 | .000 | -.193 | .041 | -.274 | -.113 |
| CoP5 | Equal variances assumed | 1.760 | .185 | -5.259 | 997 | .000 | -.207 | .039 | -.284 | -.130 |
| | Equal variances not assumed | | | -5.270 | 991.829 | .000 | -.207 | .039 | -.284 | -.130 |
| GV5 | Equal variances assumed | .698 | .404 | -4.980 | 997 | .000 | -.199 | .040 | -.277 | -.120 |
| | Equal variances not assumed | | | -4.970 | 975.945 | .000 | -.199 | .040 | -.277 | -.120 |
| Pron5 | Equal variances assumed | 28.241 | .000 | -7.032 | 997 | .000 | -.277 | .039 | -.354 | -.199 |
| | Equal variances not assumed | | | -7.054 | 994.168 | .000 | -.277 | .039 | -.353 | -.200 |

## Appendix 4. T-test results between business and non-business majors on the 14 analytic ratings

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| TFR1 | Equal variances assumed | .115 | .734 | -3.457 | 997 | .001 | -.198 | .057 | -.311 | -.086 |
| | Equal variances not assumed | | | -3.436 | 912.965 | .001 | -.198 | .058 | -.311 | -.085 |
| CoP1 | Equal variances assumed | .616 | .433 | -2.450 | 997 | .014 | -.116 | .047 | -.208 | -.023 |
| | Equal variances not assumed | | | -2.423 | 892.554 | .016 | -.116 | .048 | -.209 | -.022 |
| GV1 | Equal variances assumed | .896 | .344 | -2.808 | 997 | .005 | -.123 | .044 | -.209 | -.037 |
| | Equal variances not assumed | | | -2.789 | 910.141 | .005 | -.123 | .044 | -.210 | -.037 |
| Pron1 | Equal variances assumed | 2.926 | .087 | -.367 | 997 | .713 | -.017 | .045 | -.105 | .072 |
| | Equal variances not assumed | | | -.363 | 884.014 | .717 | -.017 | .046 | -.107 | .073 |
| GV2 | Equal variances assumed | 2.679 | .102 | -2.962 | 997 | .003 | -.148 | .050 | -.246 | -.050 |
| | Equal variances not assumed | | | -2.961 | 934.151 | .003 | -.148 | .050 | -.246 | -.050 |
| CoFlu2 | Equal variances assumed | 2.162 | .142 | -2.950 | 997 | .003 | -.131 | .044 | -.218 | -.044 |
| | Equal variances not assumed | | | -2.929 | 909.803 | .003 | -.131 | .045 | -.218 | -.043 |
| Pron3 | Equal variances assumed | 3.640 | .057 | .111 | 997 | .911 | .005 | .041 | -.076 | .085 |
| | Equal variances not assumed | | | .110 | 885.873 | .912 | .005 | .041 | -.077 | .086 |
| CoP3 | Equal variances assumed | 18.961 | .000 | -2.296 | 997 | .022 | -.094 | .041 | -.174 | -.014 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Equal variances not assumed | | | -2.262 | 877.568 | .024 | -.094 | .041 | -.175 | -.012 |
| TFR4 | Equal variances assumed | .900 | .343 | -.667 | 997 | .505 | -.055 | .083 | -.218 | .107 |
| | Equal variances not assumed | | | -.670 | 950.239 | .503 | -.055 | .083 | -.218 | .107 |
| CoFlu4 | Equal variances assumed | .209 | .647 | -1.656 | 997 | .098 | -.086 | .052 | -.187 | .016 |
| | Equal variances not assumed | | | -1.657 | 937.128 | .098 | -.086 | .052 | -.187 | .016 |
| TFR5 | Equal variances assumed | 2.583 | .108 | -1.723 | 997 | .085 | -.072 | .042 | -.155 | .010 |
| | Equal variances not assumed | | | -1.720 | 927.955 | .086 | -.072 | .042 | -.155 | .010 |
| CoP5 | Equal variances assumed | 3.009 | .083 | -3.100 | 997 | .002 | -.124 | .040 | -.202 | -.045 |
| | Equal variances not assumed | | | -3.081 | 912.835 | .002 | -.124 | .040 | -.203 | -.045 |
| GV5 | Equal variances assumed | .497 | .481 | -2.063 | 997 | .039 | -.084 | .041 | -.163 | -.004 |
| | Equal variances not assumed | | | -2.052 | 916.501 | .040 | -.084 | .041 | -.164 | -.004 |
| Pron5 | Equal variances assumed | 14.614 | .000 | -1.733 | 997 | .083 | -.070 | .040 | -.150 | .009 |
| | Equal variances not assumed | | | -1.708 | 876.588 | .088 | -.070 | .041 | -.151 | .010 |

## Appendix 5. Correlations among the original, average and rounded ratings for Task3 and Task5

### Correlations for Pron3

| | T3PronR1 | T3PronR2 | T3PronAverage | T3PronRound |
|---|---|---|---|---|
| T3PronR1 | 1 | .421$^{**}$ | .850$^{**}$ | .768$^{**}$ |
| T3PronR2 | .421$^{**}$ | 1 | .836$^{**}$ | .785$^{**}$ |
| T3PronAverage | .850$^{**}$ | .836$^{**}$ | 1 | .921$^{**}$ |
| T3PronRound | .768$^{**}$ | .785$^{**}$ | .921$^{**}$ | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

### Correlations for CoP3

| | T3COPR1 | T3COPR2 | T3CoPAverage | T3CoPRound |
|---|---|---|---|---|
| T3COPR1 | 1 | .269$^{**}$ | .797$^{**}$ | .716$^{**}$ |
| T3COPR2 | .269$^{**}$ | 1 | .797$^{**}$ | .751$^{**}$ |
| T3CoPAverage | .797$^{**}$ | .797$^{**}$ | 1 | .921$^{**}$ |
| T3CoPRound | .716$^{**}$ | .751$^{**}$ | .921$^{**}$ | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

### Correlations for TFR5

| | T5TFRR1 | T5TFRR2 | T5TFRAverage | T5TFRRound |
|---|---|---|---|---|
| T5TFRR1 | 1 | .327$^{**}$ | .820$^{**}$ | .761$^{**}$ |
| T5TFRR2 | .327$^{**}$ | 1 | .809$^{**}$ | .747$^{**}$ |
| T5TFRAverage | .820$^{**}$ | .809$^{**}$ | 1 | .926$^{**}$ |
| T5TFRRound | .761$^{**}$ | .747$^{**}$ | .926$^{**}$ | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

### Correlations for CoP5

| | T5COPR1 | T5COPR2 | T5CoPAverage | T5CoPRound |
|---|---|---|---|---|
| T5COPR1 | 1 | .335$^{**}$ | .812$^{**}$ | .736$^{**}$ |
| T5COPR2 | .335$^{**}$ | 1 | .822$^{**}$ | .765$^{**}$ |
| T5CoPAverage | .812$^{**}$ | .822$^{**}$ | 1 | .918$^{**}$ |
| T5CoPRound | .736$^{**}$ | .765$^{**}$ | .918$^{**}$ | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

### Correlations for GV5

| | T5GVR1 | T5GVR2 | T5GVAverage | T5GVRound |
|---|---|---|---|---|
| T5GVR1 | 1 | .453$^{**}$ | .853$^{**}$ | .782$^{**}$ |
| T5GVR2 | .453$^{**}$ | 1 | .852$^{**}$ | .788$^{**}$ |
| T5GVAverage | .853$^{**}$ | .852$^{**}$ | 1 | .921$^{**}$ |
| T5GVRound | .782$^{**}$ | .788$^{**}$ | .921$^{**}$ | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations for Pron5

| | T5PronR1 | T5PronR2 | T5PronAverage | T5PronRound |
|---|---|---|---|---|
| T5PronR1 | 1 | .423** | .852** | .784** |
| T5PronR2 | .423** | 1 | .835** | .769** |
| T5PronAverage | .852** | .835** | 1 | .920** |
| T5PronRound | .784** | .769** | .920** | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

# Appendix 6. Correlation matrix of the analytic scores

| | T1TFR | T4TFR | T5TFR | T1CoP | T3COP | T5COP | T1GV | T2GV | T5GV | T1Pron | T3Pron | T5Pron | T2CoFlu | T4CoFlu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T1TFR | 1 | .186** | .354** | .584** | .329** | .333** | .478** | .381** | .359** | .365** | .291** | .279** | .384** | .241** |
| T4TFR | .186** | 1 | .236** | .253** | .225** | .219** | .259** | .264** | .263** | .230** | .226** | .214** | .282** | .543** |
| T5TFR | .354** | .236** | 1 | .448** | .492** | .662** | .437** | .458** | .579** | .408** | .431** | .453** | .505** | .351** |
| T1CoP | .584** | .253** | .448** | 1 | .475** | .532** | .664** | .569** | .532** | .572** | .492** | .454** | .538** | .366** |
| T3COP | .329** | .225** | .492** | .475** | 1 | .548** | .471** | .494** | .542** | .446** | .536** | .482** | .496** | .331** |
| T5COP | .333** | .219** | .662** | .532** | .548** | 1 | .480** | .509** | .651** | .454** | .495** | .527** | .541** | .359** |
| T1GV | .478** | .259** | .437** | .664** | .471** | .480** | 1 | .690** | .651** | .642** | .555** | .531** | .543** | .382** |
| T2GV | .381** | .264** | .458** | .569** | .494** | .509** | .690** | 1 | .664** | .595** | .575** | .543** | .623** | .391** |
| T5GV | .359** | .263** | .579** | .532** | .542** | .651** | .651** | .664** | 1 | .552** | .604** | .622** | .533** | .409** |
| T1Pron | .365** | .230** | .408** | .572** | .446** | .454** | .642** | .595** | .552** | 1 | .718** | .735** | .557** | .368** |
| T3Pron | .291** | .226** | .431** | .492** | .536** | .495** | .555** | .575** | .604** | .718** | 1 | .784** | .533** | .360** |
| T5Pron | .279** | .214** | .453** | .454** | .482** | .527** | .531** | .543** | .622** | .735** | .784** | 1 | .515** | .338** |
| T2CoFlu | .384** | .282** | .505** | .538** | .496** | .541** | .543** | .623** | .533** | .557** | .533** | .515** | 1 | .401** |
| T4CoFlu | .241** | .543** | .351** | .366** | .331** | .359** | .382** | .391** | .409** | .368** | .360** | .338** | .401** | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix 7. Standardized factor loadings for the CTCU model

| | TFR | CoP | GV | Pron | CoFLu |
|---|---|---|---|---|---|
| TFR1 | 0.51* | | | | |
| TFR4 | 0.36* | | | | |
| TFR5 | 0.70* | | | | |
| CoP1 | | 0.72* | | | |
| CoP3 | | 0.70* | | | |
| CoP5 | | 0.74* | | | |
| GV1 | | | 0.80* | | |
| GV2 | | | 0.83* | | |
| GV5 | | | 0.83* | | |
| Pron1 | | | | 0.84* | |
| Pron3 | | | | 0.88* | |
| Pron5 | | | | 0.88* | |
| CoFLu2 | | | | | 0.76* |
| CoFlu4 | | | | | 0.52* |

* p<.05

Correlations among the five traits:

| | TFR | CoP | GV | Pron | CoFlu |
|---|---|---|---|---|---|
| TFR | 1.00 | | | | |
| CoP | 0.92* | 1.00 | | | |
| GV | 0.83* | 0.87* | 1.00 | | |
| Pron | 0.68* | 0.76* | 0.79* | 1.00 | |
| CoFlu | 0.96* | 0.96* | 0.88* | 0.80* | 1.00 |

*p<.05

## Appendix 8. Unique Variance and Covariance (Standard Error) Estimates

| | TFR1 | TFR4 | TFR5 | CoP1 | CoP3 | CoP5 | GV1 | GV2 | GV5 | Pron1 | Pron3 | Pron 5 | CoFlu2 | CoFlu4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | 0.73 (0.04)* | | | | | | | | | | | | | |
| TFR4 | | 0.87 (0.04)* | | | | | | | | | | | | |
| TFR5 | | | 0.51 (0.03)* | | | | | | | | | | | |
| CoP1 | 0.24 (0.02)* | | | 0.47 (0.03)* | | | | | | | | | | |
| CoP3 | | | | | 0.51 (0.03)* | | | | | | | | | |
| CoP5 | | | 0.19 (0.02)* | | | 0.46 (0.32)* | | | | | | | | |
| GV1 | 0.13 (0.02)* | | | 0.15 (0.02)* | | | 0.35 (0.02)* | | | | | | | |
| GV2 | | | | | | | | 0.32 (0.02)* | | | | | | |
| GV5 | | | 0.11 (0.02)* | | | 0.13 (0.02)* | | | 0.33 (0.02)* | | | | | |
| Pron1 | 0.07 (0.02)* | | | 0.10 (0.02)* | | | 0.10 (0.02)* | | | 0.30 (0.02)* | | | | |
| Pron3 | | | | | 0.06 (0.01)* | | | | | | 0.23 (0.02)* | | | |
| Pron5 | | | 0.05 (0.02)* | | | 0.06 (0.01)* | | | 0.06 (0.01)* | | | 0.23 (0.02)* | | |
| CoFLu2 | | | | | | | | 0.08 (0.02)* | | | | | 0.42 (0.03)* | |
| CoFlu4 | | 0.36 (0.03)* | | | | | | | | | | | | 0.80 (0.03)* |

* P<0.05

# Appendix 9. Standardized factor loadings for the CTUM model

| | TFR | CoP | GV | Pron | CoFLu | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | 0.51* | | | | | 0.45* | | | | |
| TFR4 | 0.36* | | | | | | | | 0.60* | |
| TFR5 | 0.70* | | | | | | | | | 0.39* |
| CoP1 | | 0.72* | | | | 0.50* | | | | |
| CoP3 | | 0.70* | | | | | | 0.24* | | |
| CoP5 | | 0.74* | | | | | | | | 0.45* |
| GV1 | | | 0.80* | | | 0.31* | | | | |
| GV2 | | | 0.83* | | | | 0.27* | | | |
| GV5 | | | 0.83* | | | | | | | 0.29* |
| Pron1 | | | | 0.84* | | 0.21* | | | | |
| Pron3 | | | | 0.88* | | | | 0.24* | | |
| Pron5 | | | | 0.88* | | | | | | 0.15* |
| CoFLu2 | | | | | 0.76* | | 0.27* | | | |
| CoFlu4 | | | | | 0.52* | | | | 0.60* | |

*p<0.05

Correlations among the five traits:

| | TFR | CoP | GV | Pron | CoFlu |
|---|---|---|---|---|---|
| TFR | 1.00 | | | | |
| CoP | 0.93* | 1.00 | | | |
| GV | 0.82* | 0.87* | 1.00 | | |
| Pron | 0.67* | 0.75* | 0.80* | 1.00 | |
| CoFlu | 0.96* | 0.96* | 0.88* | 0.80* | 1.00 |

*p<.05

Appendix 10. *Standardized factor loadings for the HTCU model*

| Variable | TFR | CoP | GV | Pron | CoFLu | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | 0.50* | | | | | 0.47* | | | | |
| TFR4 | 0.36* | | | | | | | | 0.60 | |
| TFR5 | 0.69* | | | | | | | | | 0.42* |
| CoP1 | | 0.72* | | | | 0.51* | | | | |
| CoP3 | | 0.70* | | | | | | -0.23* | | |
| CoP5 | | 0.73* | | | | | | | | 0.50* |
| GV1 | | | 0.80* | | | 0.30* | | | | |
| GV2 | | | 0.83* | | | | 0.26* | | | |
| GV5 | | | 0.83* | | | | | | | 0.27* |
| Pron1 | | | | 0.84* | | 0.21* | | | | |
| Pron3 | | | | 0.88* | | | | -0.23* | | |
| Pron5 | | | | 0.88* | | | | | | 0.12* |
| CoFLu2 | | | | | 0.75* | | 0.26* | | | |
| CoFlu4 | | | | | 0.52* | | | | 0.60 | |

*p<0.05

Regression coefficients from higher-order factor to first-order factors:

| | Overall speaking ability |
|---|---|
| TFR | 0.93* |
| CoP | 0.96* |
| GV | 0.92* |
| Pron | 0.82* |
| CoFlu | 0.99* |

*p<0.05

## Appendix 11. Standardized factor loadings for the HTUM model

| Variable | TFR | CoP | GV | Pron | CoFLu | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|---|---|---|---|---|
| TFR1 | 0.50* | | | | | 0.47* | | | | |
| TFR4 | 0.36* | | | | | | | | 0.60* | |
| TFR5 | 0.69* | | | | | | | | | 0.42* |
| CoP1 | | 0.72* | | | | 0.51* | | | | |
| CoP3 | | 0.70* | | | | | | -0.23* | | |
| CoP5 | | 0.73* | | | | | | | | 0.50* |
| GV1 | | | 0.80* | | | 0.30* | | | | |
| GV2 | | | 0.83* | | | | 0.26* | | | |
| GV5 | | | 0.83* | | | | | | | 0.27* |
| Pron1 | | | | 0.84* | | 0.21* | | | | |
| Pron3 | | | | 0.88* | | | | -0.23* | | |
| Pron5 | | | | 0.88* | | | | | | 0.12* |
| CoFLu2 | | | | | 0.75* | | 0.26* | | | |
| CoFlu4 | | | | | 0.52* | | | | 0.60* | |

*p<0.05

Regression coefficients from higher-order factor to first-order factors:

| | Overall speaking ability |
|---|---|
| TFR | 0.93* |
| CoP | 0.96* |
| GV | 0.92* |
| Pron | 0.82* |
| CoFlu | 0.99* |

*p<0.05

# Appendix 12. Correlation matrix of 14 items in the questionnaire

| | TFR1S | TFR4S | TFR5S | CoP1S | CoP3S | CoP5S | GV1S | GV2S | GV5S | Pron1S | Pron3S | Pron5S | CoFlu2S | CoFLu4S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TFR1S | 1 | .638$^{**}$ | .665$^{**}$ | .791$^{**}$ | .629$^{**}$ | .632$^{**}$ | .691$^{**}$ | .628$^{**}$ | .599$^{**}$ | .652$^{**}$ | .588$^{**}$ | .576$^{**}$ | .617$^{**}$ | .610$^{**}$ |
| TFR4S | .638$^{**}$ | 1 | .758$^{**}$ | .638$^{**}$ | .722$^{**}$ | .727$^{**}$ | .619$^{**}$ | .676$^{**}$ | .713$^{**}$ | .607$^{**}$ | .710$^{**}$ | .693$^{**}$ | .691$^{**}$ | .804$^{**}$ |
| TFR5S | .665$^{**}$ | .758$^{**}$ | 1 | .649$^{**}$ | .743$^{**}$ | .853$^{**}$ | .621$^{**}$ | .733$^{**}$ | .771$^{**}$ | .651$^{**}$ | .723$^{**}$ | .771$^{**}$ | .702$^{**}$ | .736$^{**}$ |
| CoP1S | .791$^{**}$ | .638$^{**}$ | .649$^{**}$ | 1 | .680$^{**}$ | .680$^{**}$ | .799$^{**}$ | .693$^{**}$ | .640$^{**}$ | .739$^{**}$ | .635$^{**}$ | .645$^{**}$ | .677$^{**}$ | .654$^{**}$ |
| CoP3S | .629$^{**}$ | .722$^{**}$ | .743$^{**}$ | .680$^{**}$ | 1 | .787$^{**}$ | .696$^{**}$ | .767$^{**}$ | .742$^{**}$ | .676$^{**}$ | .803$^{**}$ | .738$^{**}$ | .764$^{**}$ | .749$^{**}$ |
| CoP5S | .632$^{**}$ | .727$^{**}$ | .853$^{**}$ | .680$^{**}$ | .787$^{**}$ | 1 | .655$^{**}$ | .703$^{**}$ | .795$^{**}$ | .667$^{**}$ | .737$^{**}$ | .802$^{**}$ | .740$^{**}$ | .773$^{**}$ |
| GV1S | .691$^{**}$ | .619$^{**}$ | .621$^{**}$ | .799$^{**}$ | .696$^{**}$ | .655$^{**}$ | 1 | .742$^{**}$ | .707$^{**}$ | .772$^{**}$ | .693$^{**}$ | .683$^{**}$ | .654$^{**}$ | .667$^{**}$ |
| GV2S | .628$^{**}$ | .676$^{**}$ | .733$^{**}$ | .693$^{**}$ | .767$^{**}$ | .703$^{**}$ | .742$^{**}$ | 1 | .789$^{**}$ | .675$^{**}$ | .777$^{**}$ | .726$^{**}$ | .779$^{**}$ | .701$^{**}$ |
| GV5S | .599$^{**}$ | .713$^{**}$ | .771$^{**}$ | .640$^{**}$ | .742$^{**}$ | .795$^{**}$ | .707$^{**}$ | .789$^{**}$ | 1 | .716$^{**}$ | .762$^{**}$ | .835$^{**}$ | .711$^{**}$ | .726$^{**}$ |
| Pron1S | .652$^{**}$ | .607$^{**}$ | .651$^{**}$ | .739$^{**}$ | .676$^{**}$ | .667$^{**}$ | .772$^{**}$ | .675$^{**}$ | .716$^{**}$ | 1 | .703$^{**}$ | .777$^{**}$ | .665$^{**}$ | .629$^{**}$ |
| Pron3S | .588$^{**}$ | .710$^{**}$ | .723$^{**}$ | .635$^{**}$ | .803$^{**}$ | .737$^{**}$ | .693$^{**}$ | .777$^{**}$ | .762$^{**}$ | .703$^{**}$ | 1 | .787$^{**}$ | .775$^{**}$ | .757$^{**}$ |
| Pron5S | .576$^{**}$ | .693$^{**}$ | .771$^{**}$ | .645$^{**}$ | .738$^{**}$ | .802$^{**}$ | .683$^{**}$ | .726$^{**}$ | .835$^{**}$ | .777$^{**}$ | .787$^{**}$ | 1 | .723$^{**}$ | .753$^{**}$ |
| CoFlu2S | .617$^{**}$ | .691$^{**}$ | .702$^{**}$ | .677$^{**}$ | .764$^{**}$ | .740$^{**}$ | .654$^{**}$ | .779$^{**}$ | .711$^{**}$ | .665$^{**}$ | .775$^{**}$ | .723$^{**}$ | 1 | .788$^{**}$ |
| CoFLu4S | .610$^{**}$ | .804$^{**}$ | .736$^{**}$ | .654$^{**}$ | .749$^{**}$ | .773$^{**}$ | .667$^{**}$ | .701$^{**}$ | .726$^{**}$ | .629$^{**}$ | .757$^{**}$ | .753$^{**}$ | .788$^{**}$ | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

## Appendix 13. Correlation matrix of the 14 analytic scores of GSLPA SLT and 14 items in the questionnaire

| | T1TF | T4TF | T5TF | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | R | R | T1CoP | T3CoP | T5CoP | T1GV | T2GV | T5GV | T1Pron | T3Pron | T5Pron | T2CoFLu | T4CoFlu |
| TFR1 | 1 | | | | | | | | | | | | | |
| TFR4 | .230$^{**}$ | 1 | | | | | | | | | | | | |
| TFR5 | .431$^{**}$ | .205$^{**}$ | 1 | | | | | | | | | | | |
| CoP1 | .561$^{**}$ | .228$^{**}$ | .461$^{**}$ | 1 | | | | | | | | | | |
| CoP3 | .439$^{**}$ | .231$^{**}$ | .520$^{**}$ | .534$^{**}$ | 1 | | | | | | | | | |
| CoP5 | .391$^{**}$ | .215$^{**}$ | .644$^{**}$ | .573$^{**}$ | .575$^{**}$ | 1 | | | | | | | | |
| GV1 | .505$^{**}$ | .285$^{**}$ | .442$^{**}$ | .674$^{**}$ | .495$^{**}$ | .521$^{**}$ | 1 | | | | | | | |
| GV2 | .376$^{**}$ | .267$^{**}$ | .466$^{**}$ | .576$^{**}$ | .525$^{**}$ | .540$^{**}$ | .678$^{**}$ | 1 | | | | | | |
| GV5 | .413$^{**}$ | .279$^{**}$ | .595$^{**}$ | .534$^{**}$ | .531$^{**}$ | .651$^{**}$ | .671$^{**}$ | .704$^{**}$ | 1 | | | | | |
| Pron1 | .369$^{**}$ | .205$^{**}$ | .424$^{**}$ | .547$^{**}$ | .430$^{**}$ | .466$^{**}$ | .614$^{**}$ | .536$^{**}$ | .555$^{**}$ | 1 | | | | |
| Pron3 | .363$^{**}$ | .189$^{**}$ | .447$^{**}$ | .503$^{**}$ | .481$^{**}$ | .483$^{**}$ | .550$^{**}$ | .575$^{**}$ | .588$^{**}$ | .768$^{**}$ | 1 | | | |
| Pron5 | .312$^{**}$ | .200$^{**}$ | .471$^{**}$ | .433$^{**}$ | .388$^{**}$ | .506$^{**}$ | .511$^{**}$ | .501$^{**}$ | .575$^{**}$ | .784$^{**}$ | .770$^{**}$ | 1 | | |
| CoFLu2 | .395$^{**}$ | .264$^{**}$ | .479$^{**}$ | .536$^{**}$ | .505$^{**}$ | .553$^{**}$ | .528$^{**}$ | .611$^{**}$ | .577$^{**}$ | .569$^{**}$ | .548$^{**}$ | .537$^{**}$ | 1 | |
| CoFlu4 | .304$^{**}$ | .501$^{**}$ | .303$^{**}$ | .324$^{**}$ | .277$^{**}$ | .343$^{**}$ | .426$^{**}$ | .417$^{**}$ | .435$^{**}$ | .370$^{**}$ | .297$^{**}$ | .325$^{**}$ | .393$^{**}$ | 1 |
| TFR1S | .148$^{**}$ | .055 | .082 | .101 | .110$^{*}$ | .106$^{*}$ | .124$^{*}$ | .153$^{**}$ | .133$^{*}$ | .153$^{**}$ | .130$^{*}$ | .121$^{*}$ | .131$^{*}$ | .171$^{**}$ |
| TFR4S | .159$^{**}$ | .110$^{*}$ | .108$^{*}$ | .165$^{**}$ | .217$^{**}$ | .124$^{*}$ | .146$^{**}$ | .209$^{**}$ | .108$^{*}$ | .209$^{**}$ | .193$^{**}$ | .168$^{**}$ | .199$^{**}$ | .191$^{**}$ |
| TFR5S | .134$^{*}$ | .048 | .101 | .130$^{*}$ | .197$^{**}$ | .142$^{**}$ | .152$^{**}$ | .217$^{**}$ | .119$^{*}$ | .162$^{**}$ | .152$^{**}$ | .115$^{*}$ | .174$^{**}$ | .184$^{**}$ |
| CoP1S | .149$^{**}$ | .039 | .060 | .137$^{*}$ | .163$^{**}$ | .095 | .161$^{**}$ | .207$^{**}$ | .123$^{*}$ | .213$^{**}$ | .155$^{**}$ | .145$^{**}$ | .115$^{*}$ | .208$^{**}$ |
| CoP3S | .159$^{**}$ | .039 | .104 | .156$^{**}$ | .186$^{**}$ | .135$^{*}$ | .177$^{**}$ | .243$^{**}$ | .148$^{**}$ | .211$^{**}$ | .200$^{**}$ | .136$^{*}$ | .189$^{**}$ | .200$^{**}$ |
| CoP5S | .144$^{**}$ | .038 | .118$^{*}$ | .172$^{**}$ | .190$^{**}$ | .166$^{**}$ | .202$^{**}$ | .261$^{**}$ | .156$^{**}$ | .224$^{**}$ | .206$^{**}$ | .168$^{**}$ | .214$^{**}$ | .209$^{**}$ |
| GV1S | .167$^{**}$ | -.006 | .109$^{*}$ | .147$^{**}$ | .176$^{**}$ | .109$^{*}$ | .155$^{**}$ | .221$^{**}$ | .141$^{**}$ | .182$^{**}$ | .152$^{**}$ | .120$^{*}$ | .154$^{**}$ | .192$^{**}$ |
| GV2S | .141$^{**}$ | -.003 | .126$^{*}$ | .144$^{**}$ | .172$^{**}$ | .127$^{*}$ | .180$^{**}$ | .221$^{**}$ | .131$^{*}$ | .214$^{**}$ | .195$^{**}$ | .164$^{**}$ | .153$^{**}$ | .147$^{**}$ |
| GV5S | .138$^{**}$ | .031 | .112$^{*}$ | .152$^{**}$ | .199$^{**}$ | .157$^{**}$ | .154$^{**}$ | .250$^{**}$ | .126$^{*}$ | .205$^{**}$ | .179$^{**}$ | .129$^{*}$ | .221$^{**}$ | .138$^{**}$ |
| Pron1S | .206$^{**}$ | .013 | .097 | .195$^{**}$ | .152$^{**}$ | .154$^{**}$ | .179$^{**}$ | .234$^{**}$ | .152$^{**}$ | .204$^{**}$ | .193$^{**}$ | .142$^{**}$ | .209$^{**}$ | .187$^{**}$ |
| Pron3S | .193$^{**}$ | .042 | .173$^{**}$ | .222$^{**}$ | .261$^{**}$ | .178$^{**}$ | .216$^{**}$ | .279$^{**}$ | .185$^{**}$ | .239$^{**}$ | .238$^{**}$ | .198$^{**}$ | .250$^{**}$ | .197$^{**}$ |
| Pron5S | .197$^{**}$ | .039 | .158$^{**}$ | .198$^{**}$ | .251$^{**}$ | .189$^{**}$ | .221$^{**}$ | .274$^{**}$ | .183$^{**}$ | .251$^{**}$ | .244$^{**}$ | .198$^{**}$ | .241$^{**}$ | .207$^{**}$ |
| CoFlu2S | .167$^{**}$ | .025 | .118$^{*}$ | .176$^{**}$ | .211$^{**}$ | .147$^{**}$ | .178$^{**}$ | .212$^{**}$ | .161$^{**}$ | .207$^{**}$ | .177$^{**}$ | .169$^{**}$ | .202$^{**}$ | .179$^{**}$ |
| CoFLu4S | .220$^{**}$ | .109$^{*}$ | .195$^{**}$ | .230$^{**}$ | .273$^{**}$ | .202$^{**}$ | .222$^{**}$ | .292$^{**}$ | .222$^{**}$ | .271$^{**}$ | .226$^{**}$ | .240$^{**}$ | .265$^{**}$ | .272$^{**}$ |

| | TFR1S | TFR4S | TFR5S | CoP1S | CoP3S | CoP5S | GV1S | GV2S | GV5S | Pron1S | Pron3S | Pron5S | CoFlu2S | CoFLu4S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TFR1S | 1 | | | | | | | | | | | | | |
| TFR4S | .638$^{**}$ | 1 | | | | | | | | | | | | |
| TFR5S | .665$^{**}$ | .758$^{**}$ | 1 | | | | | | | | | | | |
| CoP1S | .791$^{**}$ | .638$^{**}$ | .649$^{**}$ | 1 | | | | | | | | | | |
| CoP3S | .629$^{**}$ | .722$^{**}$ | .743$^{**}$ | .680$^{**}$ | 1 | | | | | | | | | |
| CoP5S | .632$^{**}$ | .727$^{**}$ | .853$^{**}$ | .680$^{**}$ | .787$^{**}$ | | | | | | | | | |
| GV1S | .691$^{**}$ | .619$^{**}$ | .621$^{**}$ | .799$^{**}$ | .696$^{**}$ | .655$^{**}$ | 1 | | | | | | | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GV2S | .628** | .676** | .733** | .693** | .767** | .703** | .742** | 1 | | | | | | |
| GV5S | .599** | .713** | .771** | .640** | .742** | .795** | .707** | .789** | 1 | | | | | |
| Pron1S | .652** | .607** | .651** | .739** | .676** | .667** | .772** | .675** | .716** | 1 | | | | |
| Pron3S | .588** | .710** | .723** | .635** | .803** | .737** | .693** | .777** | .762** | .703** | 1 | | | |
| Pron5S | .576** | .693** | .771** | .645** | .738** | .802** | .683** | .726** | .835** | .777** | .787** | 1 | | |
| CoFlu2S | .617** | .691** | .702** | .677** | .764** | .740** | .654** | .779** | .711** | .665** | .775** | .723** | 1 | |
| CoFLu4S | .610** | .804** | .736** | .654** | .749** | .773** | .667** | .701** | .726** | .629** | .757** | .753** | .788** | 1 |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

# References

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. Firth (Eds.), *Measuring spoken proficiency* (pp.1-6). Washington, DC: Georgetown University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19* (4), 453-476.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L., Lynch, B. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language* Testing, *12,* 238–57.

Bachman, L. F. & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86.

Bachman, L. F. & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

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

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

Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380-390.

Birckbichler, D., Corl, K., & Deville, C. (1993). The dynamics of language program testing: Implications for articulation and program revision. *The Dynamics of Language Program Direction.* Heinle & Heinle, Boston, MA.

Blanche, P. & Merino, B. (1989). Self-assessment of foreign language skills: implications for teachers and researchers. *Language Learning, 39,* 313–40.

Brantmeier, C. (2005). Non-Linguistic variables in advanced L2 reading: learner's self-assessment and enjoyment. *Foreign Language Annals, 38* (4), 493–503.

Brennan, R. L. (2001a). *urGENOVA* (Version 2.1) [Computer software]. Iowa City, IA: University of Iowa.

Brennan, R. L. (2001b). *mGENOVA* (Version 2.1) [Computer software]. Iowa City, IA: University of Iowa.

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

Brown, J. D. **(**1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*(2), 217–38.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks*. (TOEFL Monograph No. MS-29). Princeton, NJ: ETS.

Brown, J. D. & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21–42.

Cai, L., du Toit, S. H. C., & Thissen, D. (2012). IRTPRO (Version 2.1). [Computer software]. Chicago, IL: Scientific Software International.

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

Carr, N. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23*(3), 269-289.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369-383.

Chalhoub-Deville, M. & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 815–832). Harlow, England: Longman.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2004). *Issues in developing a TOEFL validity argument:* Paper presented at the 26th Annual Language Testing Research Colloquium, Temecula, CA.

Chapelle, C., Enright, M., & Jamieson, J. (Eds.) (2008). *Building a validity argument for TOEFL*. New York: Routledge/Taylor and Francis Group.

Chen, Z. & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2,* 155–163.

Chiang, C. S. & Dunkel, P. (1992). The effect of speech modification, prior knowledge and listening proficiency on EFL lecture learning. *TESOL Quarterly, 26*, 345–74.

Chiu, C. W. T. (1999). *Scoring performance assessments based on judgments: Utilizing meta-analysis to estimate variance components in generalizability theory for unbalanced situations.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement, 26*(3), 321-338.

Clapham, C. (1996). The development of IELTS: A study of the effect of background knowledge on reading comprehension. Cambridge: Cambridge University Press.

Cohen, A. D. (1998). *Strategies in learning and using a second language.* New York, NY: Longman.

Cohen, A. D. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani and H. Pierson (Eds), *Learner-directed assessment in ESL* (pp. 127-150). Mahwah, NJ: Lawrence Erlbaum.

Cohen, A. D., & Olshtain, C. (1993). The production of speech acts by EFL learners. *TESOL Quarterly, 27*(1), 33-56.

Cohen, A. D., Weaver, S. J., & Li, T-Y (1996). *The impact of strategies-based instruction on speaking a foreign language*. Minneapolis: Center for Advanced Research on Language Acquisition, University of Minnesota (CARLA Working Paper Series #4).

Cronbach, L. J. (1980). Validity on parole: how can we go straight? In *New directions for testing and measurement* (Vol. 5, pp. 99–108). San Francisco: Jossey-Bass.

De Jong, J. H. A. L. & van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In J. H. A. L. De Jong (Eds.): The Construct of Language Proficiency (pp.

112-140). Philadelphia: John Benjamin.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series No. 9). Princeton, NJ: ETS.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*, 235–54.

Ellis, R. (1985). A variable competence model of second language acquisition. *IRAL 23*, 47-59.

Embretson, S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bullentin, 93,* 179-197.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data.* Cambridge, MA: The MIT Press.

Ericsson, K. A., & Simon, H. (1993). *Protocol analysis: verbal reports as data.* Cambridge: MIT Press.

Faerch, C., & Kasper, G. (1987). From product to process – introspective methods in second language research. In F. Faerch and G. Kasper (Eds.), *Introspection in second language research* (pp. 5-23). Clevedon: Multilingual Matters Ltd.

Fulcher, G. (1995). Variable competence in second language acquisition: a problem for research methodology? *System, 23*, 25-33.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208–38.

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham and D. Corson (Eds.), *Encyclopedia of language and education. Vol 7: Language testing and assessment* (pp. 75-85). New York: Springer-Verlag.

Fulcher, G. (2003). *Testing second language speaking.* Harlow: Pearson Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* New York: Routledge.

Gafni, N. (1991). *Differential item functioning: Performance by sex on reading* comprehension *tests*. ERIC Document ED 331844. Rockville, MD: Educational Resources Information Center.

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

Green, A. (1998). Verbal protocol analysis in language testing research: A handbook. Cambridge: Cambridge University Press.

Hargan, N. (1994). Learner autonomy by remote control. *System, 22* (4), 455–462.

Heilenman, K. (1991). Self-assessment and placement: a review of the issues. In Teschner, R.V. (Eds.), *Assessing foreign language proficiency of undergraduates*: *AAUSC issues in language program direction*. Heinle & Heinle, Boston, pp. 93–114.

Higgs, T. & Clifford, R. (1982). The push towards communication. In T. V. Higgs (Eds.), Curriculum, competence, and the foreign language teacher (pp. 57-79). Lincolnwood, IL: National Textbook Company,

Ingram, D. & Wylie, E. (1993). Assessing speaking proficiency in the international English language testing system. In D. Douglas & C. Chapelle. (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium.* Alexandria, VA: TESOL, Inc.

Iwashita, N., McNamara, T. & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning*, *51*, 401–36.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F. & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Jensen, C. & Hansen, C. (1995). The effect of prior knowledge on EAP listening test performance. *Language Testing, 12,* 99-119.

Jöreskog, K., & Sörbom, D. (2007). LISREL (Version 8.8) [Computer software]. Chicago, IL: Scientific Software International.

Kadir, A. K. (2008). *Framing a validity argument for test use and impact: The Malaysian public service experience*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

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

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2,* 135-170.

Kane, M. (2006). Validation. In R. L. Brennan (Eds.), *Educational measurement* (4th ed., pp. 17-64). New York: American Council on Education and Macmillan.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kenny, D.A., & Kashy, D.A. (1992). The analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165-172.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18,* 89-114.

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

Krausert, S. R., (1991). *Determining the usefulness of self-assessment of foreign language skills: post-secondary ESL students' placement contribution*. Unpublished doctoral dissertation, University of Southern California.

Kunnan, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. *Language Testing, 11*(3), 225-252.

Kunnan, A. (1995). *Test taker characteristics and test performance: A structural equation modeling approach*. Cambridge, England: Cambridge University Press.

Kunnan, A. (1998). Approaches to validation in language assessment. In Kunnan, A. (Eds.), *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, 1–16.

Lado, R. L. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. New York: McGraw-Hill.

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks* (TOEFL Res. Monograph No. MS-28). Princeton, NJ: ETS.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL Writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Res. Monograph No. MS-31). Princeton, NJ: ETS.

Linacre, J. M. & Wright, B. D. (1990). *Facets – many faceted rasch analysis*. Chicago, IL: Messa Press.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16(2),* 14-16.

Long, D. R. (1990). What you don't know can't help you. *Studies in Second Language Acquisition*, *12*, 65-80.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lumley, T. & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22(4)*, 415-437.

Lumley, T. & Qian, D. (2003). Assessing English for employment in Hong Kong. In C. A. Coombe & N. Hubley (Eds.). Assessment Practices: *Case Studies in TESOL Practice Series.* Alexandria, VA: TESOL.

Lynch, B. & McNamara, T. (1998). Using G-theory and many-faceted Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158–80.

Madsen, H. S. & Jones, R. L. (1981). Classification of Oral Proficiency Tests. In Palmer, A. S., Groot, P. J. M. & Trosper, G. A. (Eds.), *The construct validation of tests of communicative competence (pp.15-30)*. Washington, D.C.: TESOL Publications.

Marsh, H.W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52–75.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamra, T. F. & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan Publishing Company.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Mislevy, R. J. (2003). Argument substance and argument structure in educational assessment.

*Law, Probability and Risk, 2*(4), 237–258.

O'Donnell, D., Thompson, G., & Park, S. (2006). *Revisiting assessment criteria in a speaking test.* Paper Presented at JALT2006 Annual Conference, Kitakyushu, Japan.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing, *Language Testing, 19*(2), 169 – 192.

Omaggio, A. C. (1986). *Teaching language in context: Proficiency-oriented instruction*. Boston, MA: Heinle & Heinle Publishers.

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System, 30*, 143-154.

Oscarson, M. (1978). *Approaches to Self-assessment in Foreign Language Learning*. Council of Europe, Council for Cultural Cooperation, Strasbourg.

Pae, T.-I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32*, 265-281.

Pollit, A. & Murray, N. L. (1996). What raters really pay attention to. In Milanovic, M. & Saville, N. (Eds.), *Studies in Language Testing 3: Performance Testing, Cognition and Assessment.* Cambridge University Press, Cambridge.

Ross, S. J. (1998). Self-Assessment in Second Language Testing: A Meta-Analysis and Experiment with Experiential Factors. *Language Testing, 15* (1), 1-20.

Ryan, K. & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*, 12-29.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*, 95-111.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York: Lang.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks. CA: Sage.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355–90.

Sawaki, Y. (2009). Factor structure of the TOEFL Internet-based Test. *Language Testing, 26*(1), 005–030.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16(2),* 5-13.

Shin, S. K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*, 31–57.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Stricker, L. J., & Rock, D. A. (2008). *Factor Structure of the TOEFL Internet-based Test Across Subgroups* (TOEFL iBT Research Report No.7). Princeton, NJ: Educational Testing Service.

Sunderland, J. (1995). Gender and language testing. *Language Testing Update, 17*, 24-35.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302.

Takala, S. & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*, 323-340.

Tarone, E. (1988). Variation in interlanguage. London: Edward Arnold.

Toulmin, S. E. (2003). *The uses of argument (updated edition).* Cambridge, UK: Cambridge University Press.

Underhill, N. (1987). *Testing spoken English.* Cambridge, England: Cambridge University Press.

Wang, H. (2010). *Investigating the justifiability of an additional test use: An application of Assessment Use Argument to an English as a Foreign Language test*. Unpublished doctoral dissertation, University of California, Los Angeles.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–87.

Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, NJ: Prentice Hall.

Weir, C. **(**1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall International.

Wu, Y. (1998). What do tests of listening comprehension test? – A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(10), 21-44.

Xi, X. (2006). *Investigating the utility of analytic scoring for the TOEFL academic speaking test (TAST)* (TOEFL iBT Research Report No. TOEFLiBT-01). Princeton, NJ: Educational Testing Service.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136-147.